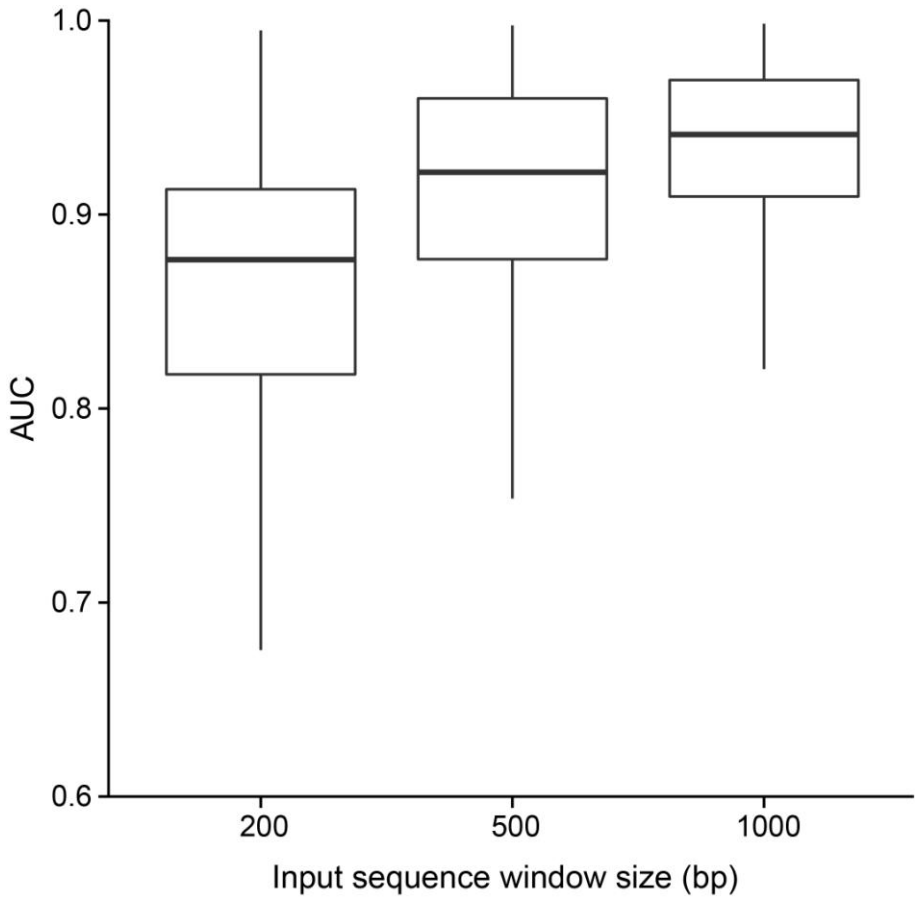
Nat. Methods doi:10.1038/nmeth.3547 (24 August 2015)

Predicting effects of noncoding variants with deep learning–based sequence model

Jian Zhou & Olga G Troyanskaya

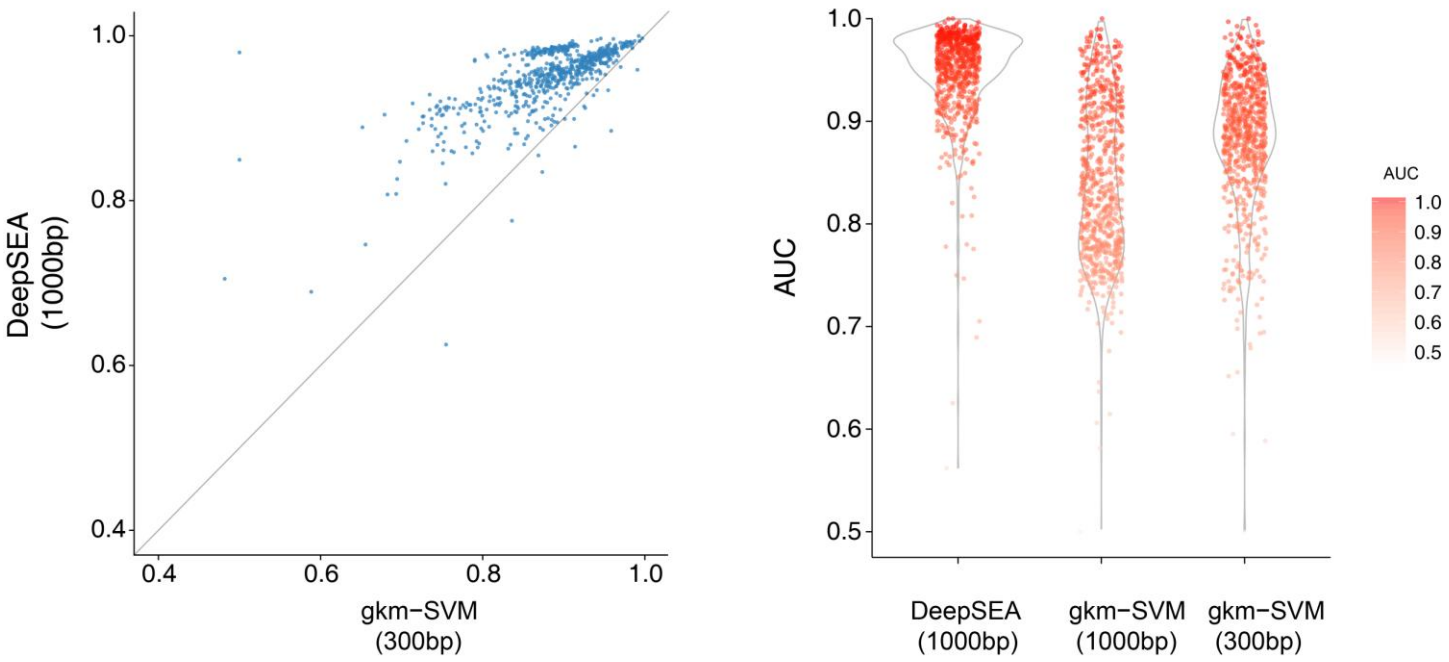
In the version of this supplementary file originally posted online, the Supplementary Note was missing. This section is now provided as of 28 August 2015.



Supplementary Figure 1

Performance comparison of DeepSEA models trained with different context sequence lengths

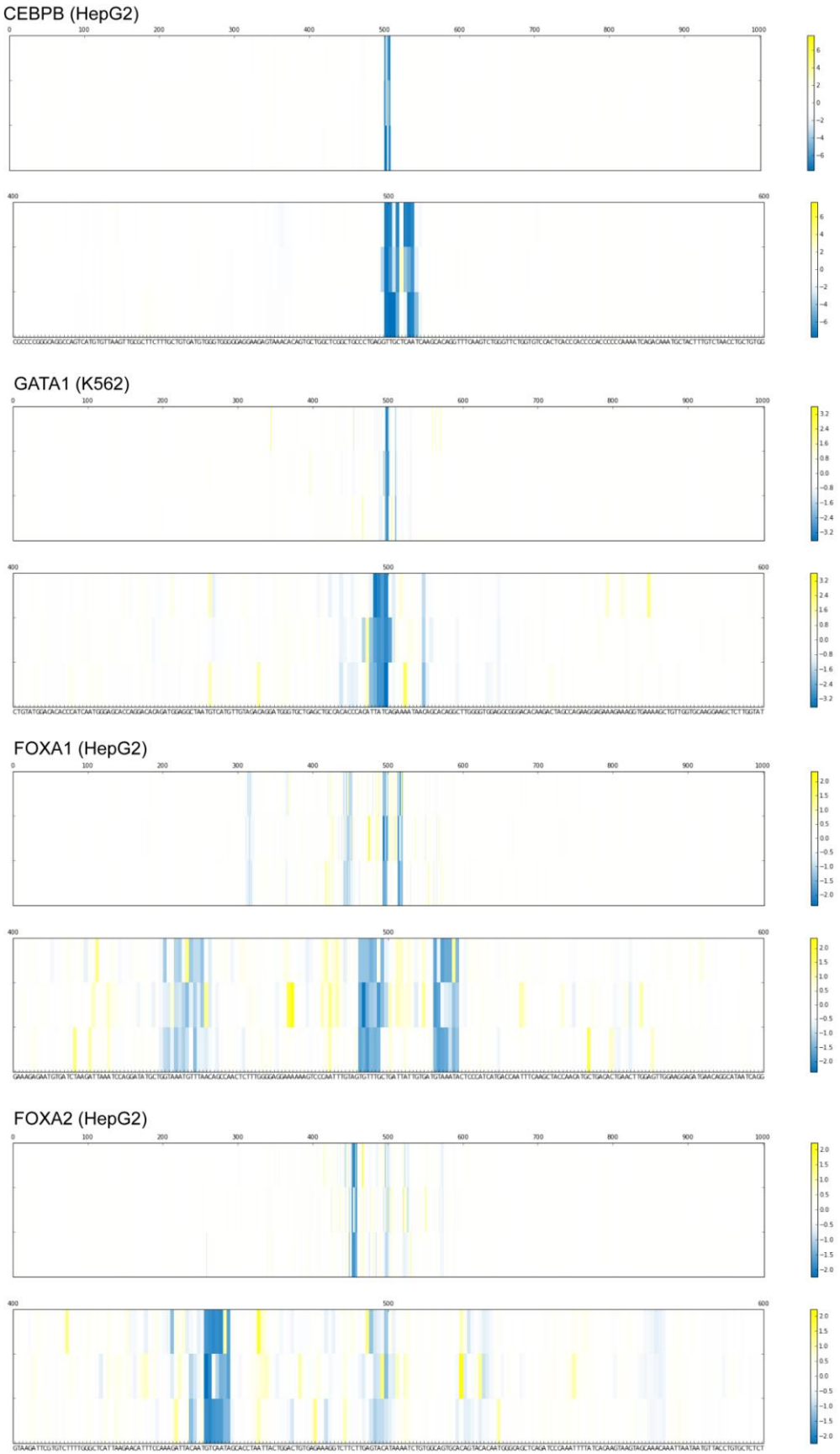
DeepSEA models with the same architecture as described in the Online Methods were trained on 200bp, 500bp, and 1000bp input sequences respectively, and the AUCs of all chromatin features were shown with box plots. While the chromatin feature labels were always determined from the central 200bp regions, increasing context sequence length significantly improved model performance (P-value < 2.2e-16 by Wilcoxon signed rank test between any pair of models).



Supplementary Figure 2

Performance comparison with gkm-SVM

Deep convolutional network model outperformed gapped *k*-mer SVM (gkm-SVM) on transcription factor binding prediction. Deep convolutional network achieved higher area under receiver operating characteristic (AUC) for almost all transcription factors (left panel). Gapped *k*-mer SVM did not gain performance from increasing size of context sequences (right panel).

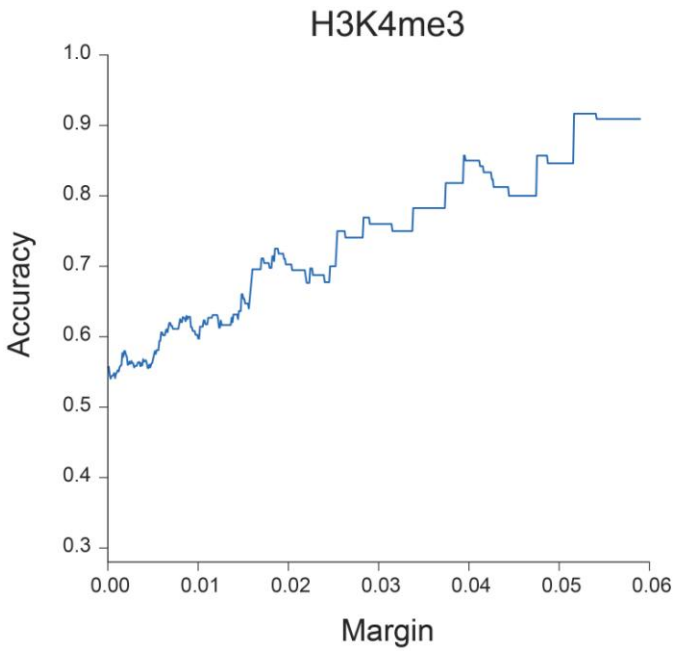


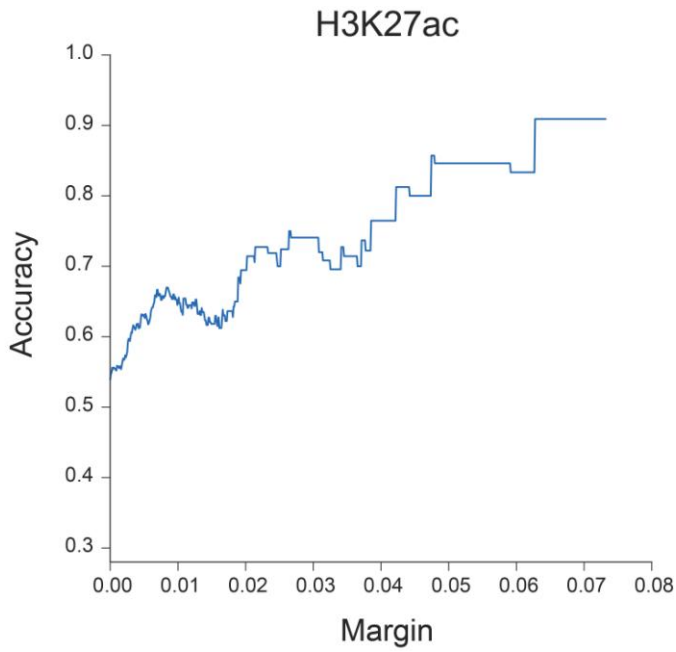
Nature Methods: doi:10.1038/nmeth.3547

Supplementary Figure 3

***In silico* saturated mutagenesis analysis for identifying predictive sequence features**

Predictive sequence features can be identified by analyzing effect on binding probability by computationally mutating each base. Each column in a heatmap represents a base position in the sequence. The three rows represent the three possible base substitutions following A>G>C>T order from bottom to top. For example, if the original sequence has base G, then the three rows represent C, T, A from bottom to top. The log2 fold change of odds (odds are computed from probability as $P/(1-P)$) are shown with the heatmap; yellow indicates increase of binding and blue indicates decrease of binding. Each sequence example is shown by two panels. The first (top) panel shows the 'mutation scanning' results on the whole 1000bp sequence. The second (bottom) panel focuses on the center 200bp in order to show the actual nucleotide sequences. Many sequence elements identified are consistent with canonical motifs such as TTGCTCAA for CEBPB, TGATAA for GATA1, GTAAATA for FOXA1 and GTACATA for FOXA2. The four example sequences shown in this figure are centered around SNPs chr1:109817590 G>T, chr16:209709 T>C, chr10:23508363 A>G, chr16:52599188 C>T respectively.

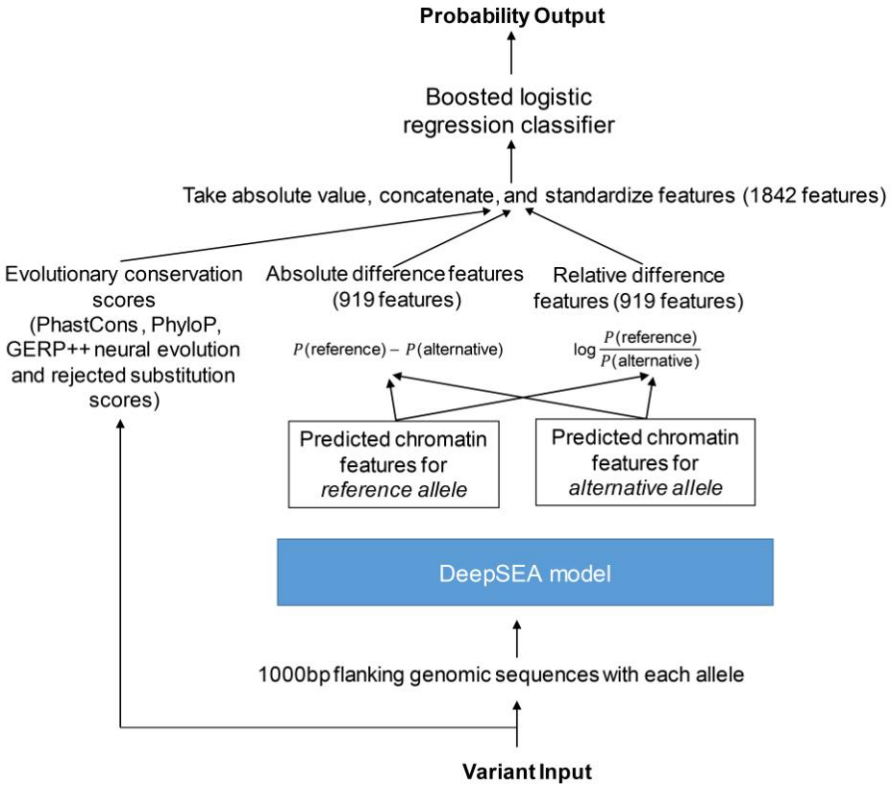




Supplementary Figure 4

DeepSEA accurately predicted histone QTL effects

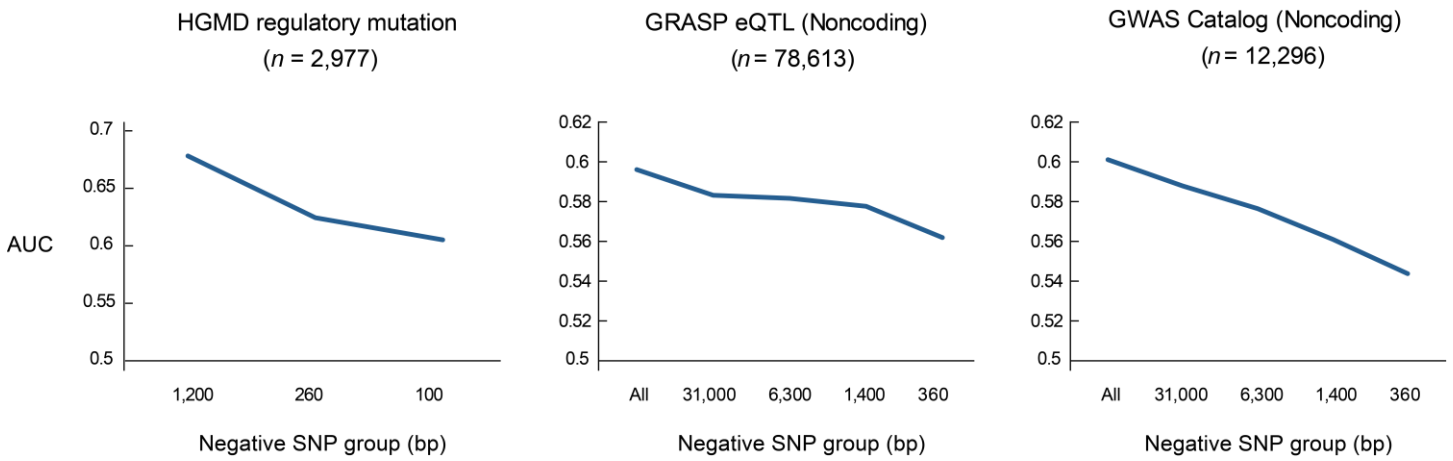
DeepSEA histone mark classifiers provided accurate prediction of allele specific effects on histone marks H3K4me3 and H3K27ac (the allele with more histone mark). The top prediction accuracies are over 0.9 for both marks. The predictions were evaluated with histone mark QTLs identified with FDR < 0.1 in Yoruba lymphoblastoid cell lines[1]. Margin shown on the *x* axis is the threshold of predicted probability differences between the two alleles for classifying high-confidence predictions. Performance is measured by accuracy of the above threshold predictions (*y* axis). 1.McVicker, G. et al. *Science* **342**, 747-749 (2013).



Supplementary Figure 5

Flow diagram for DeepSEA functional SNP prioritization

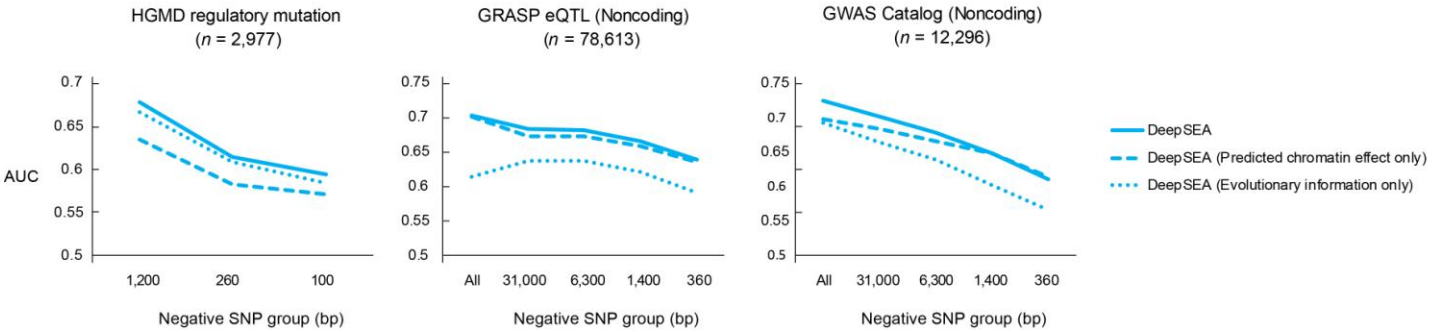
For each input variant, DeepSEA computes 1842 features, including 1838 predicted chromatin effect features and 4 evolutionary conservation features. Predicted chromatin effect features include absolute difference and relative difference computed based on predicted probability of reference and alternative sequences, for each TF / DNase / Histone chromatin feature. Evolutionary conservation scores based on multi-species genome alignments were retrieved for the variant positions. Each feature is taken the absolute value, and is then scaled to mean 0 and variance 1 before providing as input to classifier.



Supplementary Figure 6

DeepSEA functional significance score prioritizes functional noncoding variants with high performance

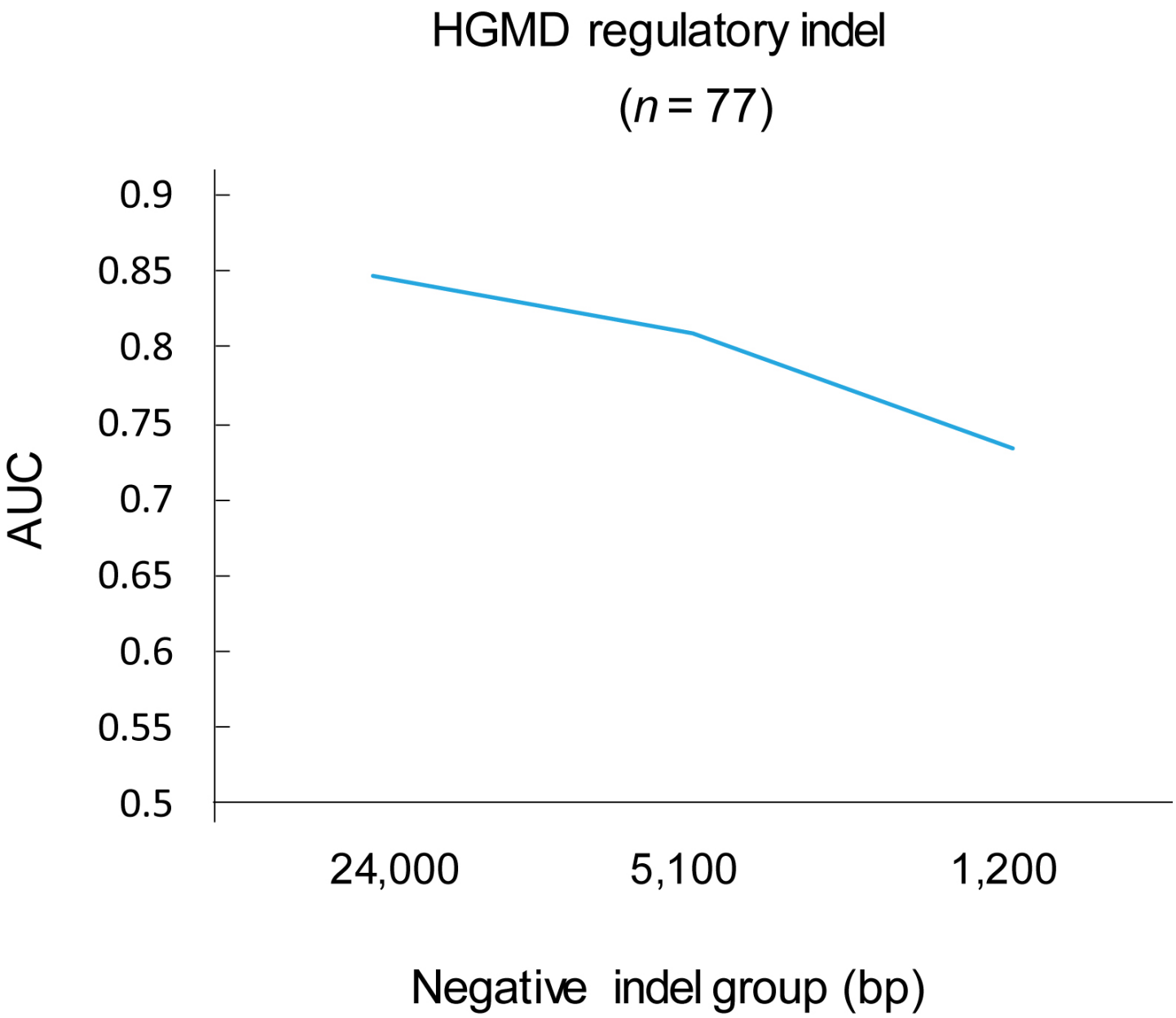
DeepSEA functional significance score measures the overall significance of predicted chromatin effects and evolutionary conservation scores, and it is unsupervised thus unbiased to any training functional variant annotation set (see Online Methods). Notably DeepSEA functional significance score still surpassed the performance of previous methods even though no supervised training was used (compare to Fig. 3). The performance was measured by area under receiver operating characteristic (AUC). *x* axis shows the average distances of negative-variant groups to a nearest positive variant. The 'all' negative-variant groups are randomly selected negative 1000 Genomes SNPs.



Supplementary Figure 7

Dissecting DeepSEA functional SNP prioritization performance with subsets of input features

DeepSEA functional SNP prioritization models performance on HGMD regulatory mutations, noncoding eQTLs, and noncoding trait-associated (GWAS) SNPs was analyzed by comparing with models trained with only predicted chromatin effect features or only evolutionary conservation features. The performance was measured by area under receiver operating characteristic (AUC). *x* axis shows the average distances of negative-variant groups to a nearest positive variant. The 'all' negative-variant groups are randomly selected negative 1000 Genomes SNPs.



Supplementary Figure 8

DeepSEA-based classifier prioritized functionally annotated indels with high performance

HGMD regulatory indels prioritization performance was evaluated against negative 1000 Genomes indel groups with different distances to positive indels (average distance shown on the x-axis). The performance was measured by area under receiver operating characteristic (AUC). The prioritization model was trained with HGMD regulatory single nucleotide substitution mutations against 1200bp average distance negative variants.

Supplementary Note. DeepSEA model configuration

Model Architecture:

1. Convolution layer (320 kernels. Window size: 8. Step size: 1.)
2. Pooling layer (Window size: 4. Step size: 4.)
3. Convolution layer (480 kernels. Window size: 8. Step size: 1.)
4. Pooling layer (Window size: 4. Step size: 4.)
5. Convolution layer (960 kernels. Window size: 8. Step size: 1.)
6. Fully connected layer (925 neurons)
7. Sigmoid output layer

Regularization Parameters:

Dropout proportion (proportion of outputs randomly set to 0):

- Layer 2: 20%
- Layer 4: 20%
- Layer 5: 50%
- All other layers: 0%

L2 regularization (λ_1): 5e-07
L1 sparsity (λ_2): 1e-08
Max kernel norm (λ_3): 0.9