Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi[1,2,6], Andrew Delong[1,6], Matthew T Weirauch[3–5] & Brendan J Frey[1–3]

Knowing the sequence specificities of DNA- and RNA-binding proteins is essential for developing models of the regulatory processes in biological systems and for identifying causal disease variants. Here we show that sequence specificities can be ascertained from experimental data with 'deep learning' techniques, which offer a scalable, flexible and unified computational approach for pattern discovery. Using a diverse array of experimental data and evaluation metrics, we find that deep learning outperforms other state-of-the-art methods, even when training on *in vitro* data and testing on *in vivo* data. We call this approach DeepBind and have built a stand-alone software tool that is fully automatic and handles millions of sequences per experiment. Specificities determined by DeepBind are readily visualized as a weighted ensemble of position weight matrices or as a 'mutation map' that indicates how variations affect binding within a specific sequence.

DNA- and RNA-binding proteins play a central role in gene regulation, including transcription and alternative splicing. The sequence specificities of a protein are most commonly characterized using position weight matrices[1] (PWMs), which are easy to interpret and can be scanned over a genomic sequence to detect potential binding sites. However, growing evidence indicates that sequence specificities can be more accurately captured by more complex techniques[2–5]. Recently, 'deep learning' has achieved record-breaking performance in a variety of information technology applications[6,7]. We adapted deep learning methods to the task of predicting sequence specificities and found that they compete favorably with the state of the art. Our approach, called DeepBind, is based on deep convolutional neural networks and can discover new patterns even when the locations of patterns within sequences are unknown—a task for which traditional neural networks require an exorbitant amount of training data.

There are several challenging aspects in learning models of sequence specificity using modern high-throughput technologies. First, the data come in qualitatively different forms. Protein binding microarrays (PBMs)[8] and RNAcompete assays[9] provide a specificity coefficient for each probe sequence, whereas chromatin immunoprecipitation (ChIP)-seq[10] provides a ranked list of putatively bound sequences of varying length, and HT-SELEX[11] generates a set of very high affinity sequences. Second, the quantity of data is large. A typical high-throughput experiment measures between 10,000 and 100,000 sequences, and it is computationally demanding to incorporate them all. Third, each data acquisition technology has its own artifacts, biases and limitations, and we must discover the pertinent specificities despite these unwanted effects. For example, ChIP-seq reads often localize to "hyper-ChIPable" regions of the genome near highly expressed genes[12].

DeepBind (**Fig. 1**) addresses the above challenges. (i) It can be applied to both microarray and sequencing data; (ii) it can learn from millions of sequences through parallel implementation on a graphics processing unit (GPU); (iii) it generalizes well across technologies, even without correcting for technology-specific biases; (iv) it can tolerate a moderate degree of noise and mislabeled training data; and (v) it can train predictive models fully automatically, alleviating the need for careful and time-consuming hand-tuning. Importantly, a trained model can be applied and visualized in ways that are familiar to users of PWMs. We explored two downstream applications: uncovering the regulatory role of RNA binding proteins (RBPs) in alternative splicing, and analyzing disease-associated genetic variants that can affect transcription factor binding and gene expression.

RESULTS

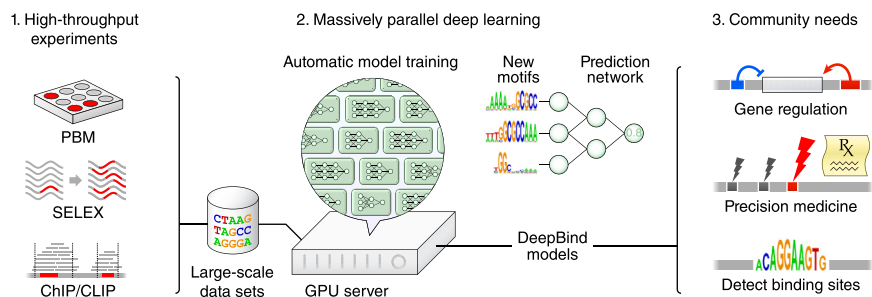### Training DeepBind and scoring sequences

For training, DeepBind uses a set of sequences and, for each sequence, an experimentally determined binding score. Sequences can have varying lengths (14–101 nt in our experiments), and binding scores can be real-valued measurements or binary class labels. For a sequence s, DeepBind computes a binding score $f(s)$ using four stages:

$$f(s) = \mathrm{net}_W(\mathrm{pool}(\mathrm{rect}_b(\mathrm{conv}_M(s))))$$

The convolution stage (conv_M) scans a set of motif detectors with parameters M across the sequence. Motif detector M_k is a $4 \times m$ matrix, much like a PWM of length m but without requiring coefficients to be probabilities or log odds ratios. The rectification stage isolates positions with a good pattern match by shifting the response of detector M_k by b_k and clamping all negative values to zero. The pooling stage computes the maximum and average of each motif detector's rectified response across the sequence; maximizing helps to identify the presence of longer motifs, whereas averaging helps to

Figure 1 DeepBind's input data, training procedure and applications. 1. The sequence specificities of DNA- and RNA-binding proteins can now be measured by several types of high-throughput assay, including PBM, SELEX, and ChIP- and CLIP-seq techniques. 2. DeepBind captures these binding specificities from raw sequence data by jointly discovering new sequence motifs along with rules for combining them into a predictive binding score. Graphics processing units (GPUs) are used to automatically train high-quality models, with expert tuning allowed but not required. 3. The resulting DeepBind models can then be used to identify binding sites in test sequences and to score the effects of novel mutations.

identify cumulative effects of short motifs, and the contribution of each is determined automatically by learning. These values are fed into a nonlinear neural network with weights W, which combines the responses to produce a score (**Fig. 2a**, **Supplementary Fig. 1** and **Supplementary Notes**, sec. 1.1).

We use deep learning techniques[13–16] to infer model parameters and to optimize algorithm settings. Our training pipeline (**Fig. 2b**) alleviates the need for hand tuning, by automatically adjusting many calibration parameters, such as the learning rate, the degree of momentum[14], the mini-batch size, the strength of parameter regularization, and the dropout probability[15].

To obtain the results reported below, we trained DeepBind models on a combined 12 terabases of sequence data, spanning thousands of public PBM, RNAcompete, ChIP-seq and HT-SELEX experiments. We provide the source code for DeepBind together with an online repository (http://tools.genes.toronto.edu/deepbind/) of 927 DeepBind models representing 538 distinct transcription factors and 194 distinct RBPs, each of which was trained on high-quality data and can be applied to score new sequences using an easily installed executable file with no hardware or software requirements.

Ascertaining DNA sequence specificities

To evaluate DeepBind's ability to characterize DNA-binding protein specificity, we used PBM data from the revised DREAM5 TF-DNA Motif Recognition Challenge by Weirauch et al.[17]. The PBM data represent 86 different mouse transcription factors, each measured using two independent array designs. Both designs contain ~40,000 probes that cover all possible 10-mers, and all nonpalindromic 8-mers, 32 times. Participating teams were asked to train on the probe intensities using one array design and to predict the intensities of the held-out array design, which was not made available to participants.

Weirauch et al.[17] evaluated 26 algorithms that can be trained on PBM measurements, including FeatureREDUCE[17], BEEML-PBM[18], MatrixREDUCE[19], RankMotif++[20] and Seed-and-Wobble[21]. For each individual algorithm, they optimized the data preprocessing steps to attain best test performance. Methods were evaluated using the Pearson correlation between the predicted and actual probe intensities, and values from the area under the receiver operating characteristic (ROC) curve (AUC) computed by setting high-intensity probes as positives and the remaining probes as negatives[17]. To the best of our knowledge, this is the largest independent evaluation of this type. When we tested DeepBind under the same conditions, it outperformed all 26 methods (**Fig. 3a**). DeepBind also ranked first among 15 teams when we submitted it to the online DREAM5 evaluation script (**Supplementary Table 1**).

To assess the ability of DeepBind models trained using *in vitro* PBM data to predict sequence specificities measured using *in vivo* ChIP-seq data, we followed the method described by Weirauch et al.[17]. Predicting transcription factor binding *in vivo* is more difficult because it is affected by other proteins, the chromatin state and the physical accessibility of the binding site. We found that DeepBind also achieves the highest score when applied to the *in vivo* ChIP-seq data (**Fig. 3b** and **Supplementary Fig. 2**). The best method reported in the original evaluation (Team_D, a *k*-mer-based model) and the best reported in the revised evaluation (FeatureREDUCE, a hybrid PWM/*k*-mer model) both had reasonable, but not the best, performance on *in vivo* data, which might be due to overfitting to PBM noise[17].
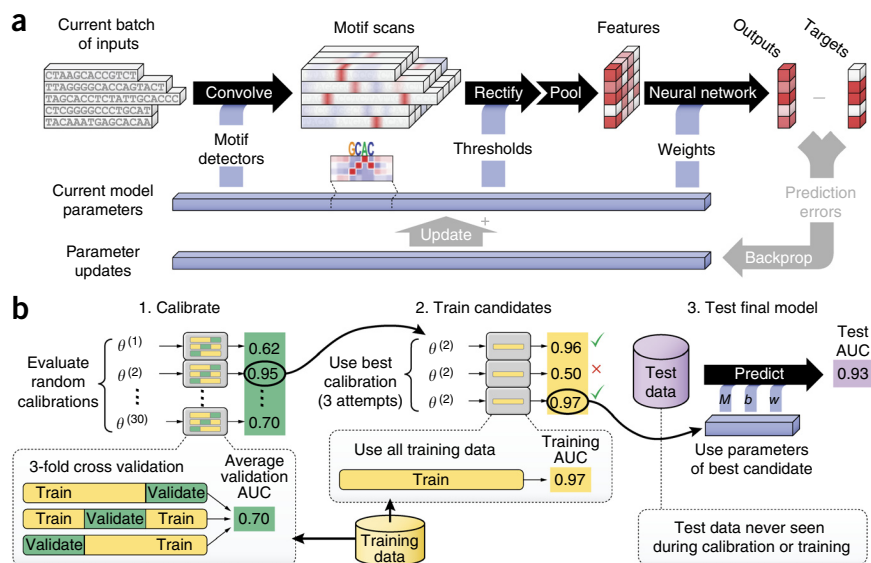
Figure 2 Details of inner workings of DeepBind and its training procedure. (**a**) Five independent sequences being processed in parallel by a single DeepBind model. The convolve, rectify, pool and neural network stages predict a separate score for each sequence using the current model parameters (**Supplementary Notes**, sec. 1). During the training phase, the backprop and update stages simultaneously update all motifs, thresholds and network weights of the model to improve prediction accuracy. (**b**) The calibration, training and testing procedure used throughout (**Supplementary Notes**, sec. 2).
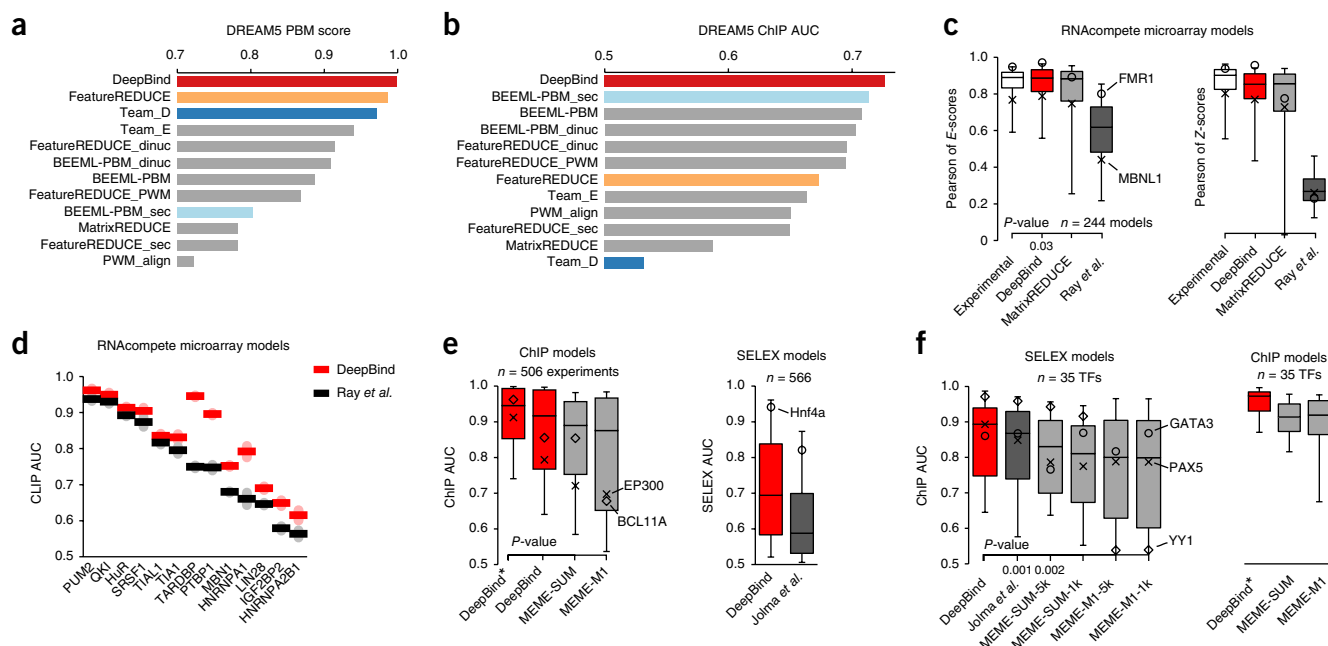
Figure 3 Quantitative performance on various types of held-out experimental test data. (**a**) Revised PBM evaluation scores for the DREAM5 *in vitro* transcription factor challenge. The DREAM5 PBM score is based on Pearson correlations and AUCs across 66 transcription factors (c.f. ref. 17, **Table 2**; **Supplementary Notes**, sec. 5.2). (**b**) DREAM5 *in vivo* transcription factor challenge ChIP AUC, using the *in vitro* models (c.f. ref. 17, Fig. 3; **Supplementary Notes**, sec. 5.3); only DeepBind ranks highly for both *in vitro* and *in vivo*. (**c**) RBP *in vitro* performance using RNAcompete data[22] (Wilcoxon two-sided signed-rank test, *n* = 244; **Supplementary Notes**, sec. 4.2); all box-plot whiskers show 95th/5th percentile. (**d**) RBP *in vivo* performance using PBM-trained models (c.f. ref. 22, **Fig. 1c**; **Supplementary Notes**, sec. 4.3). (**e**) AUCs of ChIP-seq models on ChIP-seq data, and of HT-SELEX models on HT-SELEX data (Wilcoxon one-sided signed-rank test, *n* = 506; **Supplementary Notes**, sec. 6.1 and 7.2). (**f**) Performance of HT-SELEX models when used to score ChIP-seq data (Wilcoxon one-sided signed-rank test, *n* = 35; **Supplementary Notes**, sec. 7.3).

DeepBind motifs for DREAM5 data and the data sets that follow are available in the online repository.

Ascertaining RNA sequence specificities

We next evaluated the ability of DeepBind to identify RNA binding sites. For experimental training data, we used previously published binding data for 207 distinct RBPs from 24 eukaryotes, including 85 distinct human RBPs[22], which were generated using the *in vitro* RNAcompete system[9]. These RBPs span different structural families, such as the RNA recognition motif, hnRNP K-homology domain and multiple zinc finger families. The arrays contain ~240,000 short single-stranded probes (30–41 nt) that cover all 9-mers at least 16 times and all 7-mers at least 155 times. The probes were randomly split into two similarly sized sets called SetA and SetB. Note that RNAcompete probes were designed to have only weak secondary structure and may not fully reflect the specificities of structure-selective RBPs such as Vtsp1 (ref. 23).

We trained DeepBind on RNAcompete SetA sequences and evaluated its predictive accuracy on held-out SetB sequences. DeepBind consistently had better overall performance than MartixREDUCE[19] and Ray *et al.*'s[9] PWM_align_Z method (**Fig. 3c**). This result was consistent across a panel of evaluation methods, including Pearson and Spearman correlations between the predicted and measured probe intensities, Pearson and Spearman correlations between all 7-mer *E*-scores (rank-based) and *Z*-scores (intensity-based), and AUCs computed from predicted and actual probe intensities (**Supplementary Notes**, sec. 4.2, **Supplementary Table 2** and **Supplementary Fig. 3**). *E*-score correlation is robust to outliers and array biases[21]. Correlations of 7-mer scores can be computed using measured data, providing upper bounds on performance. Strikingly, DeepBind performs as well as experimental *E*-scores (**Fig. 3c**).

We also applied models trained on *in vitro* RNAcompete data to *in vivo* CLIP-seq (crosslinking and immunoprecipitation followed by high-throughput sequencing) and RIP-seq (RNA-immunoprecipitation sequencing) data[22]. DeepBind was comparable or substantially better than published PFMs (position frequency matrices) on the *in vivo* data sets we considered, in particular TARDBP, PTBP1 and hnRNP A1 (**Fig. 3d**; **Supplementary Notes**, sec. 4.3, **Supplementary Table 3** and **Supplementary Fig. 4**). The ROC curves for **Figure 3d** are in **Supplementary Figure 5**.

We found that it is more challenging to train models to predict RBP sequence specificities than transcription factor specificities. Even though the DREAM5 transcription factor training and testing data sets were collected using different array designs, the Spearman correlations between predicted and measured probe intensities were significantly higher for DREAM5 transcription factors than for RNAcompete RBPs ($P = 1.8 \times 10^{-30}$; one-sided Mann-Whitney *U* test; 0.758 average, *n* = 66 and 0.439 average, *n* = 244, respectively). There are several potential reasons. First, RBPs usually bind to single-stranded RNA[22], which is more flexible than DNA and can fold into a relatively stable secondary structure. Second, RNA recognition motif, the most abundant RNA binding domain, is highly flexible, and usually the assembly of multiple domains are needed for proper binding[24]. Third, "indirect readout"[5] in which the structure of nucleotides guide the binding process, is more prominent in RBPs; Gupta and Gribskov[25] analyzed 211 protein-RNA structural complexes and reported that nearly three-quarters of protein-RNA interactions involve the RNA backbone, not just the bases.

In vitro models accurately identify in vivo bound sequences

We next evaluated DeepBind's performance using 506 *in vivo* ENCODE ChIP-seq data sets, which were preprocessed to remove protocol and laboratory biases[26] (**Supplementary Table 4**). Unlike experiments with *in vitro* data, these experiments were influenced by cell type–specific effects, transcription factor-nucleosome interactions, cooperation and competition between transcription factors and other cofactors, and pioneer transcription factors that can remodel chromatin and facilitate the binding of other transcription factors[27].

To train DeepBind, we used as positives the 101-bp sequences centered at the point source called for each peak, and we used shuffled positive sequences with matching dinucleotide composition as negatives (same as ENCODE's analysis[27]; **Supplementary Notes**, sec. 3).

For computational reasons, most existing methods analyze only the top few hundred peaks from among tens of thousands of peaks (the median number of peaks for ENCODE is ~17,000). However, Wang *et al.*[27] found that, for example, ~16,000 of the top ~20,000 SPI1 peaks contain the SPI1 motif. For each data set, DeepBind was able to

Figure 4 Analysis of potentially disease-causing genomic variants. DeepBind mutation maps (**Supplementary Notes**, sec. 10.1) were used to understand disease-causing SNVs associated with transcription factor binding. (**a**) A disrupted SP1 binding site in the *LDL-R* promoter that leads to familial hypercholesterolemia. (**b**) A cancer risk variant in a *MYC* enhancer weakens a TCF7L2 binding site. (**c**) A gained GATA1 binding site that disrupts the original globin cluster promoters. (**d**) A lost GATA4 binding site in the *BCL-2* promoter, potentially playing a role in ovarian granulosa cell tumors. (**e**) Loss of two potential RFX3 binding sites leads to abnormal cortical development. (**f,g**) HGMD SNVs disrupt several transcription factor binding sites in the promoters of *HBB* and *F7*, potentially leading to β-thalassemia and hemophilia, respectively. (**h**) Gained GABP-α binding sites in the *TERT* promoter, which are linked to several types of aggressive cancer. WT, wild type.

incorporate all ChIP-seq training peaks, and our results suggest that these extra peaks provide useful information. Specifically, we used the top 500 even-numbered peaks as held-out test data and trained models on either all other peaks (DeepBind*) or just the top 500 odd-numbered peaks (DeepBind) (**Supplementary Notes**, sec. 6.1 and **Supplementary Fig. 6**). For comparison, we also applied MEME-ChIP[28] to the same top 500 odd-numbered training peaks, derived five PWMs for each experiment using the ENCODE settings[27], and scored test sequences using either the top PWM (MEME-M1) or the sum of scores for all five PWMs (MEME-SUM). MEME-ChIP could not be trained using all possible training peaks due to computational limitations.

The 506 ChIP-seq experiments represent 137 unique transcription factors, so we computed one AUC for each transcription factor by taking its median test AUC across all associated experiments (antibodies, cell lines, laboratories). DeepBind trained on the top 500 odd peaks achieved higher test AUC (0.85 average) than both MEME-SUM (0.82 average; $P = 2 \times 10^{-15}$, Wilcoxon one-sided signed-rank test, $n = 137$) and MEME-M1 (0.78 average; $P = 1 \times 10^{-23}$). By training DeepBind* on all peaks not held out we achieved significantly higher test AUC (0.90 average) than training on the top 500 odd peaks either DeepBind ($P = 2.4 \times 10^{-20}$) or MEME-SUM ($P = 7 \times 10^{-23}$) (**Fig. 3e** and **Supplementary Table 5**). We additionally tested the DeepBind FoxA2 ChIP-seq model using 64 'electrophoretic mobility shift assay' (EMSA)-measured binding affinities[29] and found that it achieves the highest Spearman correlation among published methods, including JASPAR PWMs[30], TRANSFAC PWMs[31], and other models trained on ChIP-seq data (**Supplementary Notes**, sec. 6.3 and **Supplementary Fig. 7**).
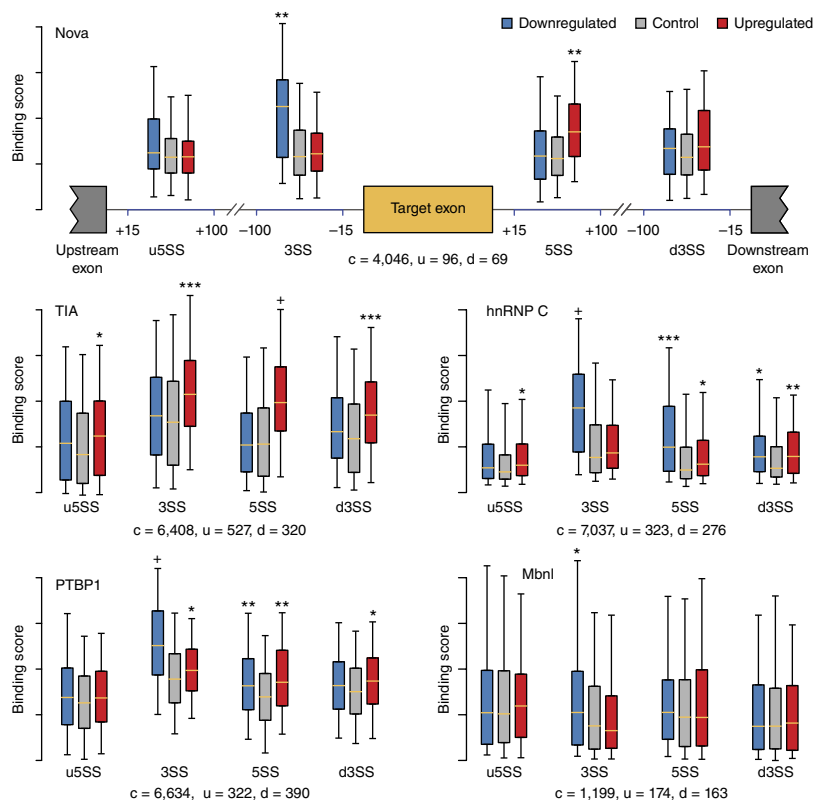
We also trained DeepBind on Jolma *et al.*'s *in vitro* HT-SELEX data for 303 human DNA-binding domains, 84 mouse DNA-binding domains and 151 full-length human transcription factors, representing 411 distinct transcription factors[32] (**Supplementary Table 6**). DeepBind achieved higher test AUC (0.71 average) than the PWMs discovered by Jolma *et al.*'s semi-automatic algorithm[32] (0.63 average; $P = 2.8 \times 10^{-94}$, Wilcoxon one-sided signed-rank test, $n = 566$) (**Fig. 3e** and **Supplementary Notes**, sec. 7.2, and **Supplementary Table 7**). We next evaluated the performance of several HT-SELEX models for which there exists corresponding *in vivo* ChIP-seq data from ENCODE. DeepBind achieved higher *in vivo* AUC (0.85 average) than both the best-possible choice of Jolma *et al.*'s PWMs[32] (0.82 average; $P = 0.001$ Wilcoxon one-sided signed-rank test, $n = 35$)

Figure 5 DeepBind models are used to describe the regulation mechanism for different RBPs. All *P* values are computed between predicted scores of upregulated and/or downregulated exons and scores of control exons (Mann-Whitney *U* test; $n = c + u$ for upregulated vs. control exons, and $n = c + d$ for downregulated vs. control exons). * $1 \times 10^{-8} < P \leq 1 \times 10^{-4}$; ** $1 \times 10^{-16} < P \leq 1 \times 10^{-8}$; *** $1 \times 10^{-32} < P \leq 1 \times 10^{-16}$; + $P \leq 1 \times 10^{-32}$. The number of up-, down- and control exons are denoted by u, d and c, respectively. All box-plot whiskers show 95th and 5th percentile. u5SS, 3SS, 5SS and d3SS: intronic regions close to upstream exon's 5′ splice site, target exon's 3′ and 5′ splice sites, and downstream exon's 3′ splice site, respectively.

and the best MEME-ChIP configuration MEME-SUM-5k (0.82 average; $P = 0.002$), suggesting that DeepBind can generalize from HT-SELEX to other data acquisition technologies despite being based on a general-purpose machine learning framework (**Fig. 3f** and **Supplementary Notes**, sec. 7.3, and **Supplementary Table 8**). The ROC curves for **Figure 3f** are in **Supplementary Figure 8**.

Identifying and visualizing damaging genetic variants

Genetic variants that create or abrogate binding sites can alter gene expression patterns and potentially lead to disease[33]. A promising direction in precision medicine is to use binding models to identify, group and visualize variants that potentially change protein binding. To explore the effects of genetic variations using DeepBind, we developed a visualization called a 'mutation map' (**Supplementary Notes**, sec. 10.1), which illustrates the effect that every possible point mutation in a sequence may have on binding affinity. A mutation map conveys two types of information. First, for a given sequence, the mutation map shows how important each base is for the DeepBind analysis by the height of the base letter. Second, the mutation map includes a heat map of size 4 by *n*, where *n* is the sequence length, indicating how much each possible mutation will increase or decrease the binding score.

We examined variants within promoters using nearly 600 DeepBind models trained using ChIP-seq and HT-SELEX data. **Figure 4a** shows how several Human Gene Mutation Database (HGMD)[34] single-nucleotide variants (SNVs) disrupted an SP1 binding site in the *LDL-R* promoter, leading to familial hypercholesterolemia[35]. It can be seen in the wild-type mutation map that all HGMD SNVs falling in the SP1 binding site decrease the score. **Figure 4b** depicts how a cancer risk variant (rs6983267) in a *MYC* enhancer[36] weakens a TCF7L2 binding site. **Figure 4c** illustrates how an inherited gain-of-function SNP creates a GATA1 binding site in the globin cluster, disrupting the original globin cluster promoters[37]. GATA4 is known to regulate *Bcl-2* (B-cell CLL/lymphoma 2) expression in ovarian

Figure 6 Comparison of motifs learned by DeepBind with known motifs. Example motif detectors learned by DeepBind models, along with known motifs from CISBP-RNA[22] (for RBPs) and JASPAR[30] (for transcription factors). A protein's motifs can collectively suggest putative RNA- and DNA-binding properties, as outlined[51], such as variable-width gaps (HNRNPA1, Tp53), position interdependence (CTCF, NR4A2), and secondary motifs (PTBP1, Sox10, Pou2f2). Motifs learned from *in vivo* data (e.g., ChIP) can suggest potential co-factors (PRDM1/EBF1) as in Teytelman *et al.*[12]. Details and references for 'known motifs' are in **Supplementary Notes**, sec. 10.2.
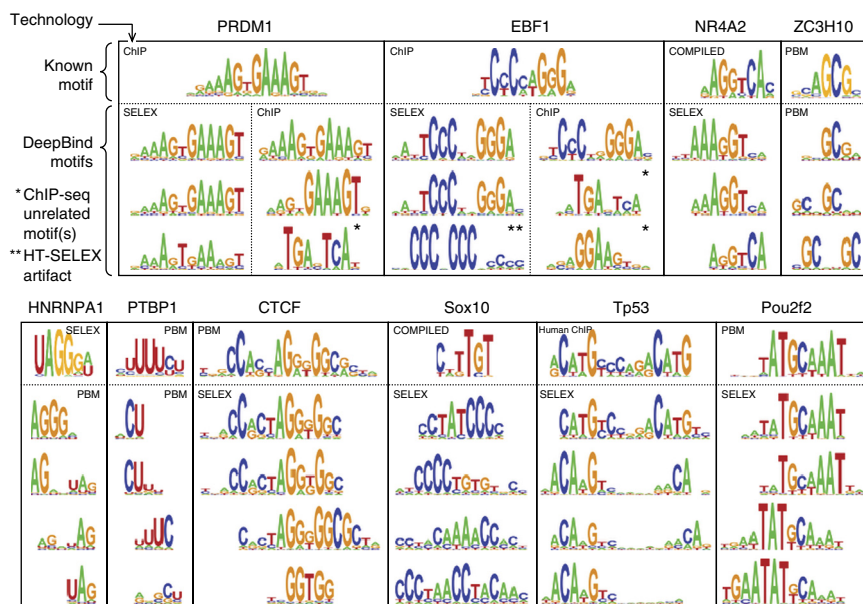
granulosa cell tumors[38]. We analyzed all COSMIC 69 (ref. 39) SNVs located in the *Bcl-2* promoters and found two that damage a single GATA4 binding site (**Fig. 4d**).

One notable example involves tandem RFX3 binding sites in the *GPR56* promoter. A 15-bp deletion of a highly conserved segment deletes one of the binding sites, resulting in abnormal cortical development[40]. DeepBind analysis discovered a third RFX3 binding site located on the opposite strand, overlapping both of the known tandem binding sites (**Fig. 4e**), and so the 15-bp deletion might destroy this third RFX binding site as well.

Two of the highest-scoring SNV-ridden promoters in HGMD belong to *HBB* (hemoglobin beta chain) and *F7* (coagulation factor VII). Disruption of these two genes results in β-thalassemia and hemophilia, respectively. **Figure 4f,g** shows how numerous HGMD SNVs might damage transcription factor binding sites in these promoters.

Lastly, it was shown that mutations in the *TERT* (telomerase reverse transcriptase) promoter are linked to many types of aggressive cancer[41], mostly by gain-of-function mutations creating ETS family[42,43] or GABP[41] transcription factor binding sites. We analyzed the *TERT* promoter using DeepBind ETS (ELK1/ELK4) and GABP-α models and the resulting mutation maps all show potential gain-of-function mutations corresponding to the literature. In **Figure 4h**, all COSMIC 69 SNVs observed in the analyzed regions are shown. It can be seen that the two highly recurrent somatic G→A SNVs (G228A and G250A) create the strongest putative GABP-α binding sites, which have each been confirmed to increase expression of *TERT*[41]. In addition to these highly recurrent mutations, DeepBind also identified a familial T→G mutation that is known to increase *TERT* expression, potentially by introducing a GABP-α binding site[42].

DeepBind models identify deleterious genomic variants

We investigated whether DeepBind can be used to predict deleterious SNVs in promoters, by training a deep neural network to discriminate between high-frequency derived alleles (neutral or negative) and simulated variants (putatively deleterious, or positive) from the CADD framework[44]. The scores of ~600 DeepBind transcription factor models for the wild type and mutant sequences were used as inputs (~1,200 inputs; **Supplementary Fig. 9**). The rationale is that a true transcription factor binding site is likely to be located with other transcription factor binding sites, and so these additional scores collectively provide context. When evaluated using held-out test data, the neural network, called DeepFind, achieved an AUC of 0.71, which increased to 0.73 when we included as inputs the distance to the closest transcription start site and a transversion/transition flag. When we included nine

conservation features, the AUC increased to 0.76. **Supplementary Figure 10** shows that distribution of DeepFind scores for derived alleles and simulated variants. For comparison, the published genome-wide CADD scores achieved an AUC of 0.64, even though they were computed using hundreds of features, including overlapping ChIP-seq peaks, transcription factor binding sites, DNA accessibility profiles and chromatin state, among others (**Supplementary Notes**, sec. 9).

In vitro models are consistent with known splicing patterns

Alternative splicing, through which multiple transcripts are produced from a single gene, is responsible for generating significant transcriptional diversity in metazoa, to the degree that more than 95% of multi-exon human genes are alternatively spliced[45]. RBPs play a crucial role in regulating splicing, having an impact on a wide variety of developmental stages such as stem cell differentiation[46] and tissue development[47]. We predicted binding scores at junctions near exons that are putatively regulated by known splicing regulators that exhibit large changes in splicing when knocked down, including: Nova (neuro-oncological ventral antigen 1 and 2), PTBP1 (polypyrimidine tract binding protein 1), Mbnl1/2 (muscleblind-like protein), hnRNP C (heterogeneous nuclear ribonucleoprotein C) and cytotoxic TIA (T-cell intracellular antigen-like 1).

Predictions were consistent with experimental CLIP-seq data and known binding profiles of studied RBPs (**Fig. 5** and **Supplementary Notes**, sec. 8, and **Supplementary Table 9**). For example, exons known to be downregulated by Nova had higher Nova scores in their upstream introns, and exons known to be upregulated by Nova had higher Nova scores in their downstream intron[48]. Similarly, TIA has been shown to upregulate exons when bound to the downstream intron[49], and PTBP1 has been shown to suppress exon inclusion when bound to upstream introns of weaker splice sites[50].

DISCUSSION

Though there is no single agreed upon metric for evaluating the quality of sequence specificity predictions[17], we found that DeepBind surpasses the state of the art across a wide variety of data sets and evaluation metrics. Importantly, our results show that DeepBind models trained *in vitro* work well at scoring *in vivo* data, suggesting an ability to capture genuine properties of nucleic acid binding

interactions. DeepBind scales well to large data sets and, for both ChIP-seq and HT-SELEX, we found that there was valuable information to be learned from sequences that other techniques discard for computational reasons.

A frequent concern with large, deep models is that they can overfit the training data. To address this, we incorporated several regularizers developed in the deep learning community, including dropout, weight decay and early stopping (**Supplementary Notes**, sec. 2). Indeed, we found that the calibration phase of DeepBind training was crucial for obtaining models that generated robust predictions on held-out test data; our experiments indicate that at least 30 calibration settings should be evaluated to obtain reliable training settings across all data sets (**Supplementary Notes**, sec. 2.2). Automatic calibration and training (**Fig. 2b**) are computationally demanding, so we implemented these stages on powerful GPUs for a 10–70× acceleration. Once a model was trained, we extracted its parameters and provided a CPU implementation of the model for easy application by downstream users. Users can browse the available RBP and transcription factor models using our online repository, which displays a familiar PWM-like representation of each model (**Fig. 6**).

DeepBind is based on deep learning, a scalable and modular pattern discovery method, and does not rely on common application-specific heuristics such as 'seed finding'. Deep learning furthermore has an extremely active research community that is garnering huge investment from academia and industry; we believe current and future insights from the deep learning community will lead to enhancements to DeepBind and to sequence analysis in general.

METHODS
Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS
B.A., A.D. and B.J.F. conceived the method. A.D. implemented DeepBind and the online database of models. B.A. designed the experiments with input from A.D., M.T.W., and B.J.F., and also implemented DeepFind. B.A., A.D. and B.J.F. wrote the manuscript with valuable input from M.T.W.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Stormo, G. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
2. Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).
3. Kazan, H., Ray, D., Chan, E.T., Hughes, T.R. & Morris, Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.* **6**, e1000832 (2010).
4. Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29**, 659–664 (2011).
5. Siggers, T. & Gordân, R. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.* **42**, 2099–2111 (2014).
6. Krizhevsky, A., Sutskever, I. & Hinton, G.E. in *Advances in Neural Information Processing Systems* (eds. Pereira, F., Burges, C.J.C., Bottou, L. & Weinberger, K.Q.) 1097–1105 (Curran Associates, 2012).
7. Graves, A., Mohamed, A. & Hinton, G. Speech recognition with deep recurrent neural networks. *ICASSP* 6645–6649 (2013).
8. Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36**, 1331–1339 (2004).
9. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* **27**, 667–670 (2009).
10. Kharchenko, P., Tolstorukov, M. & Park, P. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).
11. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).
12. Teytelman, L., Thurtle, D.M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. USA* **110**, 18602–18607 (2013).
13. LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551 (1989).
14. Cotter, A., Shamir, O., Srebro, N. & Sridharan, K. in *Advances in Neural Information Processing Systems* (Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F. & Weinberger, K.Q.) 1647–1655 (Curran Associates, 2011).
15. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
16. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
17. Weirauch, M.T. *et al.* Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31**, 126–134 (2013).
18. Zhao, Y., Stormo, G.D., Feature, N. & Eisenstein, M. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* **29**, 480–483 (2011).
19. Foat, B.C., Morozov, A.V. & Bussemaker, H.J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141–e149 (2006).
20. Chen, X., Hughes, T.R. & Morris, Q. RankMotif.: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics* **23**, i72–i79 (2007).
21. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
22. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
23. Oberstrass, F.C. *et al.* Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. *Nat. Struct. Mol. Biol.* **13**, 160–167 (2006).
24. Daubner, G.M., Cléry, A. & Allain, F.H.-T. RRM-RNA recognition: NMR or crystallography…and new findings. *Curr. Opin. Struct. Biol.* **23**, 100–108 (2013).
25. Gupta, A. & Gribskov, M. The role of RNA sequence and structure in RNA–protein interactions. *J. Mol. Biol.* **409**, 574–587 (2011).
26. Landt, S. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
27. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
28. Machanick, P. & Bailey, T.L. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
29. Levitsky, V.G. *et al.* Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genomics* **15**, 80 (2014).
30. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).
31. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
32. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
33. Lee, T.I. & Young, R.A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
34. Stenson, P. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).
35. De Castro-Orós, I. *et al.* Functional analysis of LDLR promoter and 5′ UTR mutations in subjects with clinical diagnosis of familial hypercholesterolemia. *Hum. Mutat.* **32**, 868–872 (2011).
36. Pomerantz, M.M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.* **41**, 882–884 (2009).
37. De Gobbi, M. *et al.* A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312**, 1215–1217 (2006).
38. Kyrõnlahti, A. *et al.* GATA-4 regulates Bcl-2 expression in ovarian granulosa cell tumors. *Endocrinology* **149**, 5635–5642 (2008).
39. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
40. Bae, B.-I. *et al.* Evolutionarily dynamic alternative splicing of GPR56 regulates regional cerebral cortical patterning. *Science* **343**, 764–768 (2014).

41. Bell, R.J.A. *et al.* The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science* **348**, 1036–1039 (2015).
42. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
43. Huang, F. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
44. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
45. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
46. Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241–245 (2013).
47. Fogel, B.L. *et al.* RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. *Hum. Mol. Genet.* **21**, 4171–4186 (2012).
48. Ule, J. *et al.* An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**, 580–586 (2006).
49. Del Gatto-Konczak, F. *et al.* The RNA-binding protein TIA-1 is a novel mammalian splicing regulator acting through intron sequences adjacent to a 5′ splice site. *Mol. Cell. Biol.* **20**, 6287–6299 (2000).
50. Xue, Y. *et al.* Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell* **36**, 996–1006 (2009).
51. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).

npg

ONLINE METHODS

Deep learning techniques. We implemented several of the simplest and most effective techniques now practiced within the deep learning community: mini-batch stochastic gradient descent (SGD), Nesterov momentum[14], rectified linear units[6] (ReLUs), dropout regularization[15], automatic model selection through calibration[16] ('hyper-parameter search') and, finally, GPU acceleration to facilitate fast training on large data sets. When combined, these simple techniques have achieved breakthrough performance in computer vision[6], speech recognition[7] and more. **Supplementary Notes**, sec. 1 and 2 explain the role of each technique in more detail.

Given a specific convolutional network architecture (number of motif detectors, lengths of motif detectors, width of pooling, number of layers, number of parameters), the training procedure for DeepBind is conceptually straightforward and can be outlined as follows. At the outset of training, all parameters are initialized to random values (**Supplementary Notes**, sec. 2.1). A 'mini-batch' of N randomly selected input sequences is then fed through the network, generating N (initially random) score predictions. The discrepancy between each prediction and its corresponding target is used to improve the performance of the network through a step known as 'back-propagation' (**Fig. 2** and **Supplementary Notes**, sec. 1.2). Training then continues with a new mini-batch of N randomly selected input sequences. The number of iterations (mini-batches) used in a final training run is determined by the automatic calibration phase.

Back propagation is merely an efficient way to compute the partial derivative of each parameter of a deep model with respect to a training objective. Given a random mini-batch of training sequences $s^{(1.N)}$ and corresponding target scores $t^{(1.N)}$, the performance of the network is improved by approximately minimizing the training objective

$$\frac{1}{N} \sum_{i=1}^{N} \text{LOSS}\left(f(s^{(i)}), t^{(i)}\right) + \beta_1 \, \|M\|_1 + \beta_2 \, \|W\|_1$$

where $\|\cdot\|_1$ denotes the L1 norm (sum of absolute values), weight decay coefficients $\beta_1, \beta_2 \geq 0$, and LOSS(p,t) is a function such as the squared error $(p - t)^2$ or negative log-likelihood (**Supplementary Notes**, sec. 1.3).

Automatic calibration. Deep neural networks typically require human expertise and repeated manual attempts to train correctly. The reason is that the networks themselves, and their training procedure, are sensitive to many calibration parameters, also known as 'hyper-parameters' or 'nuisance parameters' (**Supplementary Notes**, sec. 2.1). Calibration parameters must be chosen to avoid under- or overfitting to the training data, but are otherwise irrelevant to the end-user. For neural networks, typical parameters include 'learning rates', 'network depth', 'initial weights' and 'weight decay'. The complete list of hyperparameters and the way they are sampled are summarized in **Supplementary Table 10**. Similarly, for support vector machines one must likewise choose a kernel function and weight decay. For random forests one must choose a tree depth and an ensemble size.

For each DeepBind model to be trained, we randomly sampled 30 complete sets of calibration parameters[16], and for each fold of a three-way random split of the training data we trained 30 models in parallel on a single GPU. We rate each calibration by its average validation performance across those folds (threefold cross-validation AUC or mean squared error). The best calibration was then used to train a new model on all the training data (**Supplementary Notes**, sec. 2.2). The code for performing the experiments is available for download in the **Supplementary Software**.