Published online

Nucleic Acids Research, , Vol. 0, No. 0 **1–8**
doi:10.1093/nar/gkn000

QBiC-Pred: Quantitative Predictions of Transcription Factor Binding Changes Due to Sequence Variants

Vincentius Martin [1,2,+], Jingkang Zhao [2,3,+], Ariel Afek [2,4], Zachery Mielko [2,5] and Raluca Gordân [1,2,4,6*]

[1]Department of Computer Science, [2]Center for Genomic and Computational Biology, [3]Program in Computational Biology and Bioinformatics, [4]Department of Biostatistics and Bioinformatics, [5]Program in Genetics and Genomics, and [6]Department of Molecular Genetics and Microbiology, Duke University, Durham, NC 27708, USA.
[+]These authors contributed equally to this work

ABSTRACT

Non-coding genetic variants/mutations can play functional roles in the cell by disrupting regulatory interactions between transcription factors (TFs) and their genomic target sites. For most human TFs, a myriad of DNA-binding models are available and could be used to predict the effects of DNA mutations on TF binding. However, information on the quality of these models is scarce, making it hard to evaluate the statistical significance of predicted binding changes. Here, we present QBiC-Pred, a web server for predicting quantitative TF binding changes due to nucleotide variants. QBiC-Pred uses regression models of TF binding specificity trained on high-throughput in vitro data. The training is done using ordinary least squares (OLS), and we leverage distributional results associated with OLS estimation to compute, for each predicted change in TF binding, a p-value reflecting our confidence in the predicted effect. We show that OLS models are accurate in predicting the effects of mutations on TF binding in vitro and in vivo, outperforming widely-used PWM models as well as recently developed deep learning models of specificity. QBiC-Pred takes as input mutation data sets in several formats, and it allows post-processing of the results through a user-friendly web interface. QBiC-Pred is freely available at http://qbic.genome.duke.edu.

INTRODUCTION

Genetic variants and mutations play important roles in human disease (1). Most variants occur in non-coding genomic regions, where they can impact gene expression by disrupting interactions between transcription factors (TFs) and DNA. In previous work we have introduced an ordinary least squares (OLS)-based method for assessing the impact of non-coding mutations on TF-DNA interactions (2). Briefly, we used high-throughput *in vitro* TF binding data from universal protein-binding microarray (uPBM) experiments (3) to train regression models of TF-DNA binding specificity using OLS estimation. Next, we used the OLS models to predict changes in TF binding due to DNA mutations, and we showed that our binding change predictions correlate well with measured changes in gene expression.

Our approach is novel compared to previous models because, by using OLS, we obtain not only estimates of the model coefficients, but also the variance of these estimates, which allows us to compute normalized binding change scores (z-scores) and significance levels (p-values) reflecting our confidence that a mutation affects TF binding. The computed p-values implicitly take into account the quality of the model and of the training data, so in the case of poor predictive models a large change in binding is required for a mutation to be called significant (2).

Here, we introduce QBiC-Pred (Quantitative Predictions of TF Binding Changes Due to Sequence Variants), or QBiC for short, a web service that allows users to run our OLS models through a user-friendly web interface.

Input. QBiC takes as input mutation/variant data sets containing single nucleotide variants, in several formats: 1) variant files in the standard variant call format (VCF); 2) 'simple somatic mutations' files generated by the International Cancer Genome Consortium (ICGC) (4); 3) tab- or comma-separated values files with the columns: chromosome, chromosome_pos, mutated_from, and mutated_to; and 4) text files containing 17-bp DNA sequences with the mutated nucleotide in the center, followed by the 'mutated to' nucleotide, separated by a space character. The first three formats can be used with genomic coordinates from versions hg19 and hg38 of the reference human genome, while the sequence format allows users to input custom DNA sequences. For the sequence format, the context of each variant (8-bp on each side) is needed in order to assess the binding status of each allele, using uPBM 8-mer enrichment scores (E-scores) (3, 5). Examples of input mutations files are described in the About section of the website, and available for download. QBiC also takes as input a list of TF proteins of interest, from a list of 577 human TFs with available OLS models. All TF names are specified using the standard HUGO gene nomenclature (HGNC) (6). The list of available TFs and models is available on the QBiC website in the Downloads section.

Output. For each input variant, QBiC runs the OLS models for the list of specified human TFs, and it computes the predicted TF binding changes, the normalized changes (z-scores), the significance of the changes according to each

*To whom correspondence should be addressed. Tel: +1 (919)684-9881; Email: raluca.gordan@duke.edu

model (p-values), as well as the predicted changes in binding status (e.g. from specific binding, or ‘bound’, to nonspecific binding, or ‘unbound’) assessed using uPBM 8-mer data. Similarly to our previous work (7), we consider a site ‘bound’ if it contains two consecutive overlapping 8-mers with E-scores >0.4, and ‘unbound’ is it contains only 8-mers with E-score <0.35; all other sites are called ‘ambiguous’. The E-score cutoffs can be modified by the user through the QBiC interface. All computed values are reported as output, in table format. The precise models used by QBiC for each TF protein, as well as the PBM data used to train each model, are reported as part the QBiC results. The user can further process the results using the web interface (e.g. to specify a more stringent p-value cutoff for the binding change predictions) and can download the full or filtered results. The web interface also allows users to directly download models or data sets used to obtain individual predictions, and provides links to the HGNC database (6) where users can find additional information about individual TFs.

We are not aware of web servers with the same functionality as QBiC. Users interested in evaluating the putative effects of non-coding mutations on TF-DNA binding can certainly use any of the available databases of position weight matrices (PWMs) (e.g. (8, 9, 10, 11)) or deep learning models (12) of TF-DNA binding specificity, or search existing databases of annotations for non-coding variants (e.g. (13, 14)). However, such databases do not provide information on the quality of the binding models, and, as shown in the Results section below, PWM and deep learning models are not as accurate as our OLS models in predicting the *quantitative* effects of DNA variants on TF binding. The OLS models used in QBiC also have the advantage of providing a direct measure of the significance of each predicted TF binding change, given the model and the training data. This unique feature of our models facilitates interpretation of the results and allows users to prioritize variants for further analysis and validation.

MATERIALS AND METHODS

OLS models of TF-DNA binding specificity

The OLS models used by QBiC were trained on curated uPBM data from literature and our laboratory, mapped to 577 human TF proteins. Each uPBM experiment measures the binding specificity of a TF for ∼44,000 60-bp long DNA sequences, each containing a 36-bp variable region followed by a constant 24-bp primer complement (necessary for DNA double-stranding (3)). We use as features the number of occurrences of each possible 6-mer within the 60-bp sequences, and as outcomes the log-transformed fluorescence intensity signals, which reflect the levels of TF binding. The entire 60-bp sequence is used to count 6-mer occurrences, despite the fact that part of the sequence is constant, because the TF proteins can bind at any location within the 60-bp DNA molecule. We consider each 6-mer and its reverse complement as the same variable and combined their counts as one feature, resulting in a total of 2,080 features. The relationship between the outcomes Y and the features X is modeled by a multiple linear regression $Y=X\beta+\varepsilon$.

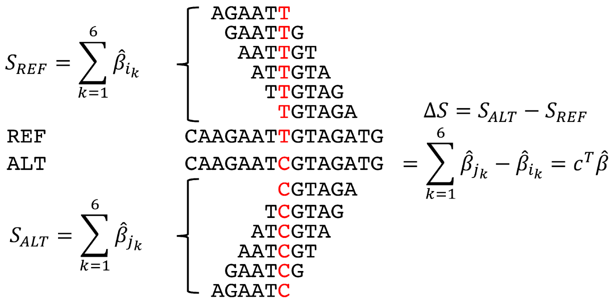

$S_{REF}=\sum_{k=1}^{6}\hat{\beta}_{i_k}$
AGAATT
GAATTG
AATTGT
ATTGTA
TTGTAG
TGTAGA
REF
CAAGAATTGTAGATG
ALT
CAAGAATCGTAGATG
$\Delta S=S_{ALT}-S_{REF}$
$=\sum_{k=1}^{6}\hat{\beta}_{j_k}-\hat{\beta}_{i_k}=c^T\hat{\beta}$
$S_{ALT}=\sum_{k=1}^{6}\hat{\beta}_{j_k}$
CGTAGA
TCGTAG
ATCGTA
AATCGT
GAATCG
AGAATC

Figure 1. The change in TF binding is computed as a linear combination of the coefficient estimates for all 6-mers overlapping the variant.

To characterize the TF binding change due to a single nucleotide variant, we define binding scores for the wild-type and the mutant sequences, as the sum of the coefficients for all 6-mers overlapping the variant, in an 11-bp window. The difference between these two scores, which represents the binding change, can be expressed as a linear combination of the regression coefficients: $c^T\beta$, where β is a vector containing the coefficients for all 2080 6-mer count features, and c is a vector of the same length containing, for each 6-mer, the difference in counts due to the variant (Figure 1). We note that most components of c are 0, as the variant affects the counts for up to twelve 6-mers.

By further assuming normality on the error term of the linear regression model $\varepsilon\sim N(0,\sigma^2 I)$, we are able to leverage the statistical properties of OLS estimation in order to test whether the binding change is statistically significant. The null hypothesis $H_0: c^T\beta=0$ can be tested using a t-statistic: $t=c^T\hat{\beta}/\sqrt{c^T\hat{\Sigma}c}$. Here, $\hat{\beta}$ is the OLS estimate for the coefficients vector: $\hat{\beta}=(X^TX)^{-1}X^TY$, and $\hat{\Sigma}$ is an unbiased estimate for the covariance matrix of $\hat{\beta}$: $\hat{\Sigma}=\hat{\sigma}^2(X^TX)^{-1}$, where $\hat{\sigma}^2=(Y-X\hat{\beta})^T(Y-X\hat{\beta})/(n-p)$, with n being the number of observations and p the number of features. Since the regression contains ∼44,000 observations and 2,080 variables, this t-statistic follows a t-distribution with ∼42,000 degrees of freedom. Thus, we can use a normal approximation to derive the z-score and calculate the p-value of the test. For each TF and variant given as input, QBiC calculates and reports the difference in TF binding, the corresponding z-score, and the associated p-value.

To select the uPBM data used in QBiC, we started with 3,342 data sets from CIS-BP (10), 245 data sets from UniPROBE (8) that were not included in CIS-BP, and 22 data sets generated in our laboratory (7). By using the information in the Human Transcription Factors database (15) for the publicly available uPBM data, and manually curating the data generated in our laboratory, we mapped 1,450 uPBM data sets to 633 human TF proteins, using both uPBM experiments that tested human TFs as well as experiments for homologous TFs with high amino-acid identity in the DNA-binding domain region, similarly to Lambert et al. (15). Next, to assess the quality of each uPBM data with respect to our task of training accurate quantitative models of TF-DNA binding specificity, we used the cross-validation accuracy of OLS models trained on each uPBM dataset. We removed data sets of poor quality (cross-validation correlation <0.2 computed for the top 10%

and top 20% sequences with the highest intensity), and for each TF we selected at most 6 uPBM data sets, including the top 3 data sets with the highest cross-validation accuracy, as well as the top 3 data sets obtained for TFs with the highest amino-acid identify to the human TFs. The final mapping, which includes 666 uPBM data sets and 577 TFs, is available on the QBiC website in the About section.

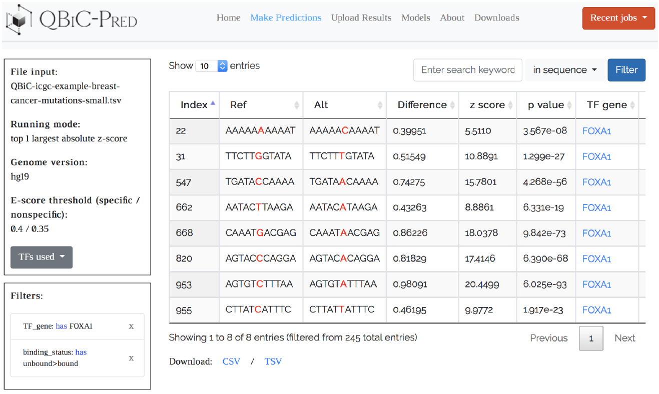
In vitro measurements of TF binding changes due to single nucleotide variants

The PBM technology can be used, with custom-designed DNA libraries, to directly measure the *in vitro* effects of single nucleotide variants on TF binding. To build custom DNA libraries we first selected, at random, DNA sequences containing binding sites for the TFs of interest, and then we introduced all possible single nucleotide variants in the binding site and the immediate flanking regions. Next, we measured the TF binding intensity for all the sequences, and we computed the log ratio of the binding signal between each mutant and the corresponding wild-type sequence to denote the TF binding change due to each variant.

We designed two such DNA libraries and used them to perform custom PBM experiments for six TFs. The DNA library for CREB1, RUNX1, and STAT3 included all single nucleotide variants in the TF binding site (10-12 bp), while the library for ETS1, ELK1, and GATA1 included all single nucleotide variants in the TF binding site and the flanking regions (36 bp). Because several TFs were tested against each DNA library, for each TF we obtain binding data both for variants in their specific binding sites, as well as variants in non-specific regions (which were present in the DNA library because they are specific to other TFs). We used all measurements to evaluate the accuracy of our predictions of TF binding changes (see Results).

In vivo allele-specific binding data

Allele-specific measurements of TF binding from *in vivo* ChIP-seq data have contributed to the identification of genetic variants that have the potential to change TF binding in the cell (16, 17). After mapping ChIP-seq reads to each allele of heterozygous variants, allele-specific binding (ASB) events can be identified as the ones with significantly different read counts between the alleles. Here, we used 32,252 ASB events and 79,827 non-ASB events across 81 TFs, as reported in (16), to compare the performance of our OLS-based models versus existing models of TF binding specificity (see Results).



File input:
QBiC-icgc-example-breast-cancer-mutations-small.tsv

Running mode:
top 1 largest absolute z-score

Genome version:
hg19

E-score threshold (specific / nonspecific):
0.4 / 0.35

Filters:

TF_gene: has FOXA1 x

binding_status: has unbound>bound x

Show 10 entries

Index	Ref	Alt	Difference	z score	p value	TF gene
22	AAAAAAAAAAT	AAAAACAAAAT	0.39951	5.5110	3.567e-08	FOXA1
31	TTCTTGGTATA	TTCTTTGTATA	0.51549	10.8891	1.299e-27	FOXA1
547	TGATACCAAAA	TGATAACAAAA	0.74275	15.7801	4.268e-56	FOXA1
662	AATACTTAAGA	AATACATAAGA	0.43263	8.8861	6.331e-19	FOXA1
668	CAAATGACGAG	CAAATAACGAG	0.86226	18.0378	9.842e-73	FOXA1
820	AGTACCCAGGA	AGTACACAGGA	0.81829	17.4146	6.390e-68	FOXA1
953	AGTGTCTTTAA	AGTGTATTTAA	0.98091	20.4499	6.025e-93	FOXA1
955	CTTATCATTTC	CTTATTATTTC	0.46195	9.9772	1.917e-23	FOXA1

Showing 1 to 8 of 8 entries (filtered from 245 total entries)

Download: CSV / TSV

Figure 2. Web server results page for a sample mutation file containing ICGC breast cancer mutation data (the example use case 'ICGC Breast Cancer Mutations - Small', available in QBiC). Output results were filtered to include only the FOXA1 transcription factor, and only mutations that create TF binding sites, i.e. 'unbound>bound' mutations.

QBiC-Pred web server implementation

QBiC-Pred was developed using the Flask web framework and it runs under Apache 2.4. Predictions of the effects of input variants on TF binding are made using pre-computed 12-mer tables encoding the predicted TF binding changes, z-scores and p-values for all possible mutations in all possible contexts (please see the QBiC About page for details). To further speed up the computations, QBiC uses asynchronous multiprocessing with the Celery framework, where 4 workers (i.e processes) are spawned for each request. Each worker extracts predictions for a subset of the input TFs. The prediction results are saved in a Redis database for two days; during this time the user can access the results using a unique job identifier, and can interactively process the results within QBiC (Figure 2). Users can also download the prediction results and re-upload them later, even after the job expired, for further processing within the QBiC framework.

Users can leave the QBiC website while the predictions are being calculated, and return to the job later using the link provided in the 'Recent Jobs' dropdown menu. Importantly, the time needed to execute a prediction job depends mostly on the number of TFs selected as input, as QBiC needs to read into memory the 12-mer table corresponding to each TF. Adding more variants to the input mutation/variant file will have an almost negligible impact on the processing time. After all predictions are computed, they are displayed in a table format with filtering capabilities. Users can post-process the results and downloaded them as csv or tsv files.

RESULTS

In previous work we showed that our OLS model-based predictions of TF binding changes due to DNA mutations correlate well with measured changes in gene expression (2). We also analyzed a large set of pathogenic non-coding variants, showing that these variants lead to more significant differences in TF binding between alleles, compared to common variants, which indicates that there is a strong regulatory component to pathogenic non-coding variants (2). Here, we complement our previous evaluations of the OLS models by assessing their accuracy in predicting *in vitro* and *in vivo* TF binding changes, and by comparing our OLS models to PWMs and deep learning models of TF binding specificity.

OLS models of TF binding specificity outperform PWMs and DeepBind models in predicting *in vitro* TF binding changes

As described in Materials and Methods, we designed custom DNA libraries for PBM experiments to test the effects of all single nucleotide variants within binding sites of six human TFs. We used the log ratio of the binding intensity between a mutant and its corresponding wild-type site to represent the TF binding change. Next, we made predictions of these

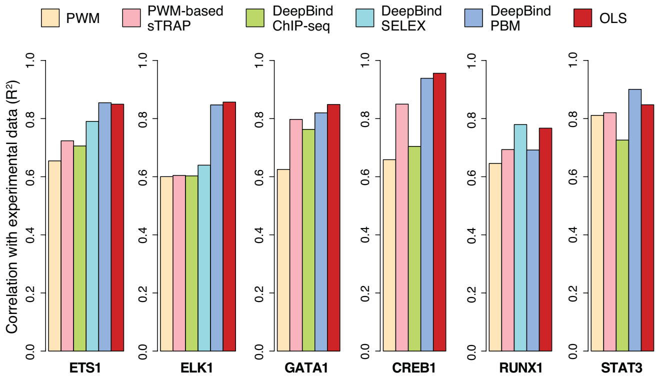


Figure 3. Performance of OLS models in predicting *in vitro* TF binding changes, compared to PWM and DeepBind models. When multiple PWM models are available for a TF, we choose the one that gives the best prediction result. We note that DeepBind ChIP-seq models are not available for RUNX1, and DeepBind SELEX models are not available for GATA1, CREB1, and STAT3. The *in vitro* binding data used in this analysis is available in Supplementary Table 1.

binding changes using six types of models: OLS models, PWM models used in (16), PWM-based sTRAP models (18), and DeepBind models (12) trained on *in vivo* ChIP-seq data, *in vitro* HT-SELEX data, and *in vitro* uPBM data. The uPBM data sets used to train DeepBind and OLS models were the same. The PWMs were obtained from the JASPAR (11) and HOCOMOCO (19) databases. For TFs with multiple PWMs available, the results we report below are for the PWM that performed best in our evaluation (ETS1: HOCOMOCO ETS1_HUMAN.H11MO.0.A, ELK1: HOCOMOCO ELK1_HUMAN.H11MO.0.B, GATA1: JASPAR MA0035.2, CREB1: JASPAR MA0018.2, RUNX1: JASPAR MA0002.2, STAT3: HOCOMOCO STAT3_HUMAN.H11MO.0.A). For DeepBind ChIP-seq and SELEX models, we used the v0.11 tools made available for download by the authors (12). For DeepBind PBM models, the authors kindly provided assistance training the models on our uPBM data.

OLS models can directly predict the TF binding change due to a variant in a fixed-length or variable-length sequence. In contrast, for PWM and DeepBind models we computed likelihood scores for the wild-type and mutant sequences, based on fixed-length window scores. For these models, we predicted the binding change as the difference between the maximum of all wild-type window scores and the maximum of all mutant window scores. This definition is the same as delta track metric defined in Wagih et al. (16), which performed best in their study.

The correlations between model predictions and the TF binding changes measured using custom PBM experiments across the six TFs are shown in Figure 3. Except for RUNX1, for which the DeepBind SELEX model was slightly better than the rest of the models, DeepBind PBM models and our OLS models outperformed the other models in predicting TF binding changes *in vitro*. Compared to DeepBind PBM models, our OLS models are simpler and much faster for training and for predictions. In addition, OLS models can be used to assess the statistical significance of the TF binding changes predicted for each variant.

Figure 4 shows a detailed comparison of five models (OLS, PWM, sTRAP, DeepBind SELEX, and DeepBind PBM) for a binding site of TF ELK1. The input mutation file used in QBiC to generate the ELK1 binding change predictions shown in Figure 4 is available as Supplementary Table 2, and can also be downloaded from the QBiC website as the sample input file in sequence format.

The cross-validation accuracy of OLS models correlates with their accuracy in predicting *in vitro* TF binding changes

A TF can have multiple PWM models and DeepBind models available, and it is often difficult to choose which model to use for prediction. In contrast, for our OLS-based approach, we are able to rank the models based on cross-validation accuracy on the uPBM training data set. As expected, we found that there is a positive relationship between the in-sample cross-validation accuracy and the TF binding change prediction accuracy on independent *in vitro* data (Figure 5). Thus, when a TF has multiple OLS models, we recommend choosing the model with the highest cross-validation accuracy. Detailed information on the available OLS models for each human TF can be found in the About section of the QBiC website.

OLS models of TF binding specificity outperform PWMs and DeepBind models in predicting *in vivo* allele-specific binding variants

To test the performance of OLS models on *in vivo* data, we used the allele-specific binding (ASB) and non-ASB variants in (16). We compared the performance of OLS models, PWM models, and DeepBind models in distinguishing ASB variants from non-ASB variants. The performance of each model was assessed using the area under the Receiver Operating Characteristic curve (AUROC) measure. For PWMs and DeepBind ChIP-seq models, we used the binding change scores reported by Wagih et al. (16). For DeepBind SELEX and PBM models we derived the binding change scores similarly to Wagih et al. (16), and used them for the classification. For OLS models we used the z-score outputs to classify the variants. The DeepBind PBM and OLS models were trained on the same sets of PBM data. To illustrate how QBiC can be used to analyze ASB and non-ASB variants, in Supplementary Table 3 we provide the input mutation file corresponding to the ASB data for TF MAFK, in VCF format. This file is also available on the QBiC website, as the sample input file for the VCF format.

A total of 14 human TFs have PWM models, OLS models, and DeepBind models available. For these TFs we divided their ASB variants into gain-of-binding and loss-of-binding variants (for which the TF binding changes have opposite signs), and for each set we used the different TF binding models to distinguish between ASB and non-ASB variants. OLS models clearly outperformed PWMs (Figure 6a), which was expected given the limitations of PWM models in capturing TF binding specificity (7, 20, 21, 22). OLS models also outperformed DeepBind SELEX models trained on *in vitro* binding data from HT-SELEX experiments (Figure 6b) and DeepBind PBM models trained on *in vitro* data from PBM experiments (Figure 6c) demonstrating that, when using only DNA sequence information for training, OLS models perform best in predicting *in vivo* allele-specific binding variants.

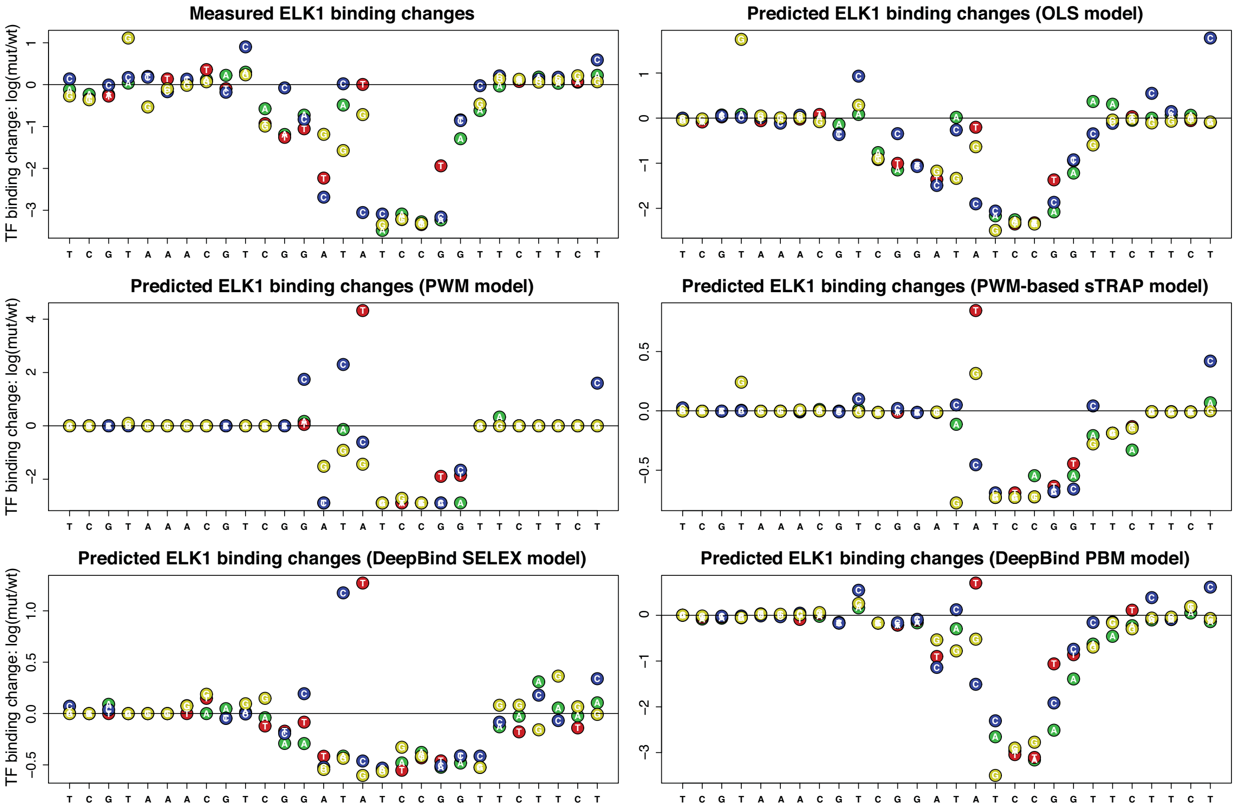


Figure 4. Measured and predicted effects of single nucleotide mutations in an ELK1 binding site and its flanking regions. Since the wild-type sequence contains an ELK1 binding site, most of the variants decrease binding. The A to T mutation in the middle generates a perfect match to core ELK1 motif TTCC. This, however, does not increase the binding signal, likely because the flanking regions already made the ATCC site very strong. Both the PWM and DeepBind models incorrectly predict a dramatic increase in binding due to the A to T mutation. The OLS model, however, correctly predicts the TF binding to be nearly unchanged. There are also positions where the magnitude of the TF binding change seems to be overestimated by our OLS model but not so much by PWM-based and DeepBind models, such as the T to C mutation at the last position. We note, however, that in this case the correctness of the magnitude of the predicted increase is difficult to assess. For the PWM and the DeepBind SELEX models, the largest predicted increases are incorrect, so we cannot compare them directly to predicted increase at the last position. For the PWM-based sTRAP model and the DeepBind PBM model, the magnitude of the predicted increase at the last position is larger than for other correctly-predicted increases, similarly to our OLS model. Thus, it is difficult to judge which model performed best at predicting this particular increase. Nevertheless, over all mutations tested, the OLS model performs best (see also Figure 3).

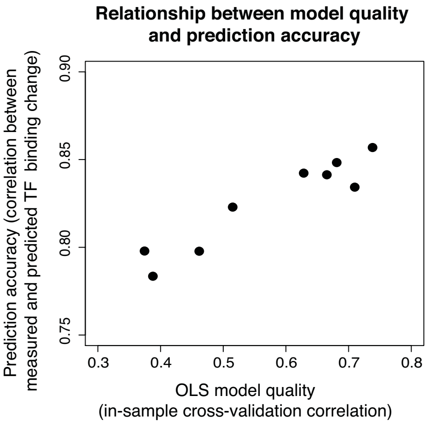


Figure 5. Relationship between OLS model quality (assessed as the in-sample cross-validation correlation) and the prediction accuracy on independent *in vitro* mutation data. Figure shows the performance of OLS models trained on 9 different uPBM data sets for TF ELK1.

We also compared the performance of OLS models to DeepBind models trained on *in vivo* ChIP-seq data (Figure 6d). Using OLS models we obtained larger AUROC values for about half of the TFs, and overall the two models had similar power in distinguishing ASB from non-ASB variants. Nevertheless, we note that the DeepBind ChIP-seq models were trained on ChIP-seq data from the same cell type as the ChIP-seq data from which the ASB variants were called. Therefore, OLS models managed to reach similar performance to models trained on the ChIP-seq data itself, despite the fact that OLS models do not use any cell type specific information.

DISCUSSION

Quantitative predictions of TF binding changes can help us understand the functional roles of genetic variants, and prioritize variants that are likely to have regulatory effects. QBiC-Pred provides a fast and accurate approach to predict TF binding changes due to genetic variants, based solely on their sequence context. QBiC-Pred models are trained on *in vitro* high-throughput universal PBM data, and they outperform current PWM-based models and DeepBind models, which are also based mainly on DNA sequence information. In addition, QBiC-Pred offers a way to statistically test the significance of each variant, taking the quality of the predictive models into account. The quality measure of the models also helps circumvent the problem of deciding which model to use when

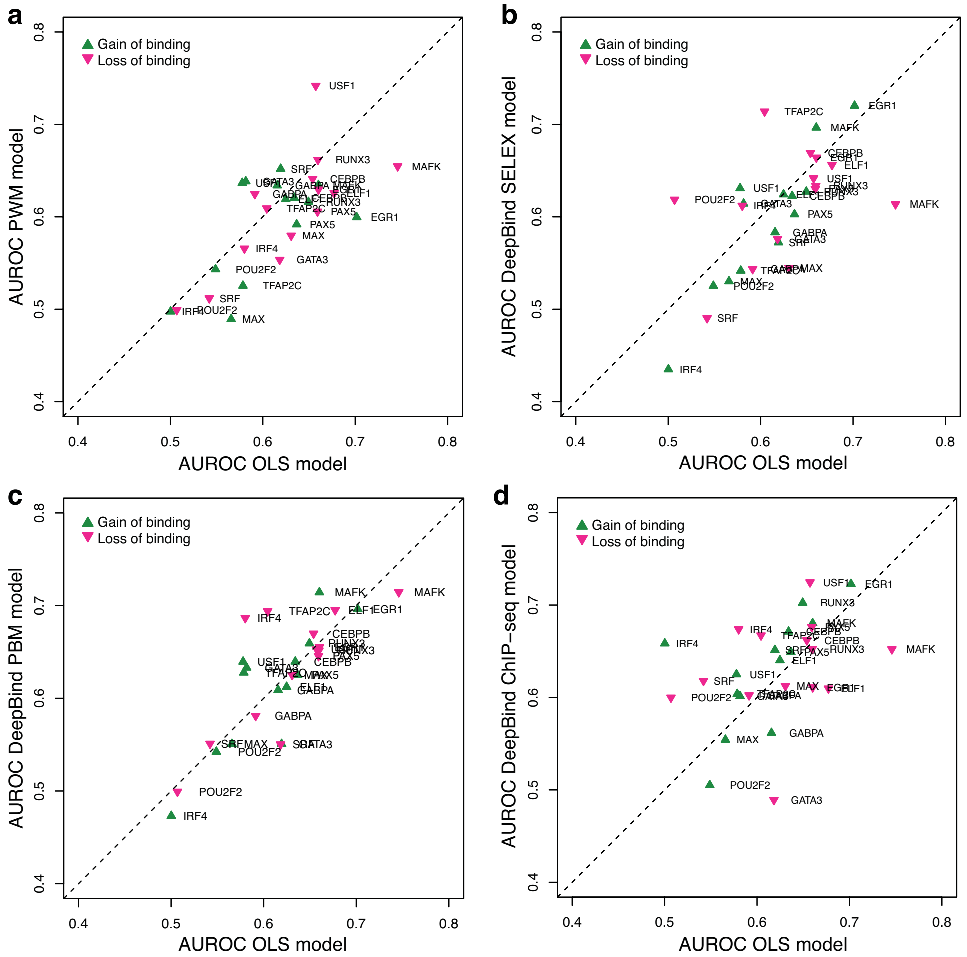


Figure 6. Performance of OLS, DeepBind, and PWM models in distinguishing between ASB and non-ASB variants identified from *in vivo* ChIP-seq data.

multiple models are available, which is often encountered when making predictions using PWMs.

Several recent methods, including Sasquatch (23), DeepSEA (24), and deltaSVM (25), predict the impact of non-coding variants by taking advantage of cell- and tissue-specific information, oftentimes beyond TF binding data. These methods are complementary to ours: they focus on overall functional changes caused by non-coding variants, while we examine more specifically the potential binding changes for each individual TF. For example, Sasquatch predicts the change in the DNase footprint due to a variant, but does not directly pinpoint the binding of which TF(s) is affected by the variant (unless one post-processes the results using specific TF binding models). In contrast, QBiC-Pred can make quantitative predictions in a TF-specific manner, for a large number of TFs, although it cannot predict the effect of the variant in any specific cell type. Using these methods together would give us a better understanding of the functional impact of non-coding variants in the cell.

Annotation-based methods such as rVarBase (26), INFERNO (27), HaploReg (28), and RegulomeDB (14) can also be used to investigate potential regulatory variants. These methods test whether the input variants fall within known regulatory regions annotated, for example, using PWM models and cell type-specific data. Thus, predictions made by annotation-based methods depend on the quality of the existing annotations, and, in the case of TF binding sites, these methods are unlikely to detect variants that lead to the creation of new binding sites in the genome. In addition, we note that none of the methods mentioned above provides a direct measure of the confidence in the predicted changes in TF binding, based on the quality of the binding data and model, which is a distinguishing feature of QBiC-Pred.

In summary, QBiC-Pred uses OLS models of TF-DNA binding specificity to make accurate predictions of TF binding changes due to single nucleotide variants. In addition to the current functionalities of QBiC-Pred, a natural extension would be to allow input sequences containing multiple variants. As shown in our previous work, OLS models perform very well on data containing multiple variants, being able to predict $\sim 50\%$ of the resulting variation in gene expression (2). Another extension would be to include models trained on other types of high-throughput *in vitro* TF binding data, such as HT-SELEX data (29, 30). This would extend the list of human TFs that can be analyzed using QBiC-Pred beyond the 577 TFs with available high-quality uPBM data. This extension, however, will require the development of new methodology that takes into account the statistical properties of the HT-SELEX data, in order to allow us to use the data directly to compute significance levels (p-values) reflecting our confidence in the predicted effects of mutations on TF binding.

ACKNOWLEDGEMENTS

The authors would like to thank Brendan Frey, Leo J. Lee, and Alice Gao (University of Toronto) for help with training DeepBind PBM models on the same data sets used for our OLS models. This research was supported in part by the National Institutes of Health (grant R01-GM117106 to RG) and the National Science Foundation (grant MCB-1715589 to RG). High-performance computing was partially supported by the Duke Center for Genomic and Computational Biology.

DATA AVAILABILITY

The raw TF-binding data used in our analyses of the *in vitro* effects of single nucleotide variants on the binding specificities of six human TFs is being deposited to GEO and will be freely available upon publication. The processed data is available in Supplementary Table 1.

Conflict of interest statement.

None declared.

REFERENCES

1. Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. (Feb, 2016) Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.,* **17**(2), 93–108.
2. Zhao, J., Li, D., Seo, J., Allen, A. S., and Gordan, R. (May, 2017) Quantifying the Impact of Non-coding Variants on Transcription Factor-DNA Binding. *Res Comput Mol Biol,* **10229**, 336–352.
3. Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (Nov, 2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.,* **24**(11), 1429–1435.
4. Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., et al. (Apr, 2010) International network of cancer genome projects. *Nature,* **464**(7291), 993–998.
5. Berger, M. F. and Bulyk, M. L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc,* **4**(3), 393–411.
6. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., and Bruford, E. A. (Jan, 2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.,* **43**(Database issue), D1079–1085.
7. Shen, N., Zhao, J., Schipper, J. L., Zhang, Y., Bepler, T., Leehr, D., Bradley, J., Horton, J., Lapp, H., and Gordan, R. (Apr, 2018) Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. *Cell Syst,* **6**(4), 470–483.
8. Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., and Bulyk, M. L. (Jan, 2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.,* **43**(Database issue), D117–122.
9. Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (Jan, 2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.,* **34**(Database issue), D108–110.
10. Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., et al. (Sep, 2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell,* **158**(6), 1431–1443.
11. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S. R., Tan, G., et al. (Jan, 2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.,* **46**(D1), D260–D266.
12. Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (Aug, 2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.,* **33**(8), 831–838.
13. Ward, L. D. and Kellis, M. (Jan, 2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.,* **44**(D1), D877–881.
14. Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., and Snyder, M. (Sep, 2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.,* **22**(9), 1790–1797.
15. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (02, 2018) The Human Transcription Factors. *Cell,* **172**(4), 650–665.
16. Wagih, O., Merico, D., Delong, A., and Frey, B. J. (2018) Allele-specific transcription factor binding as a benchmark for assessing variant impact predictors. *bioRxiv,* [bioRxiv:253427v1].
17. Shi, W., Fornes, O., Mathelier, A., and Wasserman, W. W. (Dec, 2016) Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.,* **44**(21), 10106–10116.
18. Thomas-Chollier, M., Hufton, A., Heinig, M., O'Keeffe, S., Masri, N. E., Roider, H. G., Manke, T., and Vingron, M. (Nov, 2011) Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat Protoc,* **6**(12), 1860–1869.
19. Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., Kolpakov, F. A., and Makeev, V. J. (Jan, 2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.,* **46**(D1), D252–D259.
20. Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S.,

Bussemaker, H. J., Gordan, R., and Rohs, R. (Apr, 2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.,* **112**(15), 4654–4659.

21. Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordan, R., and Rohs, R. (Sep, 2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.,* **39**(9), 381–399.
22. Siggers, T. and Gordan, R. (Feb, 2014) Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.,* **42**(4), 2099–2111.
23. Schwessinger, R., Suciu, M. C., McGowan, S. J., Telenius, J., Taylor, S., Higgs, D. R., and Hughes, J. R. (10, 2017) Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. *Genome Res.,* **27**(10), 1730–1742.
24. Zhou, J. and Troyanskaya, O. G. (Oct, 2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods,* **12**(10), 931–934.
25. Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., and Beer, M. A. (Aug, 2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.,* **47**(8), 955–961.
26. Guo, L., Du, Y., Qu, S., and Wang, J. (Jan, 2016) rVarBase: an updated database for regulatory features of human variants. *Nucleic Acids Res.,* **44**(D1), D888–893.
27. Amlie-Wolf, A., Tang, M., Mlynarski, E. E., Kuksa, P. P., Valladares, O., Katanic, Z., Tsuang, D., Brown, C. D., Schellenberg, G. D., and Wang, L. S. (Sep, 2018) INFERNO: inferring the molecular mechanisms of noncoding genetic variants. *Nucleic Acids Res.,* **46**(17), 8740–8753.
28. Ward, L. D. and Kellis, M. (Jan, 2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.,* **40**(Database issue), D930–934.
29. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpaa, M. J., et al. (Jun, 2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.,* **20**(6), 861–873.
30. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (Jan, 2013) DNA-binding specificities of human transcription factors. *Cell,* **152**(1-2), 327–339.