GRAM: A generalized model to predict the cell-specific experimental deleterious effect of non-coding variants

Abstract: [[SKL: need work]]

Identification and prioritization of disease associated variants become an increasing demand as next-generation sequencing data rapidly grows and accumulated. Luciferase assay and multiplex assay are widely used to verify the deleterious effect of a putative phenotypic associated variant. But the biological mechanism of experimental validation is not clear. To investigate the possible mechanism of the experimental assays, we proposed GRAM, a generailzed model using machine-learning algorithm to study the biological significance of experimental deleterious effect. We found the TF binding features are the most predictive features, and both ChIP-Seq and SELEX derived features has high contributions to the model. Just using *in vitro* SELEX features can achieve similar prediction power as using all the TF binding features. In our multi-phase classification model, we incorporate a novel set of cell-specific effect features, with TF binding features, the AUROC reachess 0.728 and outperforms all the state-of-the-art tools. Finally, we a generalized model and assess it using luciferase assay in a different cell line, resulting in high predictive performance. Our study has shed a deeper insights into the underlying biological implications of genetic variants on the middle-layer assay type.

Introduction

Genomic sequencing data and genetics-related data

Biological datasets are being generated from numerous experiments in unprecedented amounts. Next generation sequencing (NGS) has revolutionized the study of genetic diseases at the population level and boosted the field of genomic medicine, a core domain that study diseases at a molecular level. With current advancements in NGS that enables the entire human genome be sequenced in one day \cite{PMCID: PMC3841808, costseq:PMC3245608}, the emerging field of personalized medicine seem to be on the horizon.

Cancer Analysis can identify driver based recurrence

Studying the genetic components of disease focuses on the genetic variation across the population of interest. As a genetic disease and a leading cause of fatalities worldwide \cite{NCI: https://www.cancer.gov/about-cancer/understanding/statistics}, cancer has become a major disease to be studied utilizing the next-generation sequencing techniques. Consortia like PCAWG \cite{pcawg} and TCGA \cite{tcga} have orchestrated the collection of thousands of genomic datasets across cancer types and demographics, and variants with causal relationship to different cancers are being continuously identified. However, power analysis indicates not all cohort study has accumulated sufficient amount of data, thus studying the phenotype consequence of genetic variants remains a major. Generally, , identification of disease driver

variants or elements is biased by various influential factors of the study, such as number of available samples, genomic region of focus, and the availability of multi-omics study \cite{larva}.

Model and prioritization. Funseq2, GWAWA, Deepsea, CADD, LINSIGHT & Problem

As NGS rapidly become primary technique for identifying and characterizing genetic variants associated with phenotype consequences, there is an increasing need for computational methods to effectively predict the deleterious effect of variants. However, variant effect prediction and prioritization is challenging due to the high complexity of the genome, molecular process and cellular interactions, and also limitations of available technologies. A myriad of variants occur in every human genome, and noncoding regions host the majority of them \cite{Hindorff et al., 2009; Frazer et al., 2009 from GWAVA}. Computationally, a host of approaches have been developed to address the problem of variant prioritization from different perspectives. We study the performance of five major models that target noncoding variants, namely GWAVA \cite{gwava}, DeapSEA \cite{deepsea}, CADD \cite{cadd}, Funseq2\cite{funseq2}, and LINSIGHT \cite{linsight}. Most of these models combine a diverse set of genomic annotations to perform prioritization. While LINSIGHT attempts to identify the effects of noncoding variants on evolutionary fitness, GWAVA aims to prioritize causal disease variants and distinguish them from benign ones. CADD and Funseq2 combine a diverse set of annotations to prioritize noncoding variants in the human genome. DeepSEA, on the other hand, makes *de novo* predictions on noncoding variant effects on numerous molecular phenotypes mainly related to chromatin structure and accessibility. These computational methods have made many efforts to prioritize noncoding variants and some of them have already been applied to every position on the genome, However, the experimental validations are still required to further verify the prediction.

Experimental methods, such as luciferase assay and GFP assay, work as a middle-layer assay and a proxy to bridge the genotype alteration with phenotype consequence. Luciferase assay is originally used to measure the regulatory effects of noncoding elements\cite{here}. By comparing the difference of the assay readout of the elements with and without the mutation, we can estimate the experimental deleterious effect of non-coding variants. Using high throughput methods, i.e. microarray and NGS, multiplexed assays (MPRA) \cite{27701403} has extended the scales to genome wide levels [starseq and MPRA paper]. Measuring the effects of variants from the molecular level can be used to verify predicted deleterious effect of variants, but these assays usually are mediated by plasmid or virus, and thus cannot reflect a actual *in vivo* environment where the variance located. More sophisticated technique, like organoid level or in situ CRISPR genome editing, may provide more promising and reliable evaluation, but their applications are still limited and out of our scope.

Motivation[[explain the experiment results]]

These middle-layer assays mentioned above represent a crucial step closer to measuring variant effects with relative low cost, precision, high throughput experimental nature, However,

underlining biological significance of these experimental results is not clear. In this paper, we approach data mining methods from a new perspective to further bridge the gap between the genotype and phenotype, and try to predict the experimental deleterious effect of variants by the middle layer assays, which is not limited to MPRA and Luciferase. We have built a regression model to maximally use all of information available from the MPRA and identified highly associated transcription factors using a comprehensive feature selection framework; and then we developed a multi-stage classifier, which considers a novel set of cell-specific effect features and transcription factor binding, and has achieved the highest performance compared with the state-of-the-art models. Finally, we build a generalized model and assess it using luciferase assay in a different cell line **(multiple cell lines if NCVARG data still can be used)**, resulting in high predictive performance. Our study has shed a deeper insights into the underlying biological implications of genetic variants on the middle-layer assay type.

Results

Flowchart

To study the non-coding variants effect in vivo is difficult because of two major reasons: firstly, it is costly and time consuming to introduce a point mutation to the genome, though CRISPR technology can potentially solve this problem; secondly, it is impossible to evaluate the effect because there is no direct metrics for indicating the deleterious effect of a mutation on the genome. So plasmid-based or virus–based experimental assays can compromise this problem by a non-integration or randomly integration of genome to detect expession level of reporter gene. In this study, as described in Fig1a, we firstly collected dataset from paper \cite{Ryan paper}, which is the largest dataset so far for estimation of expression modulation differences between wild-type and mutants in GM12878. The predictor features are extracted according to Ryan's cell paper along with the knowledge from the other variants prioritization studies, which include evolutionary feature, cell-specific ChIP-Seq and TF binding feature for the SNV, CAGE features and motif binding features. The motif binding features are generated using PWM-based binding affinity change or Deepbind bind scores. We firstly trained a regression mode with 10-fold cross-validation and found the log-skew (fold change of mutation over the fold change of wild-type) can be well predicted using the above features. Because transcription factor bindings are thought to be the most impacting factor that affect the regulatory activity of element and the fact that the chromatin environment context on a plasmid will be lost, we further investigate the importance of TF binding scores from 515 Deepbind models. Then we consider generalizing our model to other assay platform, like Luciferase assay. In order to make log skew values from MPRA comparable to florescence readouts of luciferase assay, we discretized the logskew value to expression modulating variants (emVAR) and non-expression modulating variants (non-emVAR) as described in cell paper and developed a multiple-phases classification model. In the end, luciferase experiments are then used to evaluate the model.

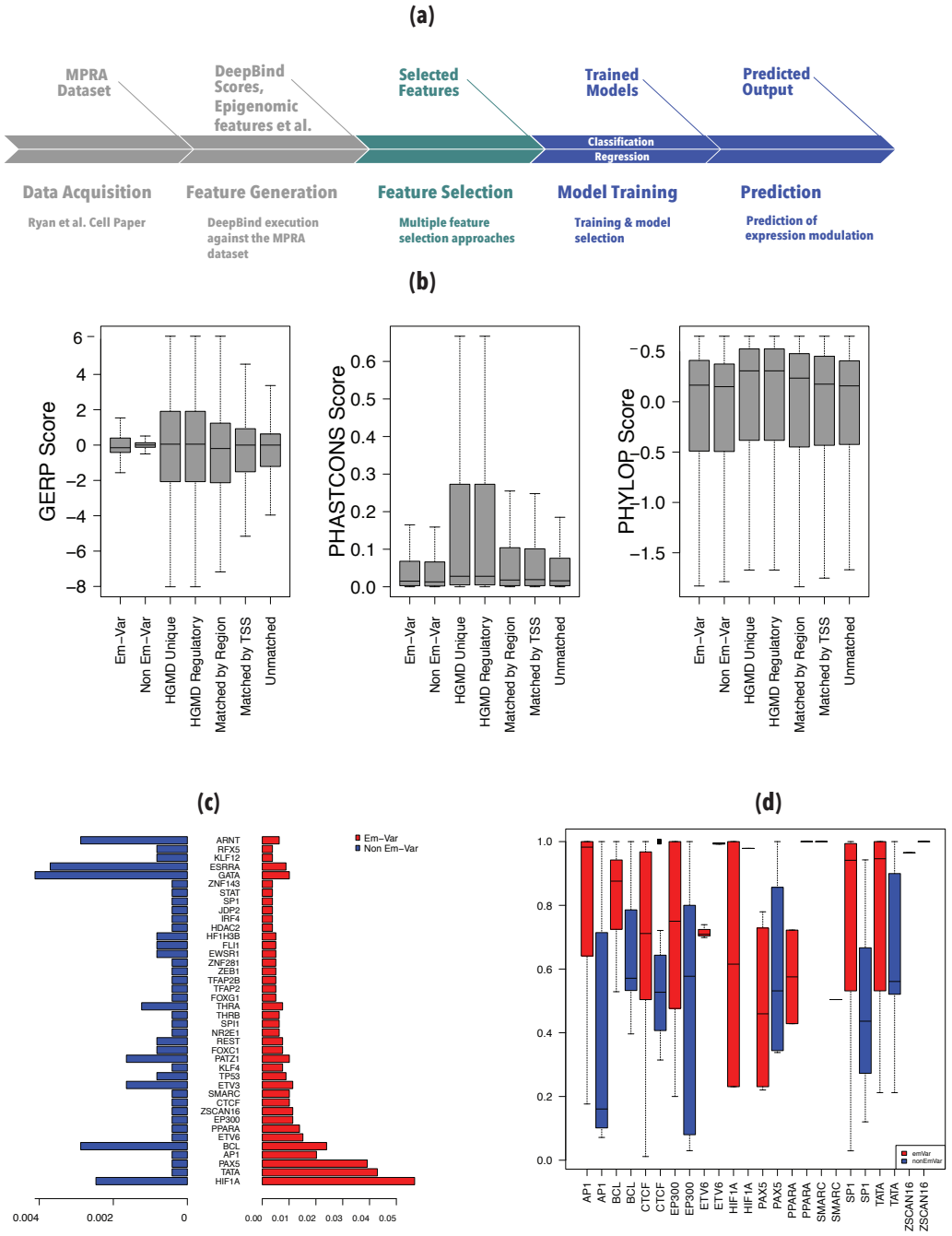


Figure 1 (a) flowchart of our study. (b) Conservation scores***(c)*** **|** MOTIFBR - motif-based - P-value (bottom- sorted up increasing order) ***(d) motif score changes***

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 and Phastcons in LINSIGHT

and CADD, GERP in Funseq2. However, because the difference in assay-based experiment is that chromatin context was diminished or randomized (Lenti-virus integration), we then questioned whether the experimental deleterious effect of these assays are still associated with the evolutionary conservation features. 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 benign variant. GERP score show similar pattern but more centered in emVAR and nonEmVar compared to other datasets, with a slightly larger values for emVAR.[[SKL: maybe need re-draw GERP one, for the y-scales are different]]. Since no different patterns found between emVar and non-emVAR, we further found 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 very minor contributions to experimental deleterious effect.
[[SKL: later, we will only consider GERP, here need mention why we only choose GERP later]]

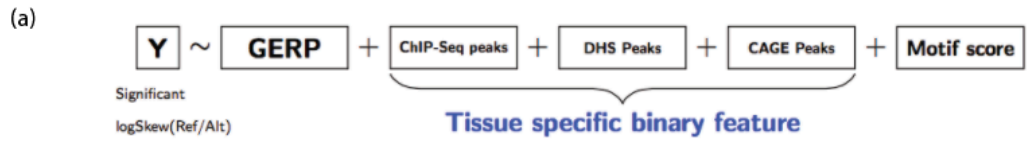
Transcription factor binding can link the deleterious effect of noncoding variants to a cascade regulatory network, which is thought to be an important factor for regulatory effect (cadd, funseq, deepsea and deepbind). In Ryan's paper, they found the log skew positively associate with TF binding scores. To thoroughly look into the effect of TF binding, we tested all xxx TF motif break events or peaks overlap with the SNVs in the dataset. Two set variants: emVAR and non-emVAR, were annotated and analyzed by Funseq2 \cite{funseq2}. The enrichment of transcription factors' motifs in both sets, with ones with lowest p-values according to the hypergeometric distribution test are shown in a bottom-up increasing order in Figures 1c, respectively. emVAR set have more TF binding events compared with non-emVAR set. The top highly enriched TF in emVAR are: xxxxx, . Besides the TF binding enrichment, we also further look at the motif break scores for these TFs, especially top enriched TFs. The largest differential scores correspond to AP1 and EP300 motifs. In addition, for a smaller subset of motifs with lowest p-values, the distribution of difference between 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 comparision, the emVAR set not only tend to have more TF binding events, but also have larger binding alteration compared with non-emVAR set.

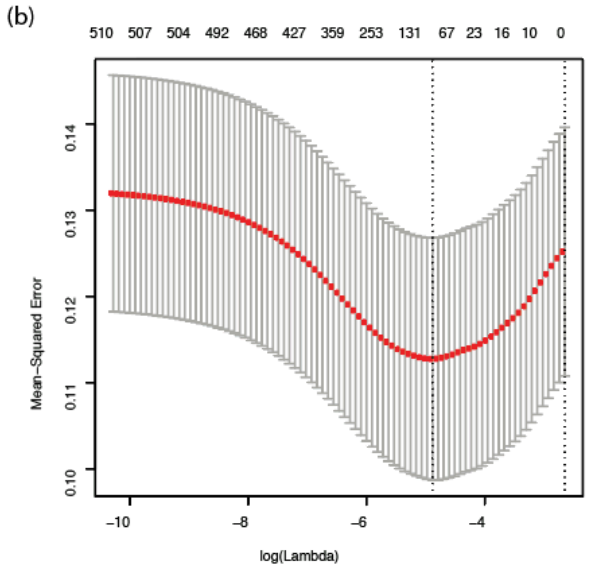
To learn the underlying patterns of variant modulated expression, we trained a host of models using a combination of epigenetic and evolutionary features. We formulate the problem as two predictive tasks: (1) a regression to predict the log-skew difference in expression modulation fold change between wild type and variant sequences, and (2) a generalized model which classifies the variant effect as two experimental deleterious class: expression modulating (emVar, label 1) or non-modulating (nonEmVar, label 0).

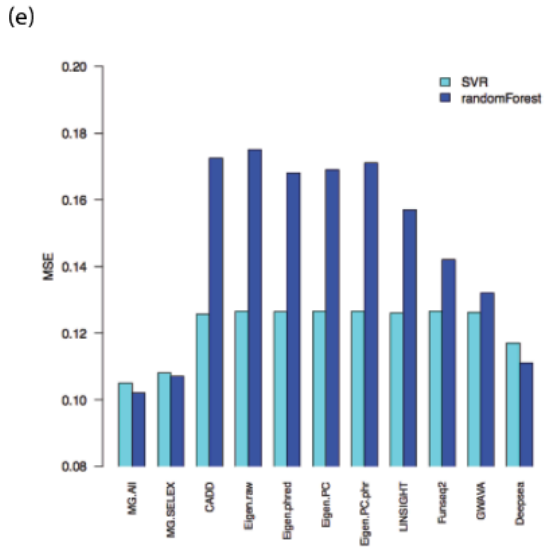
Directly predict the expression modulating changes-logSkew

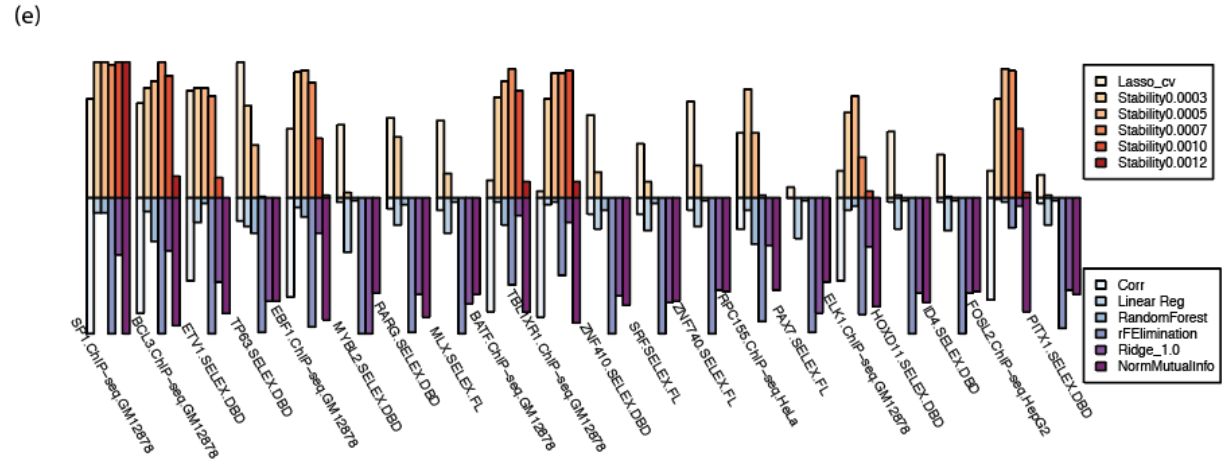
In Ryan's paper, they found histone mark and CAGE highly enriched in emVAR, which indicate these feature are potentially useful to predict the expression modulating effect. Besides, we also

add evolutionary feature and motif binding changes in our regression models We carefully consider tissue specificity differences in 1678 training samples from GM12878 cell line by removing variants without overlap with any ChIP-seq peaks and incorporating features related to the CAGE, TFs, histone marks, and DNase I hypersensitivity sites corresponding to each tissue under study. A schematic representation of the regression task is shown in Fig 2a.









(d)

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

Fig2. Regression model to predict logSkew. (a): diagram of feautures in regression model (b) Lasso regression with 10-fold cross-valiation (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.

We firstly learned Lasso regression model with 10-fold cross-validation. The fine-tuning of ƛ, the penalty parameter in the cost function of this model, determined according to the mean-squared error (MSE) values is shown in Fig 2b, with a the best performance log(ƛ) ≅ -5.The R-square for prediction is 0.39 and 0.29 with and without considering histone mark, TF ChIP-Seq and CAGE peaks respectively. However, the most important features according to Lasso regression are TF binding features, why GERP scores just show very minor contributions. But it is still not clear how the ChIP-Seq and CAGE peaks contribute on the model, since there indeed no epigenomic context on the plasmid. The only possible explanation may be these features can retein some cell-specific information, e.g. expression and regulations \cite{paper to predict expression of gene by MBG lab}.
[[SKL: why we just consider TF feature?, from biological significant, data availability and feature selection, add group comparison]]

Since TF-based binding features are top-ranked and more biologically-explainable, we then prioritized these features across models with different feature selection methods, namely: Lasso, ridge, and linear regression methods, stability selection (with five $ƛ_{stability}$ values), random forest feature importance prioritization, 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 are shown in decreasing order in Fig 2c. Expectedly, applying various methods on data with multiple dimensions leads to relatively varied results *w.r.t.* Importance to each feature across the method spectrum. However, two main conclusions can be drawn from these results. First, both ChIP-Seq and SELEX deepbind features show higher importance, with the top two being GM12878 ChIP-Seq features, and thus cell line specific. Second, almost all top features are assocated with TF-binding, what emphasizes the significance of TF binding features in studying variant expression modulation. The top two features are SP1 And BCL3 (both from cell line GM12878) , followed by a number of SELEX features starting with ETV1 and ETP63.
[[SKL: how to explain these features effect? Network degree or regulatory effect, or may also some biases]]

After considering feature importance values as per different criteria, we assess the accuracy of each of SVR (support vector repressor), Lasso, and Random forest regression models. Interestingly, the incorporation of Chip-Seq-based DeepBind features, which are cell-specific, does not boost the accuracy significantly for all three models. MSE values of both models, with and without Chip-Seq-based features, are shown in Fig 2d. Results suggest that we can reliably deploy the model trained on cell-line-independent SELEX features to predict logSkew of modulation value on samples from cell lines different from GM12878 used in training. Deepbind TF Chip-Seq model are cell line-specific, and adding them to the model shows no dramatic improvement. Thus, we can rely on the model and use it in task involving cell line independent

features only to build a generalized model since not all the cell lines have TF ChIP-Seq experiment that can be used to infer ChIP Deepbind binding model.

We then compare a Support Vector Regressor and Random Forest performance when trained on all DeepBind features, DeepBind SELEX features, and the feature values generated by each of CADD, Funseq2, DeepSEA, GWAVA, LINSIGHT, Eigen decomposition, PCA, and Eigen.PC.phr. As shown in Fig 2e, our model with DeepBind features lead to the best trained model with the lowest mean squared error for both models. In confirmation of the previous findings, the removal of ChIP-Seq Deepbind features does not cause a significant deterioration in models' predictive quality yet simplifies training. As for other methods, results show that DeepSEA features result the third best set of models (SVR and RF). For Deepsea's prediction of deleterious effect according the label of emVar and non-emVAR, The AUC is 0.5, which indicate either the experimental deleterious effect is not equivalent with phenotypic consequence or generality limitation of the model.

Build a generalized model by multi-phase learning

Our regression model has shown promising performance. However, instead of estimating the log skew value based on reads count as in MPRA, other different types of assay, such as Luciferease assay, GFP assay, and Lenti-virus based platforms, may use fluorescence readouts and different statistics methods or cutoff to decide the effects of the variants. Thus although these different platforms may have consistent result \cite{xxx}, translation of MPRA between the outputs of these assays would be difficult. In order to build a generalized model, we need tackle two challenges: firstly, making a unified target that can be used for comparison cross these different assay types, and a classification model will be a good choice since different assay platform use different statistics to distinguish the experimental deleterious variants; secondly, considering the cell-specific information that are more easily obtained compared with ChIP-Seq experiments which is not widely tested in all the cell lines. We will use gene expression data of transcription factor to represent cell-specific context. As a result, we developed a three-phase model to predict SNP effects.

For the Phase one, we will predict whether an element has enhancer-like regulatory activity. An element with or without mutation that inserted into plasmid are tested as enhancer-like element if the fold change between the element with the control is large than a statistically significant cutoff. For example, for MPRA study, statistical test based on DESeq2 will indicate which it is significance; while for Luciferase assay, testing element that has the fold change with control (empty plasmid or eGFP) is greater than 1.5 or 2 will be thought as enhancer-like. Using the Deepbind TF binding feature as predictor, whether is a enhancer-like element as target, a randomForest classifier was trained to predict enhance-likeness. 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)..

For phase two, we want to consider the cell specific effect in the study. 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 reflected by gene expression; and loci specifity between different loci in the same cell line or tissue which is denoted by TF binding preference and TF's expression. We found the variance or standard devitation of log odds (Vodds) to be a suitable indicator. By comparing 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 the emVAR with non-emVAR variants, the higher variance group tends to have more non-emVar. (Chi-square test p-value: 0.0002021), which indicates the emVAR class tends to have lower variances. We use TF binding score and expression ranking matrix to predict high and low Vodds classes defined by top and bottom quartile value (fig3e). TF binding score can predict the high and low classes with high AUC 0.80, and expression ranking have a AUC (0.65) is higher than a random effect (fig 3g-h).

The final phase is to predict whether the variants has 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 target. The AUROC of 10-fold cross-validation for the optimal model is 0.728 and AUPRC is 0.505, which higher than the state-of-the-art for the study using the same dataset (AUROC: 0.684, AUPRC: 0.478)\cite{David Gifford lab paper}. For a generalized model, we redo phase one and two on the same dataset by excluding Deepbind 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.

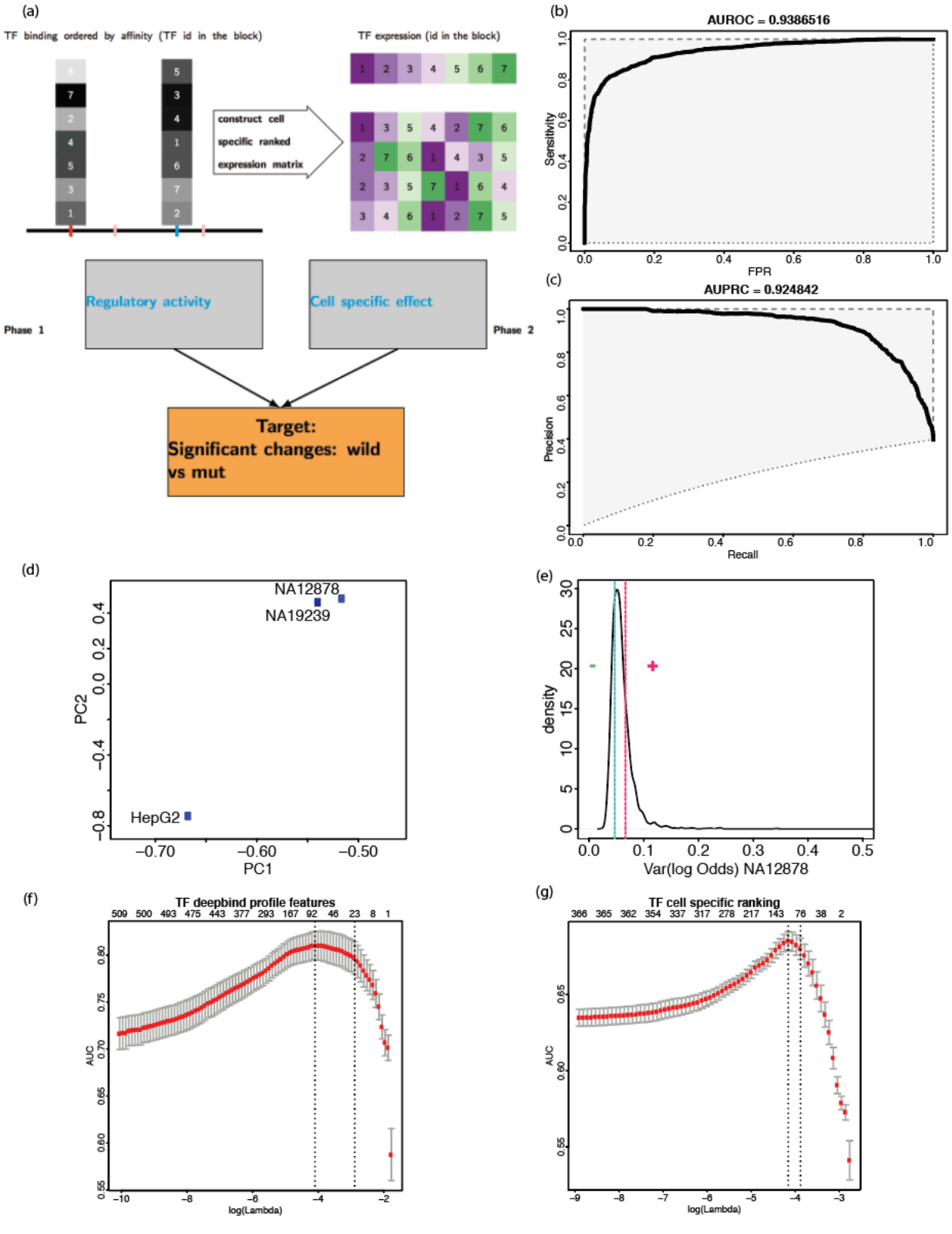

(a)
TF binding ordered by affinity (TF id in the block)
TF expression (id in the block)
construct cell specific ranked expression matrix
Regulatory activity
Cell specific effect
Phase 1
Phase 2
Target: Significant changes: wild vs mut
(b)
AUROC = 0.9386516
Sensitivity
FPR
(c)
AUPRC = 0.924842
Precision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

Fig3

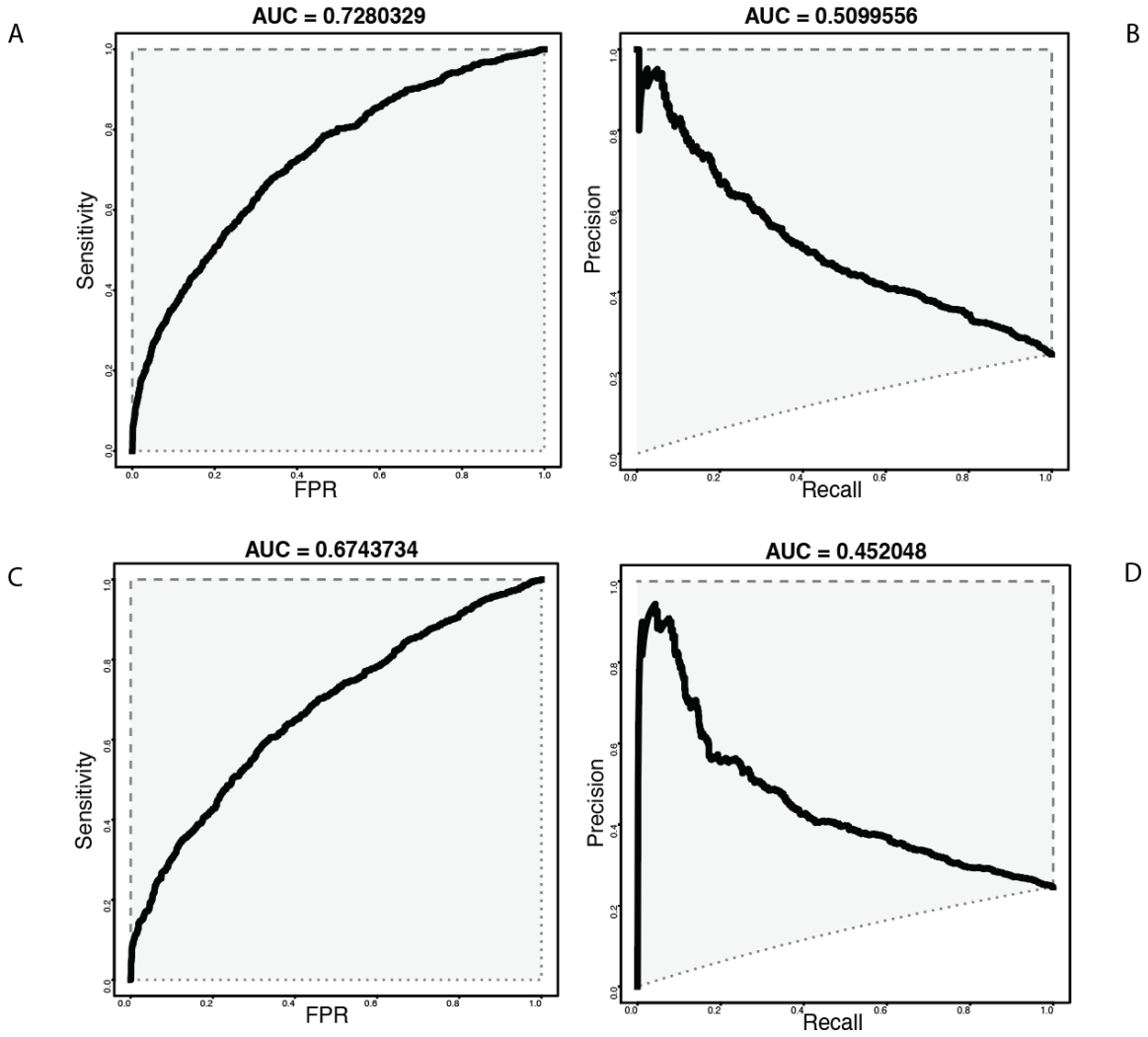


Fig4 Performance of classification model. (A,B ROC and PRC for model including tissue-specific ChIP-Seq Deepbind scores, C, D ROC and PRC for generalized model)

The generalized model is trained on Gm12878 and MPRA dataset. To evaluate the performance of the model on the other cell lines using Luciferase assay. We did luciferase assay on two different cell lines: MCF7 ~~and K562(NCVARG data quality is pool, may not used in the manuscript)~~. We select 8 potential regulatory elements from MCF7 cell line, each one with a mutation as decribed in our study \cite[ENCODEC]. We predict the enhancer-like regulatory activity for both wild-type and mutant alleles, and expression modulating deleterious differences between wide-type and mutant. For enhancer-like activity, the predicted probability to be an active regulator is positive correlated with luciferase assay fold change. The results are perfectly predict (AUROC=1) for different luciferase fold change cutoffs from 1.2 – 2 that is used to define a active enhancer (fig5a). For the prediction of deleterious effect, the significant differences

between mutant and wild-type is defined by using absolute log2(fold change) cutoff. The predicted probability also showed 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]. This indicates our model perform very well on the testing luciferase assay on a different cell line.

[[NCVARG results will hold to resolve after discussion with Sandy and Jin, may not be useful]]

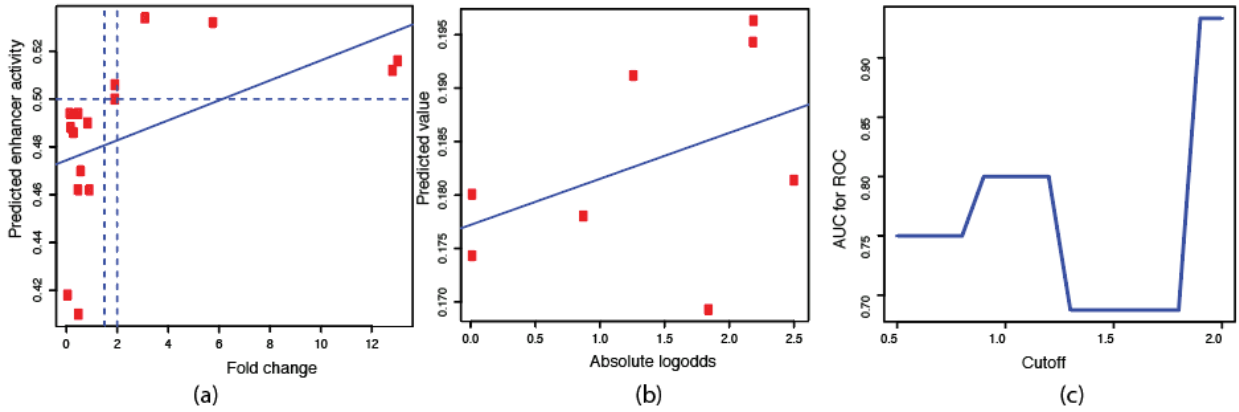


Fig5 (a) enhancer-likeness prediction. x-axis: fold change from experiment, the vertical dot lines represent the cut off (1.5, or 2) to determine positive (enhancer) and negative, the horizontal dot line is predicted probability cutoff (0.5). (b): predicted probability for emVar and non-emVAR versus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

Discussion [[need work]]

There is an increasing number of computation methods that can prioritize non-coding variants, as well as high-throughput whole-genome sequencing data that become the primary technique for identifying and characterizing disease associated variants. Experimental-based methods can bridge the gap between genetics variants and phenotypic prediction and verify the experimental deleterious effect, which works as a middle-layer type between genotype and phenotype. In this paper, we performed a thorough analysis of deleterious effect modeling on this middle-layer assay type, trained both regression and classification models using MPRA data from Gm12878 cell lines. By taking advantage of non-cell-specific SELEX TF binding feature, and easily obtained cell-specific TF expression data, we built a generalized model that can be potentialy applied to any cell lines and tissues, and predict the significant expression modulation changes for all types of experiment assay. Expeirmental validation using luciferase assay on MCF7 cell lines to further verified the generality and robustness of the model.

[[SKL: GERP score and features selection, why use the TF features,]]
[[SKL: classification and expression tissue-specific feature]]

In regression model, we tested features that maybe associated with the experimental deleterious effect. In spite of the biological insight evolutionary features provide, Lasso regression indicates that they do not rank high in significance when predicting the output of middle layer assays. The Histone Mark and CAGE features are chosen because of enrichment analysis between emVAR and non-emVAR, however, how these features works 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 imitate the application of model. While the transcription factor binding is more biological 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 are still the most important factor, but re-ordered TF expression matrix also associate with cell specific effect. however, features from a re-ordered TF expression matrix can also be problematic for some worse cases. The idea to re-order TF expression according to its binding strength or rank in its binding preference is inspired by the study of TF binding waiting time\cite{}. The waiting time of TF binding is thought to be related to TF binding free energy, which is further related to the binding scores. In our study, we just 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. However, our results indeed showed that the re-ordered expression matrix have association with the cell-specificity effect.

Though our model achieve 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 are very diverse, which will result in the overfiting of our model. Even for our experimental validation, it only includes 8 elements (suppose we will not use NCVARG data) which is far from enough to make strong conclusion of our model's robustness and generality, but at least from these very few pilot tests, our model shows an acceptance and even better performance. 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 [[SKL: need more work]]

Dataset

The data was downloaded from Ryan cell paper. 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 .

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 \cite{Alipanahi2015} 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) .

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 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(\cite{Meinshausen2010}), 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.

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 expression rank matrix.

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 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 [[SKL: from Jin Liang,but may need change to ENCODEC one if no NCVARG experiment can be used]]

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.

References:

##the end ##
[[SKL: to delete later]]

We then formulate the study of modulated expression effect of variants using classification tasks. In the first task, whose results are shown in Fig3a, we extract two main features: predicted regulatory activity by a Random Forest classification model using DeepBind TF scores, and cell specific bias, calculated using a Lasso regressor trained on TF expression data of lymphoblastoid GM12878 cell line obtained from ENCODE \cite{encode}. A more detailed description of both matrices is shown in Fig3b. For each variant position, the TF score of the

wild type at this specific position is calculated. Then, the second matrix is generated with the TF expression values from ENCODE according to the order of wild type score rank. Thus, the order of TFs in the second matrix differs for each variant according to log odd values, and the color/name of each TF is shown in each row corresponding to the variant.

Once both features are extracted, they're used by the larger Lasso classification model to predict the variant's regulatory activity on a luciferase assay (1 for active, 0 for inactive) with respect to the wild type. Results are assessed *w.r.t.* to the same emVar and nonEmVar dataset used in the previous tasks described in this paper.

For the second classification task, we train a model to predict regulatory activity based on the fold change (*fc*) in luciferase expression. Unlike in the previous task that compares expression levels of plasmids with and without the variant, we here normalize the fold change by empty plasmid. To define activity, we consider two ranges of fc values: 0.3 < *fc* as inactive (class 0) and *fc* > 1.5 as active (class 1).

CADD	http://krishna.gs.washington.edu/download/CADD/v1.2/whole_genome_SNVs.tsv.gz	