Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks.

David R. Kelley[1]
Jasper Snoek[2]
John L. Rinn[1]

1 Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
2 School of Engineering and Applied Science, Harvard University, Cambridge, MA 02138, MA

Contact information:
David R. Kelley
dkelley@fas.harvard.edu
732-849-4305
7 Divinity Ave.
Cambridge, MA 02138

Jasper Snoek
jsnoek@seas.harvard.edu
School Of Engineering and Applied Science
Harvard University
Cambridge, MA 02138

John L. Rinn
john_rinn@harvard.edu
7 Divinity Ave.
Cambridge, MA 02138

Abstract

The complex language of eukaryotic gene expression remains incompletely understood. Thus, most of the many noncoding variants statistically associated with human disease have unknown mechanism. Here, we address this challenge using an approach based on a recent machine learning advance—deep convolutional neural networks (CNNs). We introduce an open source package Basset (https://github.com/davek44/Basset) to apply deep CNNs to learn the functional activity of DNA sequences from genomics data. We trained Basset on a compendium of accessible genomic sites mapped in 164 cell types by DNaseI-seq. Basset predictions for the change in accessibility between two variant alleles were far greater for GWAS SNPs that are likely to be causal relative to nearby SNPs in linkage disequilibrium with them. With Basset, a researcher can perform a single sequencing assay in their cell type of interest and simultaneously learn that cell's chromatin accessibility code and annotate every mutation in the genome with its influence on present accessibility and latent potential for accessibility. Thus, Basset offers a powerful computational approach to annotate and interpret the noncoding genome.

Introduction

The process of identifying genomic sites that show statistical relationships to phenotypes holds great promise for human health and disease (Hindorff et al. 2009). However, our current inability to efficiently interpret noncoding variants impedes progress toward using genomes for personalized medicine. Coordinated efforts to survey the noncoding genome have shown that sequences marked by genome accessibility and certain histone modifications are enriched for variants that are statistically related to phenotypes (Consortium 2012; Kundaje et al. 2015). The first stages of a mechanistic hypothesis can now be assigned to variants that directly overlap these annotations (Ritchie et al. 2014; Kircher et al. 2014; Fu et al. 2014).

However, simply comparing variant and annotation positions underutilizes these data; more can be extracted by understanding the protein interaction as a function of the underlying sequence. Proteins that recognize specific signals in the DNA influence genome accessibility and histone modifications (Voss and Hager 2014). Given training data, machine learning models can learn to predict protein binding, DNA accessibility, histone modifications, and DNA

methylation from the sequence (Benveniste et al. 2014; Lee et al. 2015; Pinello et al. 2014; Das et al. 2006; Setty and Leslie 2015; Whitaker et al. 2015; Arnold et al. 2013). A trained model can then predict these regulatory attributes across the genome and annotate the influence of every nucleotide (and variant).

In the last few years, artificial neural networks with many stacked layers have achieved breakthrough advances on benchmark datasets in image analysis (Krizhevsky et al. 2012) and natural language processing (Collobert et al. 2011). Rather than choose features manually or in a preprocessing step, convolutional neural networks (CNNs) adaptively learn them from the data during training. They apply nonlinear transformations to map input data to informative high-dimensional representations that trivialize classification or regression (Bengio et al. 2013). Early applications of CNNs to DNA sequence analysis predict protein binding and accessibility from DNA sequence and surpass more established algorithms, such as support vector machines or random forests (Alipanahi et al. 2015; Zhou and Troyanskaya 2015). More accurate models can precisely dissect the influence of regulatory sequences, thus improving noncoding variant interpretation. However, to fully exploit the value of these models, it is essential that they are technically and conceptually accessible to the researchers who can take advantage of their potential.

Here, we introduce Basset, an open source package to apply deep CNNs to learn functional activities of DNA sequences. We used Basset to simultaneously predict the accessibility of DNA sequences in 164 cell types mapped by DNaseI-seq by the ENCODE and Epigenomics Roadmap projects (Consortium 2012; Kundaje et al. 2015). From these datasets, CNNs simultaneously learn the relevant sequence motifs and the regulatory logic with which they are combined to determine cell-specific DNA accessibility. We show that a model achieving this level of accuracy provides meaningful nucleotide-precision measurements. Subsequently, we assign GWAS variants cell-specific scores that reflect the accessibility difference predicted by the model between the two alleles. These scores are predictive of the causal SNP among sets of linked variants. Importantly, Basset puts CNNs in the hands of the genome biology community, providing tools and strategies for researchers to train and analyze models on new datasets. In conjunction with genomic big data, Basset offers a promising future for human genetic research.

Deep convolutional neural networks predict genome accessibility

DNA sequence codes for the chromatin shifts that transform cells in development and disease. We focused here on DNA accessibility due to the abundance of available data and significant association with conserved segments (Thurman et al. 2012), disease variation (Maurano et al. 2012), and eQTLs (Degner et al. 2012). The ENCODE consortium performed DNaseI-seq on 125 cell types (Thurman et al. 2012), and the Epigenomics Roadmap consortium curated an additional 39 (Kundaje et al. 2015). We collected and merged these sets, resulting in 2 million sites across all cells (Figure 2A). These sites are 17% promoters, 47% intragenic, and 36% intergenic according to GENCODE v18 hg19 annotations. 4.1-19.0% (median 8.2%) are accessible in any individual cell type, with 3.8% constitutively expressed in >50% of the cells. For each DHS, we extracted 600 base pairs (bp) around the midpoint of each site as input to the model.

To learn the DNA sequence signals of open versus closed chromatin, we applied a deep convolutional neural network (CNN). CNNs have proven highly effective in a number of diverse tasks; this set recently includes biological sequence analysis (Alipanahi et al. 2015; Zhou and Troyanskaya 2015). As opposed to manually specifying features or performing a preprocessing step to statistically learn them, CNNs perform adaptive feature extraction to map input data to informative representations during training. The convolution operation is the engine of the CNN. In a convolution layer, the algorithm scans a set of weight matrices called filters across the input; these weight matrices learn to recognize relevant patterns (Figure 1). Prior work has demonstrated that with a sufficiently large dataset, deep neural networks can learn far more expressive and accurate models than other common approaches like random forests or kernel methods (Bengio et al. 2013).

For DNA sequences, the initial convolution layer corresponds to optimizing the weights of a set of position weight matrices (PWMs), which are a well-studied tool in bioinformatics (Stormo 2000). These PWM filters search for their motifs along the sequence and output a matrix with a row for every filter and column for every position in the sequence (Figure 1). Computing nonlinear functions of the information flowing through the network allows for more expressive models. In each convolution layer, we apply a rectifier operation (i.e. set negative values to 0) to the matrix of filter output (Nair and Hinton 2010). Finally, we pool adjacent values by taking the maximum in a small window. This operation reduces the dimension of the input to the next

layer (and thus the computation required in training). It also provides invariance to small sequence shifts to the left or right.

Subsequent convolutional layers operate analogously on the output of the prior layer. Thus, for DNA sequences, they capture spatial interactions between the initial PWM filter outputs. The full architecture of the neural network includes three convolutional layers and two layers of fully connected hidden nodes. In general, deeper networks are able to learn more abstract representations; here, depth allows the algorithm to learn sophisticated regulatory codes that combine the recognized sequence motifs. The final layer outputs a prediction for the probability that the sequence is accessible in each of the 164 cell types. During training, we compare this prediction to the experimentally measured accessibilities and update the model parameters to improve the prediction (see Methods).

We have released open-source software implementing all procedures described in this manuscript, including routines to preprocess common functional genomics data, train the network, and extract the knowledge it has learned. We have named the package Basset, as an allusion to the extraordinary abilities of this hound dog to learn a scent that they are subsequently able to pursue and detect.

We trained Basset and a recently published advance based on gapped kmer support vector machines called gkm-SVM (Lee et al. 2015) to predict the accessibility of a set of test sequences in 164 cell types (see Methods). To synthesize sensitivity and specificity, which inherently trade off, we assessed the models using the area under the receiver operating characteristic curve (AUC), which plots the false positive rate versus the true positive rate. By this measure, Basset is exceptionally accurate, achieving a mean AUC of 0.892 over all cells, relative to 0.760 for gkm-SVM (Figure 2B). Basset substantially improved the AUC for every cell type. At a false positive rate of 10%, Basset identifies 55-80% of true positive DHS sequences (Figure 2C). In addition to predicting constitutive sites, Basset effectively captures cell and lineage-specific accessibility (Supplementary Figure 1).

Basset recovers known protein binding motifs

Though deep neural networks are exceptionally accurate, the principles that they learn are not trivially interpretable as they are in simpler linear models. However, substantial information can

be extracted from the model by examining its parameters, modulating its actions, and exploring its predictions on purposefully chosen sequences.

Nucleosome-free regions of DHSs are influenced by protein binding to naked DNA and occluding the nucleosome (Voss and Hager 2014; Sherwood et al. 2014). Thus, we expect that a predictive model of accessibility will capture this dependence by learning the DNA binding sites of a variety of universal and cell-specific proteins. The first convolution layer of the model scans the DNA sequence with a set of pattern-recognizing filters whose weights are optimized during training. These filters are ideal for capturing this protein-binding information.

The model's 300 convolution filters recovered an extensive repertoire of known DNA binding protein motifs (Figure 3B). To aid in their interpretation, we nullified each filter by setting its output to the mean value, which obstructs it from passing any information forward through the network. We computed the L2 norm of the change in predicted accessibility to measure its influence.

The AP-1 complex, consisting of the proteins JUN and FOS, emerged in 4 filters as the most influential transcription factor (TF) in the set. AP-1's important role in opening chromatin has been established (Biddie et al. 2011). The model dedicated 12 filters to comprehensively represent the extraordinarily long DNA recognition site of the critical chromatin protein CTCF. Each filter focused on separate portions and variations of the motif. At a q-value threshold of 0.1, 45% of the filters aligned significantly to protein motifs in the CIS-BP database, which were originally acquired by independent ChIP-seq or in vitro experiments (Gupta et al. 2007; Weirauch et al. 2014). This set included highly similar weight matrices for many developmental regulators; a sample biased towards longer and more influential motifs is depicted in Figure 3B. Many more filters capture partial coverage of known motifs, which were deemed insignificant matches to the database after multiple testing correction.

Most unrecognized filters captured lower-order sequence composition. One example directly detects CpG's, which can be methylated and are a well-studied feature of regulatory modules (Bird and Wolffe 1999). Others measure AT-richness or the absence of CpG's. Their influence emphasizes the importance of the local sequence context of binding motifs to their function (Figure 3A) (Levo and Segal 2014).

Several unrecognized filters had high information content, which indicates that they may refer to unannotated proteins or alternative binding modes of annotated ones. To further explore these filters, we computed the influence of nullifying each filter on the downstream cell accessibility predictions (Supplementary Figure 2). Clustering these influence profiles reveals modules of filters matching proteins known to regulate development to their active cell types. For example, filters matching database motifs for known epithelial regulators P63, GRHL1, and KLF factors encourage accessibility in a variety of epithelial cells (Pignon et al. 2013; Wilanowski et al. 2008; Ray and Pollard 2012). The unrecognized filters span a range of cell preferences, and future work will be necessary to determine their role and potential binding proteins.

in silico saturation mutagenesis pinpoints nucleotides driving accessibility

Saturation mutagenesis experiments are a powerful tool for dissecting the exact nucleotides driving some activity (Patwardhan et al. 2009; Melnikov et al. 2012). Our trained model can be used to predict the accessibility of arbitrary sequences. Thus, by computing the expected accessibility of all possible mutations to a sequence, Basset can be used to perform an in silico saturation mutagenesis.

We constructed heat maps that display the change in predicted accessibility from mutation at every position to each alternative nucleotide. These maps highlight the individual nucleotides most critical to a sequence's activity. We assigned two scores to every position: (1) the loss score measures the largest possible decrease and (2) the gain score measures the largest increase.

High loss scores mark positions with existing functional motifs where mutations can damage the motif and decrease accessibility. For example, the sequence mapped in Figure 4 is accessible in embryonic stem cells and contains a motif recognized by AP-1. ChIP-seq of JUN and JUND in H1-hESCs support the binding site, and PhyloP conservation statistics support the precise relevance of the TGASTCA motif (Pollard et al. 2010). Mutations within the motif and numerous flanking nucleotides result in decreased predicted accessibility. Genome-wide, a synthesis of loss scores from all cell types significantly correlated with PhyloP (Pearson 0.187; p-value 3.8e-52) (see Methods) (Figure 4b).

In contrast, the gain score suggests latent potential of a sequence; the corresponding mutation typically introduces a functional motif and increases the predicted accessibility. Although such a position does not pinpoint a present functional motif, there may be negative selection against one forming and rearranging accessibility in the region. The addition of gain scores increased the Pearson correlation with PhyloP to 0.219 (p-value 1.4e-71), supporting this effect (Figure 4d).

Basset predicts greater accessibility changes for likely causal GWAS SNPs

Genome-wide association studies (GWAS) have uncovered ample noncoding variants associated with physical traits and disease in human populations (Welter et al. 2014). DHSs are highly enriched for GWAS SNPs, which can modulate the accessibility of the site to affect local gene expression (Maurano et al. 2012; Degner et al. 2012). Basset captures the sequence signals driving accessibility and ought to have predictive power for prioritizing noncoding variants and suggesting mechanistic hypotheses for further investigation into their causal role for the phenotype. For this purpose, we defined SNP Accessibility Difference (SAD) profiles as the difference in predicted accessibility between two alleles across all cell types.

The scarcity of confirmed positive examples of noncoding causal variants challenges a thorough assessment of the value of SAD scores for GWAS prioritization. Instead, we studied probabilistic assessments of causality assigned by an orthogonal method. 8741 GWAS SNPs associated with auto-immune disease were analyzed with a statistical method called PICS that leverages dense genotyping data to assign a probability of being the causal SNP among a nearby set of SNPs in linkage disequilibrium (LD) (Farh et al. 2015). In many cases, PICS identifies the causal SNP with high probability using only population linkage data.

We focused on a set of 7252 GWAS SNPs for which no SNP in LD affects a protein coding gene and classified 232 high-PICS SNPs that were assigned causal probability >0.5. SAD profile means were significantly greater for the set of high-PICS SNPs (Mann-Whitney U Test, p-value 1.8e-6). 5 times more high-PICS SNPs than low were predicted to change accessibility change by an average of >0.1 over the cell types (Figure 5A). Coverage of the SNPs in this set is wide—34% of all index SNPs had at least one SNP in its LD set for which the model predicted a >10% change in probability of accessibility in some cell type.

Among the agreements with PICS was rs4409785, associated with vitiligo (Jin et al. 2012), rheumatoid arthritis (Okada et al. 2013), and immune mechanisms in multiple sclerosis (Sawcer et al. 2011). PICS assigned rs4409785 85.3% probability of causality for vitiligo. The SNP is located in a 559 Kb gene desert. However, it has been hypothesized to regulate *TYR*, which although 6.28 Mb away, offers a plausible mechanism for the skin color disease vitiligo. *TYR* catalyzes conversion of tyrosine to melanin, the pigment that gives skin its color (Jin et al. 2012).

Basset predictions support this hypothesis; the more prevalent T allele is devoid of activity, but the C allele creates a motif recognized by the model's CTCF filters (Figure 5B). Although this sequence imperfectly matches the CTCF database motif, Basset predicts dramatically increased accessibility in all cell types, including an increase in H1-hESCs from 0.8% probability to 73.24%. To assess experimental evidence for allele-specific CTCF binding, we downloaded 258 CTCF ChIP-seq datasets performed in a variety of cell types by ENCODE (Consortium 2012). Only 22 unique cell types had >2 reads aligned to rs4409785, which supports the default inaccessibility of the region (Figure 5C). All 10 cells that have >6 reads aligned sequenced only the C allele, thus confirming the allele specificity.

Recent work has demonstrated that modulating CTCF binding at topologically associating domain boundaries can dramatically affect gene expression by altering the architecture of the genomic region (Lupiáñez et al. 2015; Guo et al. 2015). By modeling the underlying sequence driving the experimental measurement, Basset's highlight of rs4409785 suggests a functional hypothesis to accelerate study of this widely influential variant.

Leveraging large-scale models allows accurate and efficient prediction of new datasets. To make the most of machine learning to identify causal SNPs, researchers must train models on data from the appropriate (perhaps specialized) cell types (Lee et al. 2015). Such cells will often be overlooked by large-scale mapping projects, but experimentally characterized by individual labs. We designed Basset to accommodate this frequent and critical scenario. We also hypothesized that Basset can rapidly learn to accurately predict new data by leveraging a model pre-trained on public data.

To test this hypothesis, we removed 15 datasets spanning the range of AUC from the pool of 164 studied above. We re-trained on the remaining 149 "public" datasets to establish a “pre-trained” model. This model achieved an average AUC of 0.892 across the 149 cells, on par with the full model analyzed above.

For each remaining dataset individually, we sampled a matching number of sites from the “public” 149 to serve as negative examples. Rather than train from scratch on the dataset alone, we initialized the model parameters with those from the “pre-trained” model. By providing this head start, we need only perform one training pass through the new data to achieve models with predictive accuracies rivaling those from the full 164 cell model above (Figure 6). We could complete this training procedure in an average of 6.5 hours on a Macbook 2.8 GHz Intel core i7 or 18 minutes on an NVIDIA Tesla K20m GPGPU. Thus, this approach allows researchers to train highly accurate deep CNN models on common computer hardware in a few hours.

Discussion

In this work, we introduced Basset—an open source package to apply deep CNNs to learn DNA sequence activity. Basset effectively learned the complex code of DNA accessibility across many cell types and substantially surpassed the predictive accuracy of the present state-of-the-art. We demonstrated how our model precisely implicates the nucleotides driving activity, highlighting genomic positions with either fragile activity that can be lost by mutation or latent potential activity that can be unlocked by mutation. These sites are more conserved and their mutation is more likely to cause disease.

The most successful prior approaches to analyze noncoding variants compare them to the broad regions that functional genomics experiments have annotated to have reproducible accessibility, protein binding, and/or histone modifications (Kircher et al. 2014; Fu et al. 2014; Ritchie et al. 2014). Basset has two primary benefits over this approach. By directly modeling the mapping from sequence to activity, Basset implicates the precise nucleotides influencing accessibility, providing a finer resolution view than mere overlap with a broad region. Many nucleotides overlapped by these regions are assigned low SAD profiles by Basset, calling into question their causal role.

Furthermore, if the genome sequenced in the original experiment included the inaccessible allele, there will be no indication that accessibility is relevant to the SNP. Basset readily identifies these gain-of-function mutations, as demonstrated for rs4409785 (Figure 5). As statistical searches for influential variants in human populations continue, more rare variants will be uncovered. Especially for these rare variants, the functional genomics experiment will be unlikely to have been performed in the necessary genetic background. This growing trend makes methods like Basset even more important if we are to interpret these variants.

With Basset, a researcher can perform a single sequencing assay in their cell type of interest and simultaneously learn that cell's chromatin accessibility code and annotate every mutation in the genome with its influence on present accessibility and latent potential for accessibility. By leveraging large-scale public data, one can train highly accurate models on common computational hardware. Researchers continue to discover noncoding variants in human populations with phenotypic effects, and such annotation will be indispensable for interpreting how those variants function. As the tide of functional genomics data continues to flow, novel machine learning approaches such as deep convolutional neural networks have great power to aid this goal.

Methods

DNaseI hypersensitivity data

We downloaded DNaseI-seq peak BED format files for 125 cell types from the ENCODE project and 39 cell types from the Epigenomics Roadmap project. For both sets, the previous groups called peaks using the HotSpot algorithm and performed a simulation procedure to establish a set with 1% false discovery rate.

To merge the peaks into one set, we first extended each one from its midpoint to 600 bp. We greedily merged peaks based on their distance to an adjacent peak until no peaks overlapped by more than 200 bp. During a merger of peaks, we specified the activity of the new peak as the union of the sets of active cell types for each individual peak. We specified the new limits by extending from a weighted average of the two peak midpoints, weighted by the number of cells each individual peak was active in. This produced a set of 2,071,886 peaks of which 4.1-19.0% (median 8.2%) were active in the individual cell types.

Deep convolutional neural network

Deep convolutional neural networks (CNNs) are a variant of deep neural network that are specifically parameterized to take advantage of known spatial structure. They were originally developed to recognize handwritten digits in images (LeCun et al. 1998). Convolutional networks have since become the gold standard for numerous image analysis tasks (Krizhevsky et al. 2012; Szegedy et al. 2014). Recently, convolutional networks have been modified for use within natural language processing and text analysis by applying a one dimensional convolution temporally over a sequence (Hu et al. 2014; Zhang and LeCun 2015).

We implemented a deep convolutional neural network (CNN) using Torch7 (http://torch.ch). Initially, we map the DNA sequence to 4 rows of binary variables representing the presence or absence of an A, C, G, and T at each nucleotide position. The first convolutional layer of the network scans position weight matrices across the sequence (Figure 1). The matrix weights are parameters learned from the data. These are typically referred to as filters in the CNN literature. After convolving the matrix across the sequence, we applied a rectified linear ReLU nonlinearity ($f(x) = \max(0,x)$), which has been found helpful in avoiding the vanishing gradient problem that plagued early deep learning research (LeCun et al. 1998; Nair and Hinton 2010).

Subsequent convolutional layers operate on the outputs of these filters across the sequence. We “pool” adjacent positions by taking the max from a small window after each nonlinearity to reduce the number of parameters and achieve shift invariance. After three convolutional layers, we placed two standard fully connected artificial neural network hidden layers and a final fully connected Sigmoid transformation to 164 outputs, representing activity in each cell type. We trained to minimize the binary cross entropy loss function, summed over these 164 outputs.

We trained the network using the RMSprop algorithm for stochastic gradient descent (Tieleman and Hinton 2012). We randomly initialize the parameters before training. During training, the network computes predicts for small batches of sequences. We compare these predictions to the true experimental measurements using the loss function. We then update the model parameters to improve those predictions and move on to the next batch. Over time, the model begins to recognize specific sequence motifs indicative of accessibility and project this

recognition through the network to the cell predictions. We continue training until accuracy ceases to increase on a held out validation set.

The user must specify the number of each type of layer, number of filters per convolution layer, filter sizes, pooling widths, fully connected layer units, and numerous regularization and training optimization parameters. We experimented with various model architectures and hyper-parameter settings using Bayesian optimization, implemented in the package Spearmint (available from https://github.com/HIPS/Spearmint) (Snoek et al. 2012). The top performing architecture is depicted in Supplementary Figure M. Importantly, we apply batch normalization after every layer, which substantially stabilized training optimization (Ioffe and Szegedy 2015).

Source code implementing all steps—date preprocessing, training, and downstream analysis—is available in the package Basset from http://www.github.com/davek44/Basset.

Training and test datasets

From the 2,071,886 total sites, we reserved 71,886 for testing and 70,000 for validation, leaving 1,930,000 for training. We performed optimization directly on the training set. We used the validation set for "early stopping" after 12 epochs of unimproved validation loss and Bayesian optimization. We performed all assessment and analysis on the test set.

gkm-SVM

We downloaded gkm-SVM v1.3 from *http://www.beerlab.org/gkmsvm/* (Lee et al. 2015; Ghandi et al. 2014). Because the code computes the full Gram matrix, we could only feasibly train on a 100k subsample of the full dataset. For each cell type, we downsampled the inactive sequences to match the number of active sequences. Though Basset easily handles imbalanced datasets, we found the natural imbalances of this DHS dataset significantly decreased gkm-SVM accuracy. Using default options, gkm-SVM required 16 days to train and test on 50 randomly selected cell types.

Motif comparison

We identified the likely binding protein for Basset-learned first convolution layer filters by querying the CIS-BP motif database (Weirauch et al. 2014) using the TomTom search tool (Gupta et al. 2007) and requiring an FDR q-value < 0.1.

Comparison of Basset predictions to PhyloP

We used Basset to compute loss and gain scores for every nucleotide. We compute the loss score as the predicted activity with the reference nucleotide subtracted by the minimum predicted activity after mutating the position to the alternative three nucleotides. We compute the gain score as the maximum predicted activity after mutation subtracted by the reference nucleotide activity. To ask whether these scores have a significant statistical relationship with nucleotide conservation, we required a method to distill the 328 scores per nucleotide down to one to compare to PhyloP. We applied a linear regression model with ridge penalty, training on 80% of the nucleotides and testing on the remaining 20%. We limited the analysis to nonrepetitive regions where more confident PhyloP statistics can be assigned and to the center 100 bp of the DHSs, which Basset learned to emphasize.

PICS

We downloaded 8741 PICS SNP annotations from the supplement of the authors' manuscript (Farh et al. 2015). For SAD profile comparison, we focused on an unquestionably noncoding set of 7252 SNPS by removing all SNPs linked to a SNP in a protein coding gene. Without sufficient training data to learn weights for the various cell types studied, we resorted to comparing the SAD profile means to the PICS causal SNP probabilities.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

The authors acknowledge Abbie Groff, David Hendrickson, Marta Mele, Stephanie Fine Sasse, Chinmay Shukla, and Michael Ziller for feedback on the manuscript. DK was supported by NIH K25 award ES022984-02.

References

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotech* **33**: 831–838.

Arnold P, Schöler A, Pachkov M, Balwierz PJ, Jørgensen H, Stadler MB, van Nimwegen E, Schübeler D. 2013. Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res* **23**: 60–73.

Bengio Y, Courville A, Vincent P. 2013. Representation Learning: A Review and New Perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **35**: 1798–1828.

Benveniste D, Sonntag H-J, Sanguinetti G, Sproul D. 2014. Transcription factor binding predicts histone modifications in human cell lines. *Proc Natl Acad Sci U S A* **111**: 13367–13372.

Biddie SC, John S, Sabo PJ, Thurman RE, Johnson TA, Schiltz RL, Miranda TB, Sung M-H, Trump S, Lightman SL, et al. 2011. Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding. *Mol Cell* **43**: 145–155.

Bird AP, Wolffe AP. 1999. Methylation-Induced Repression— Belts, Braces, and Chromatin. *Cell* **99**: 451–454.

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research* **12**: 2493–2537.

Consortium TEP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH, Zhang MQ. 2006. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A* **103**: 10713–10716.

Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al. 2012. DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**: 390–394.

Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJH, Shishkin AA, et al. 2015. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**: 337–343.

Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M. 2014. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**: 480.

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features ed. Q. Morris. *PLoS Comput Biol* **10**: e1003711.

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, et al. 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**: 900–910.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W. 2007. Quantifying similarity between motifs. *Genome Biol* **8**: R24.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**: 9362–9367.

Hu B, Lu Z, Li H, Chen Q. 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. *Advances in Neural Information Processing Systems*. pp. 2042–2050.

Ioffe S, Szegedy C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167*.

Jin Y, Birlea SA, Fain PR, Ferrara TM, Ben S, Riccardi SL, Cole JB, Gowan K, Holland PJ, Bennett DC, et al. 2012. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nat Genet* **44**: 676–680.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310–315.

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*. pp. 1097–1105.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

LeCun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**: 2278–2324.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47**: 955–961.

Levo M, Segal E. 2014. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics* **15**: 453–468.

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al. 2015. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**: 1012–1025.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190–1195.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotech* **30**: 271–277.

Nair V, Hinton GE. 2010. Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning.*

Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, et al. 2013. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**: 376–381.

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotech* **27**: 1173–1175.

Pignon J-C, Grisanzio C, Geng Y, Song J, Shivdasani RA, Signoretti S. 2013. p63-expressing cells are the stem cells of developing prostate, bladder, and colorectal epithelia. *Proc Natl Acad Sci U S A* **110**: 8105–8110.

Pinello L, Xu J, Orkin SH, Yuan G-C. 2014. Analysis of chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. *Proc Natl Acad Sci U S A* **111**: E344–53.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**: 110–121.

Ray S, Pollard JW. 2012. KLF15 negatively regulates estrogen-induced epithelial cell proliferation by inhibition of DNA replication licensing. *Proc Natl Acad Sci U S A* **109**: E1334–43.

Ritchie GRS, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. *Nat Methods* **11**: 294–296.

Sawcer S, Hellenthal G, Pirinen M, Spencer CCA, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, et al. 2011. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**: 214–219.

Setty M, Leslie CS. 2015. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps ed. Z. Weng. *PLoS Comput Biol* **11**: e1004271.

Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotech* **32**: 171–178.

Snoek J, Larochelle H, Adams RP. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*. pp. 2951–2959.

Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* **16**: 16–23.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. 2014. Going Deeper with Convolutions. *arXiv:1409.4842*.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.

Tieleman T, Hinton G. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *Neural Networks for Machine Learning*. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf (Accessed September 26, 2015).

Voss TC, Hager GL. 2014. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics* **15**: 69–81.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**: 1431–1443.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**: D1001–6.

Whitaker JW, Chen Z, Wang W. 2015. Predicting the human epigenome from DNA motifs. *Nat Methods* **12**: 265–272.

Wilanowski T, Caddy J, Ting SB, Hislop NR, Cerruti L, Auden A, Zhao L-L, Asquith S, Ellis S, Sinclair R, et al. 2008. Perturbed desmosomal cadherin expression in grainy head-like 1-null mice. *EMBO J* **27**: 886–897.

Zhang X, LeCun Y. 2015. Text Understanding from Scratch. *arXiv:1502.01710.*

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods*.

Figure Legends

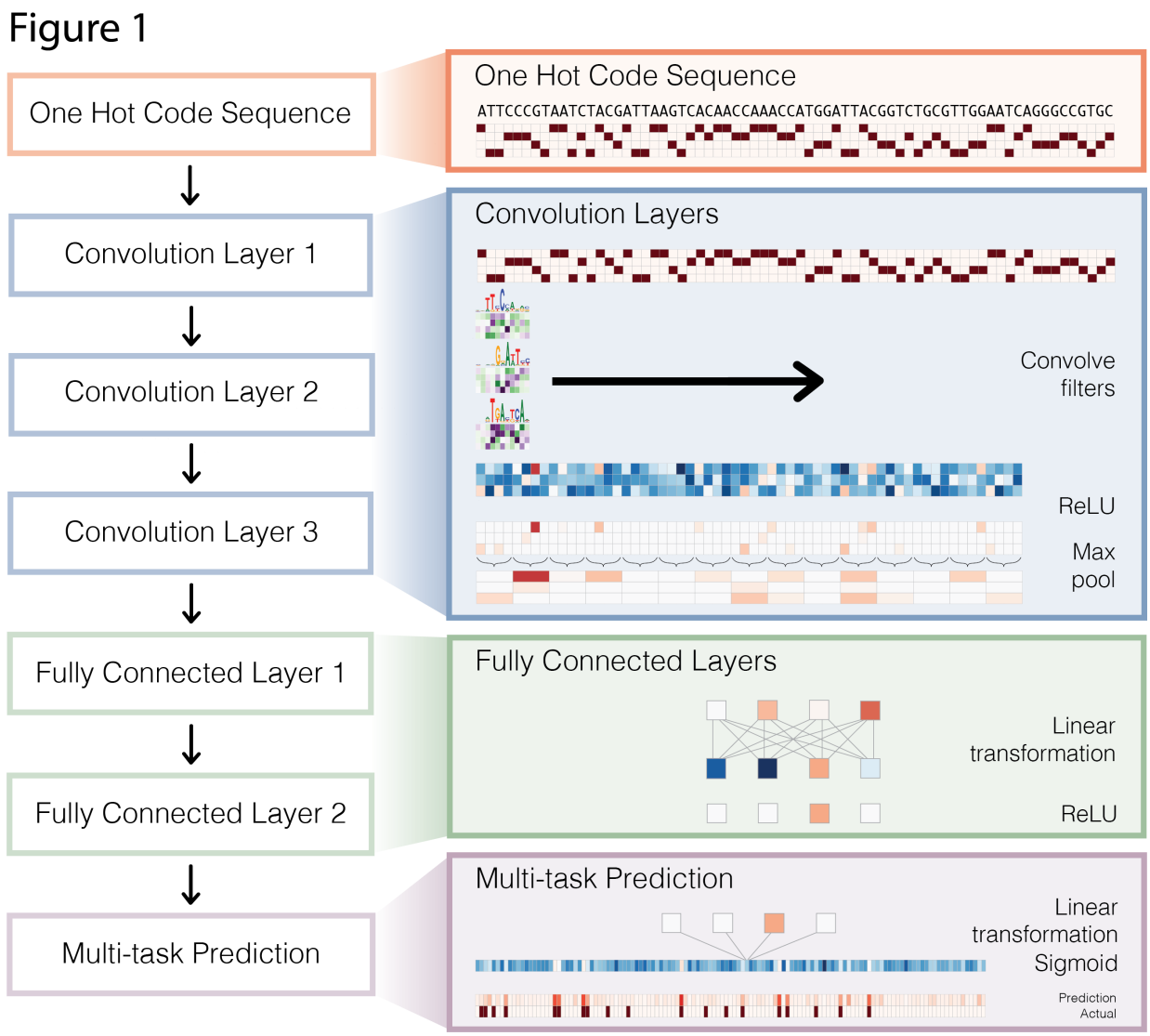


Figure 1 - Deep convolutional neural network for DNA sequence analysis

Basset predicts the cell-specific DNaseI hypersensitivity of sequences. First, we convert the sequence to a "one hot code" representation, where each position has a four element vector with one nucleotide's bit set to one. Convolution layers proceed by scanning weight matrices across the input matrix to produce an output matrix with a row for every convolution filter and a column for every position in the input (minus the width of the filter). We apply a rectified linear unit (ReLU) nonlinear transformation to the convolution output and pool by taking the max across a window of adjacent positions. The first convolution layer operates directly on the one hot coding of the input sequence, making the convolution filters akin to the common bioinformatics tool position weight matrices. Subsequent convolution layers consider the

orientations and spatial distances between patterns recognized in the previous layer. Fully connected layers perform a linear transformation of the input vector and apply a ReLU. The final layer performs a linear transformation to a vector of 164 elements that represents the target cells. A sigmoid nonlinearity maps this vector to the range 0-1, where the elements serve as probability predictions of DNaseI hypersensitivity, to be compared via a loss function to the true hypersensitivity vector.

Figure 2

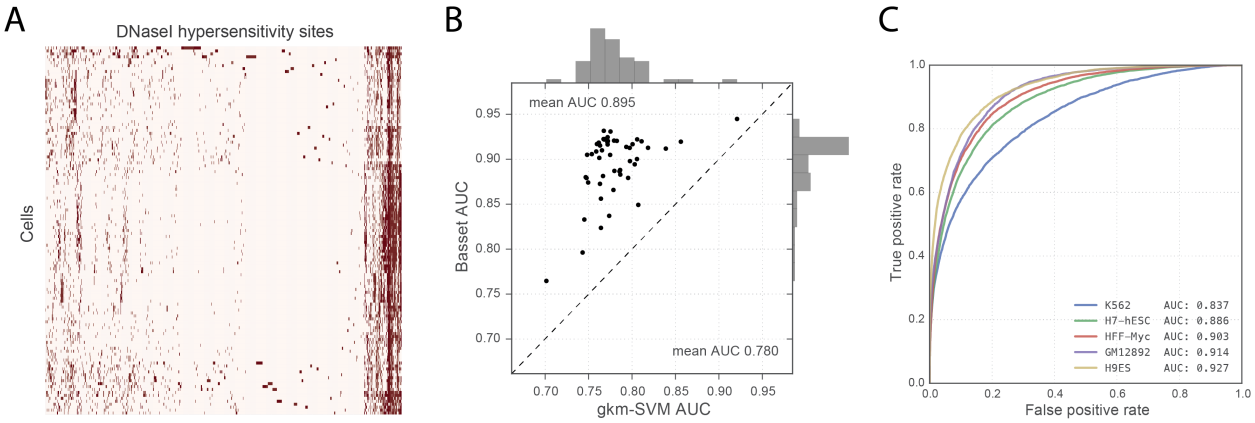


Figure 2 - Basset accurately predicts cell-specific DNA accessibility.
(A) The heat map displays hypersensitivity of 2 million DHSs mapped across 164 cell types. We performed average linkage hierarchical clustering using Euclidean distance to both cells and sites. (B) The scatter plot displays AUC for all cell types achieved by Basset and the state-of-the-art approach gkm-SVM, which uses support vector machines. (C) The ROC curves display the Basset false positive rate versus true positive rate for five cells, selected to represent the .05, .33, .50, .67, and .95 quantiles of the AUC distribution.

Figure 3

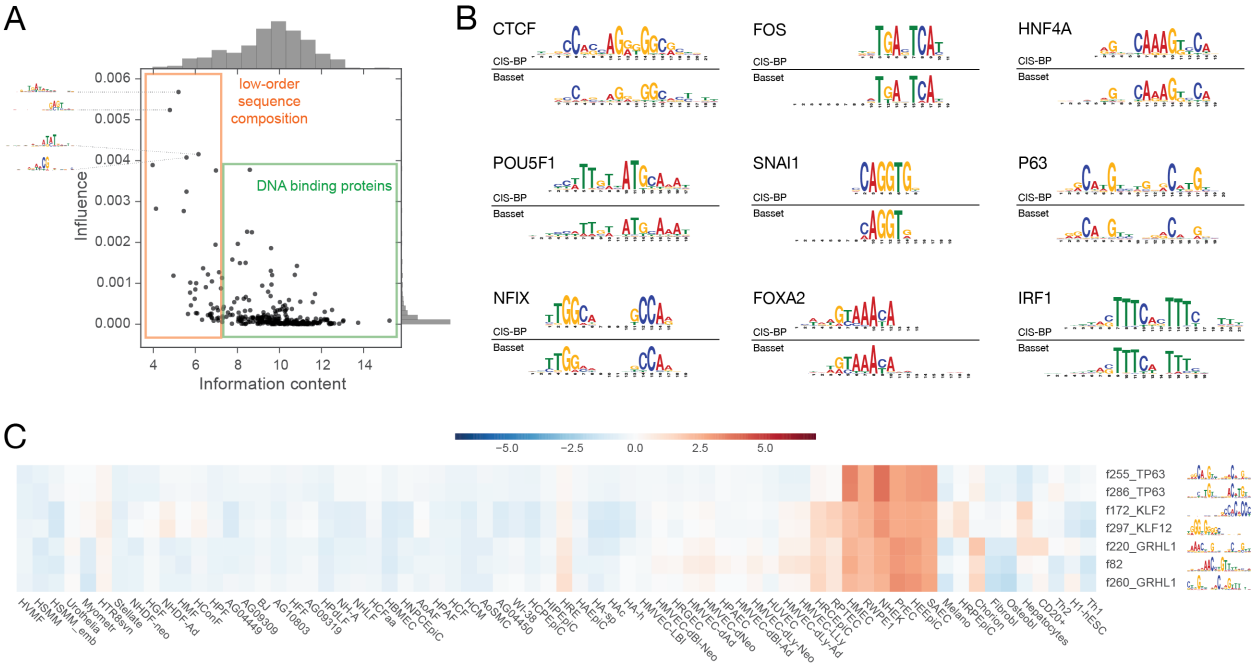


Figure 3 - Basset initial convolutional layer discovers known and novel sequence motifs (A) In the scatter plot, the x-axis describes the information content for the PWMs represented by the 300 first layer convolution filters. The y-axis describes an influence score, which we compute by setting all output from the filter to its mean (nullifying the filter) and taking the L2 norm of the vector of accessibility prediction changes over all cells. (B) At a q-value threshold of 0.1, the Tomtom motif comparison tool annotated 45% of the filters to known TF motifs in the human CIS-BP database, including a variety of exceptionally accurate alignments shown here.

Figure 4

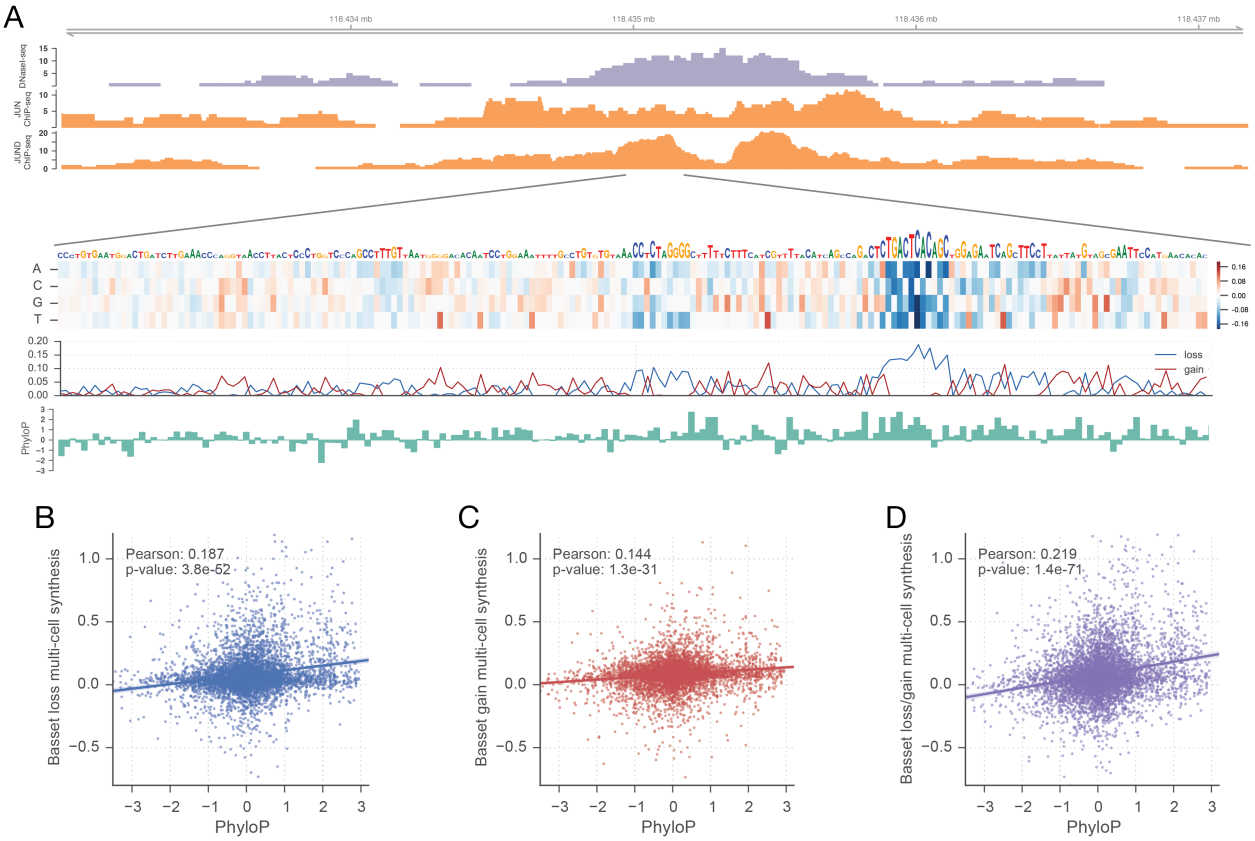


Figure 4 - in silico saturated mutagenesis for DNaseI hypersensitivity

(A) We used Basset to predict the effect of every mutation on the accessibility of the region chr9:118,434,976-118,435,175 in H1-hESCs. The heat map displays the accessibility predictions for mutated sequences. Each column corresponds to a position in the sequence. Each row represents mutation to the corresponding nucleotide. We normalize every prediction by subtracting the prediction for the true sequence. In the line plot below, loss scores measure the maximum decrease among all mutations from the true nucleotide. Gain scores measure the maximum increase. We drew nucleotides to be proportional to the loss score, beyond a minimum height. At this locus, the model highlights the TGASTCA motif of the AP1 complex. ChIP-seq of JUN and JUND in H1-hESCs confirm binding of the complex. The bound motif displays high conservation according to PhyloP. (B) Genome-wide, loss scores had a strong relationship with PhyloP (see Methods). (C) Gain scores alone had a weaker relationship, but the (D) combination achieved the strongest relationship.

Figure 5

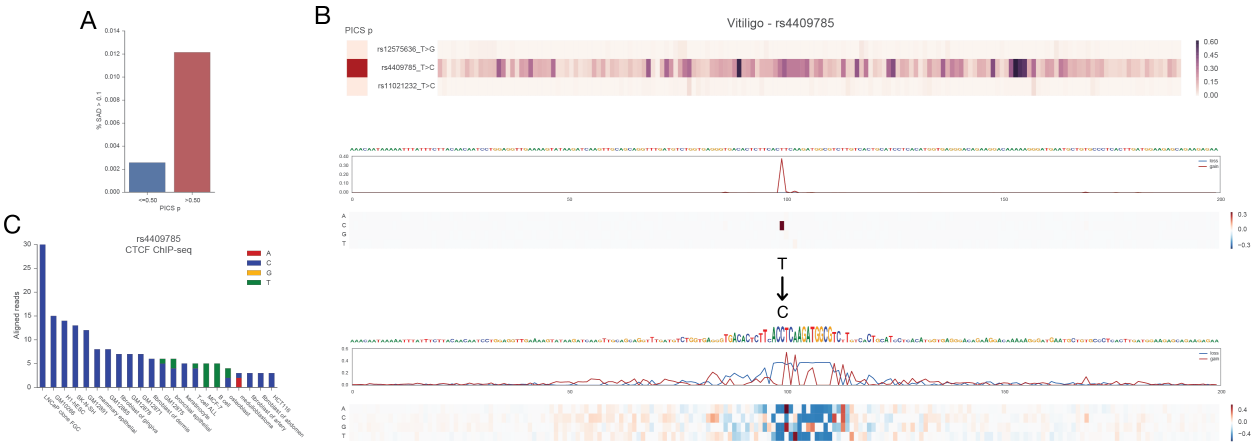
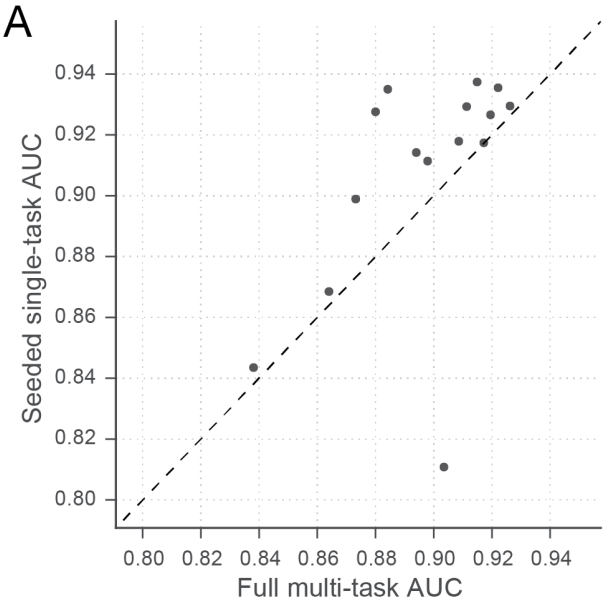


Figure 5 - SNP accessibility difference scores enable genomic variant interpretation

(A) Basset assigned greater scores to GWAS SNPs that PICS determined to have greater than 50% probability of being the causal SNP using population fine mapping data. The bars measure the proportion of SNPs assigned a SAD profile mean across all cell types >0.1. (B) We annotated rs4409785 among the highest SAD scores, in agreement with the PICS view of this haplotype block. Basset predicts the more common T allele to be completely dormant, but the region transforms with the C allele into a site deemed by Basset to have very high accessibility due to a CTCF binding site. (C) CTCF ChIP-seq confirms the allele-specificity of CTCF at this site. Hundreds of cells indicate no binding, and the few cells where CTCF binds sequence the C allele. We plotted only cells with >=3 reads aligned to the site.

Figure 6

A



B

	GPU	CPU
Full multi-task	85 h	-
Seeded single-task	18 m	6 h 37 m

Figure 6 – Basset leverages large-scale public data to inform additional dataset learning
The scatter plot shows AUC for 15 datasets achieved by the full model trained on all 164 cell types on the x-axis and AUC achieved by a procedure to simulate studying that dataset alone on the y-axis. To study the dataset alone, we pre-train a model on 149 cells (after removing these 15), seed training of the additional cell with that model's parameters, and perform a single training pass through the new data. This rapid procedure was effective for all but one dataset (HRCEpiC, renal cortical epithelial cells), for which multi-task training with the many other similar epithelial cells was critical. The AUC improvement for many cells suggests that our full model may benefit from increased capacity or decreased regularization.