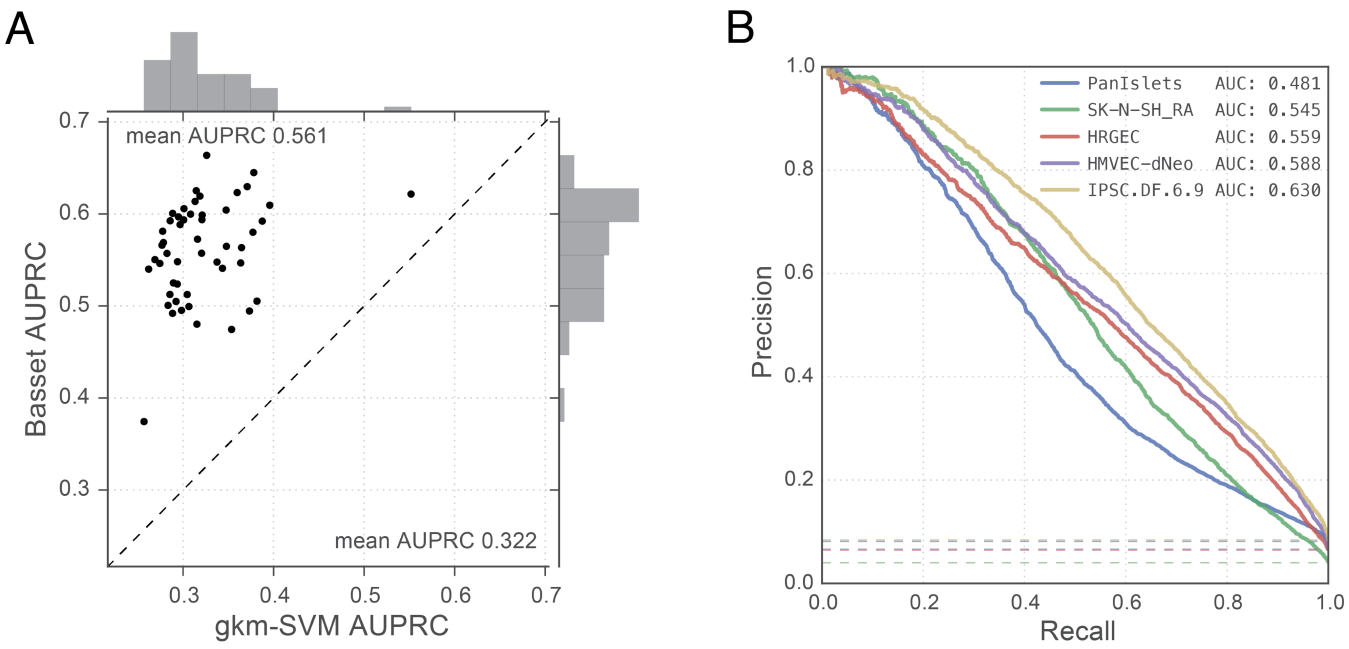
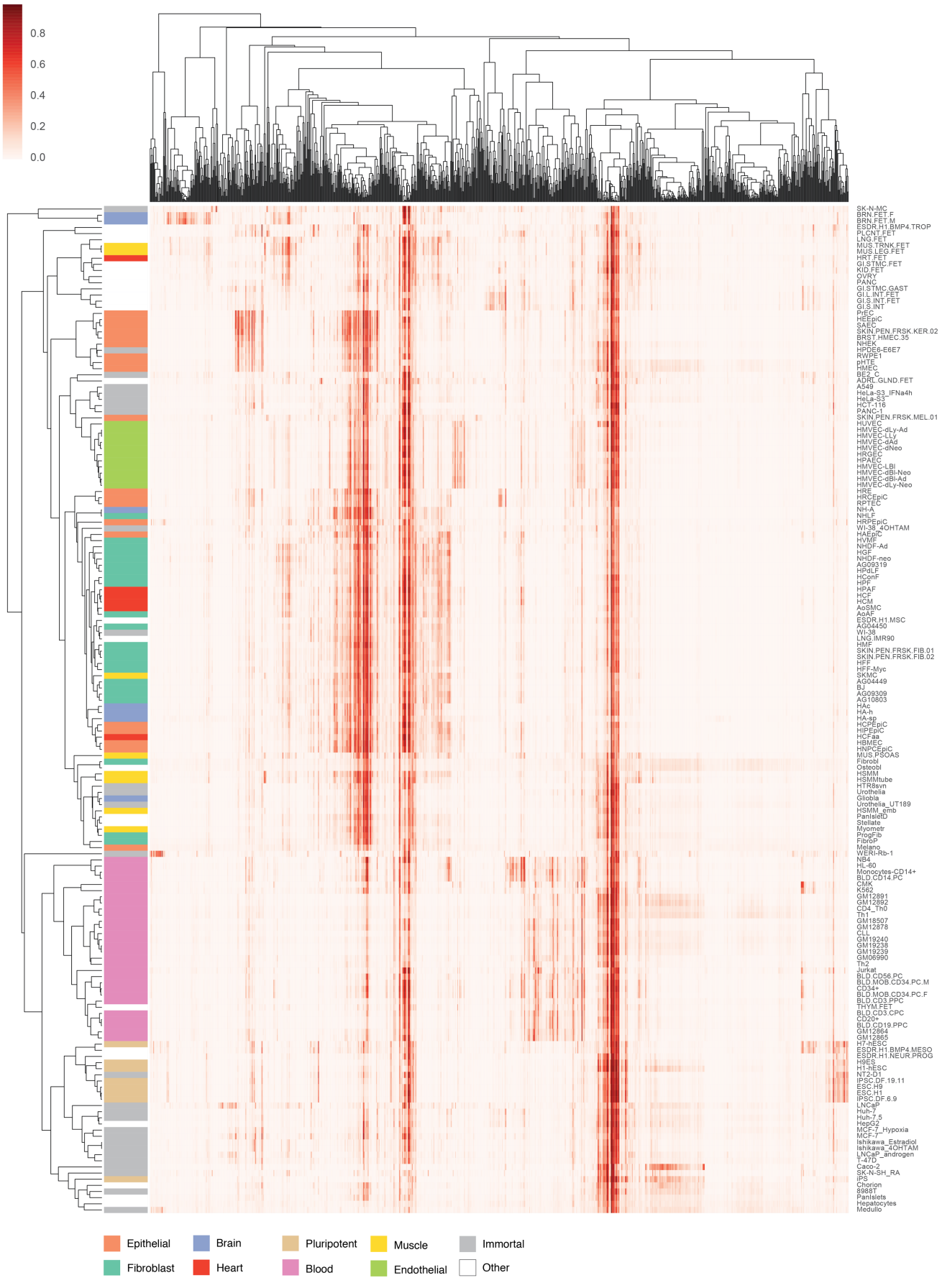


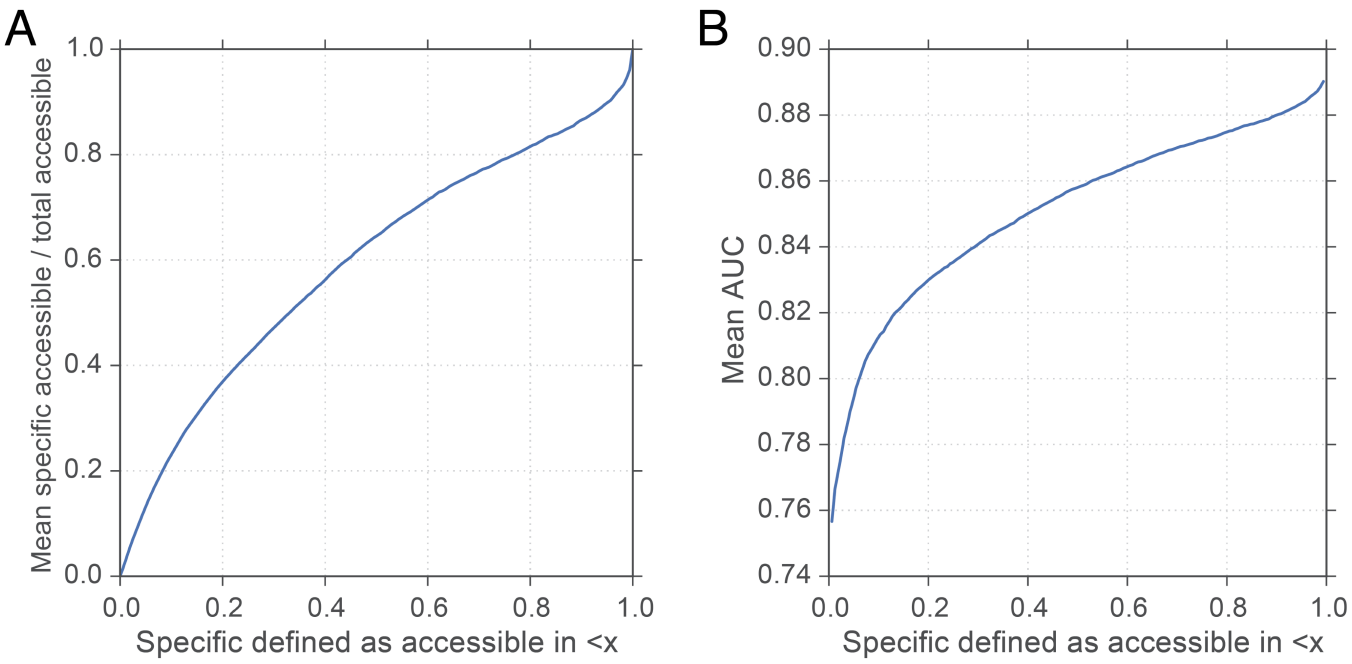
Supplementary Figure 1

(A) The scatter plot displays the area under the precision-recall curve AUPRC for 50 randomly selected cell types achieved by Basset and the state-of-the-art approach gkm-SVM, which uses support vector machines. (B) The curves display Basset's recall versus precision for five cells, which were selected to represent the .05, .33, .50, .67, and .95 quantiles of the AUPRC distribution. Dotted lines represent a null model of random guessing.


0.8
0.6
0.4
0.2
0.0
SK-N-MC
BRN.FET.F
BRN.FET.M
ESDR.H1.BMP4.TROP
PI.CNT.FET
LNG.FET
MUS.TRNK.FET
MUS.LEG.FET
HRT.FET
GI.STMC.FET
KID.FET
OVRY
PANC
GI.STMC.GAST
GI.LINT.FET
GI.S.INT.FET
GI.S.INT
PrEC
HEEpiC
SAEC
SKIN.PEN.FRSK.KER.02
BRST.HMEC.35
NHEK
HPDE6-E6E7
RWPE1
phTE
HMEC
BE2_C
ADRL.GLND.FET
A549
HeLa-S3_IFNa4h
HeLa-S3
HCT-116
PANC-1
SKIN.PEN.FRSK.MEL.01
HUVEC
HMVEC-dLy-Ad
HMVEC-LLY
HMVEC-dAd
HMVEC-dNeo
HRGEC
HPAEC
HMVEC-LBl
HMVEC-dBl-Neo
HMVEC-dBl-Ad
HMVEC-dLy-Neo
HRE
HRGEpiC
RPTEC
NH-A
NHLF
HRPEpiC
WI-38_4OHTAM
HAEpiC
HVMF
NHDF-Ad
HGF
NHDF-neo
AG09319
HPdLF
HConF
HPF
HPAF
HCF
HCM
AoSMC
AoAF
ESDR.H1.MSC
AG04450
WI-38
LNG.IMR90
HMF
SKIN.PEN.FRSK.FIB.01
SKIN.PEN.FRSK.FIB.02
HFF
HFF-Myc
SKMC
AG04449
BJ
AG09309
AG10803
HAc
HA-h
HA-sp
HCPEpiC
HIPEpiC
HCFaa
HBMEC
HNPCEpiC
MUS.PSOAS
Fibrobl
Osteobl
HSMM
HSMMtube
HTR8svn
Urothelia
Gliobla
Urothelia_UT189
HSMM_emb
PanIsletD
Stellate
Myometr
ProgFib
FibroP
Melano
WERI-Rb-1
NB4
HL-60
Monocytes-CD14+
BLD.CD14.PC
CMK
K562
GM12891
GM12892
CD4_Th0
Th1
GM18507
GM12878
CLL
GM19240
GM19238
GM19239
GM06990
Th2
Jurkat
BLD.CD56.PC
BLD.MOB.CD34.PC.M
CD34+
BLD.MOB.CD34.PC.F
BLD.CD3.PPC
THYM.FET
BLD.CD3.CPC
CD20+
BLD.CD19.PPC
GM12864
GM12865
H7-hESC
ESDR.H1.BMP4.MESO
ESDR.H1.NEUR.PROG
H9ES
H1-hESC
NT2-D1
iPSC.DF.19.11
ESC.H9
ESC.H1
iPSC.DF.6.9
LNCaP
Huh-7
Huh-7.5
HepG2
MCF-7_Hypoxia
MCF-7
Ishikawa_Estradiol
Ishikawa_4OHTAM
LNCaP_androgen
T-47D
Caco-2
SK-N-SH_RA
iPS
Chorion
8988T
Panislets
Hepatocytes
Medullo
Epithelial
Fibroblast
Brain
Heart
Pluripotent
Blood
Muscle
Endothelial
Immortal
Other

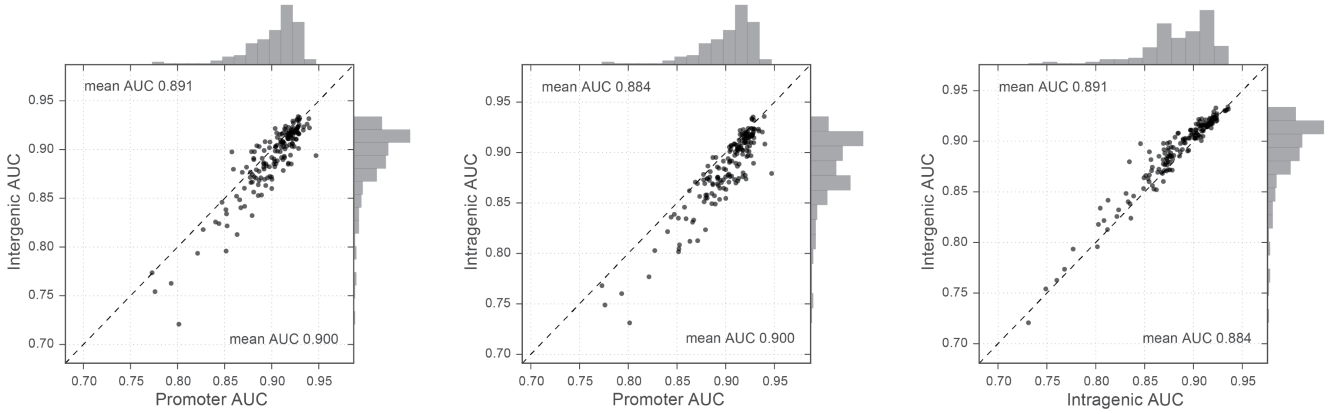
Supplementary Figure 2

We plotted Basset accessibility predictions for 1000 random test sequences in the 164 cell types and performed average linkage hierarchical clustering using cosine distance. We manually annotated cells using ENCODE and Epigenomics Roadmap descriptive tables. Cell and lineage-specific regulatory programs appear as coherent clusters of cell types.



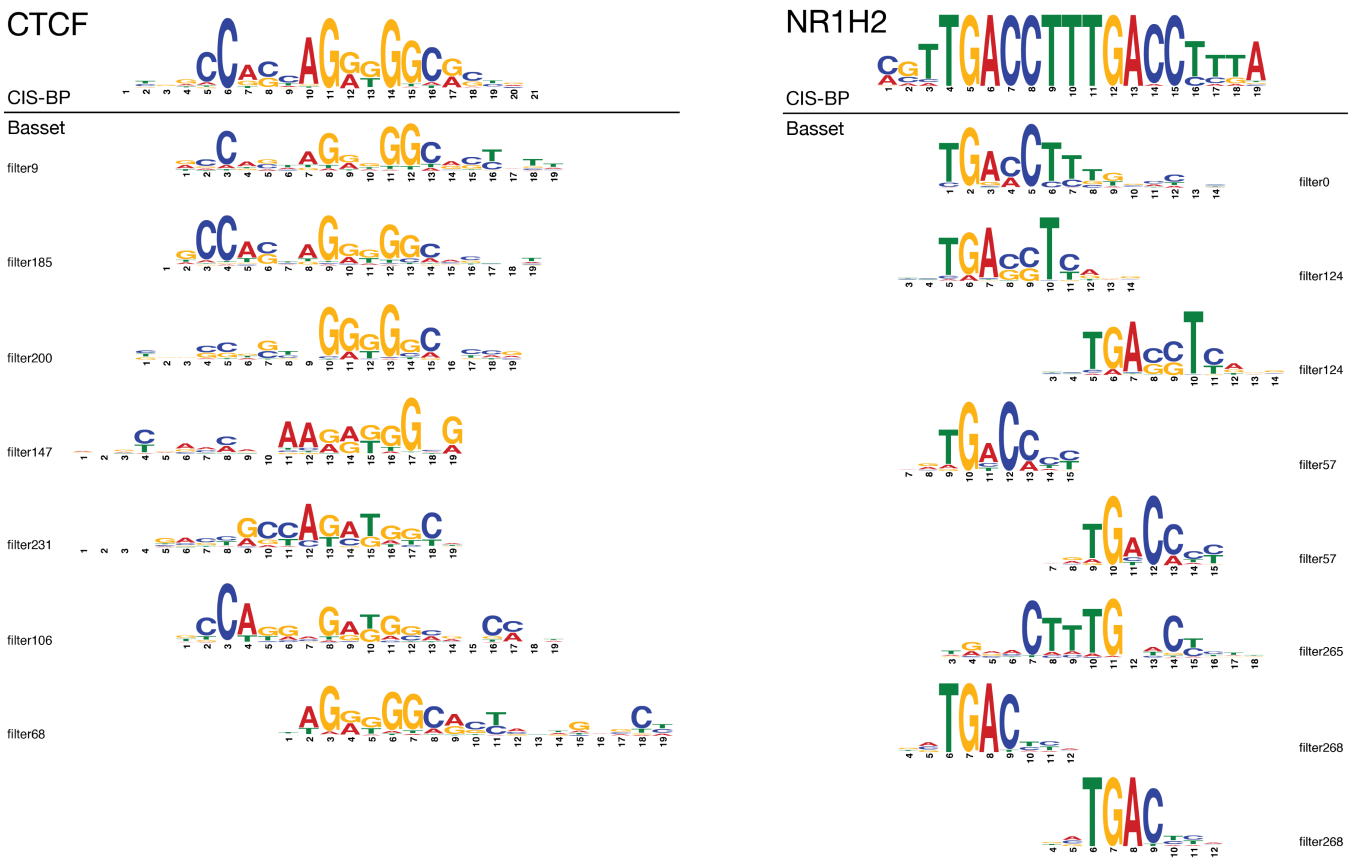
Supplementary Figure 3

One can separate the sites into *specific* versus *constitutive* based on what proportion of cells the site is accessible in. Then in each cell type, some proportion of the accessible sites are *specific* and the rest are *constitutive*. (A) We plotted the average specific proportion across cell types as we vary the threshold to separate *specific* from *constitutive*. I.e. when we consider sites that are accessible in >50% of the cells to be *constitutive*, those *constitutive* sites make up an average of 35% of the accessible sites in each cell type. (B) We can also compute accuracy separately on the two subsets as we vary the threshold. Sites that are only accessible in few cell types are more challenging to predict.



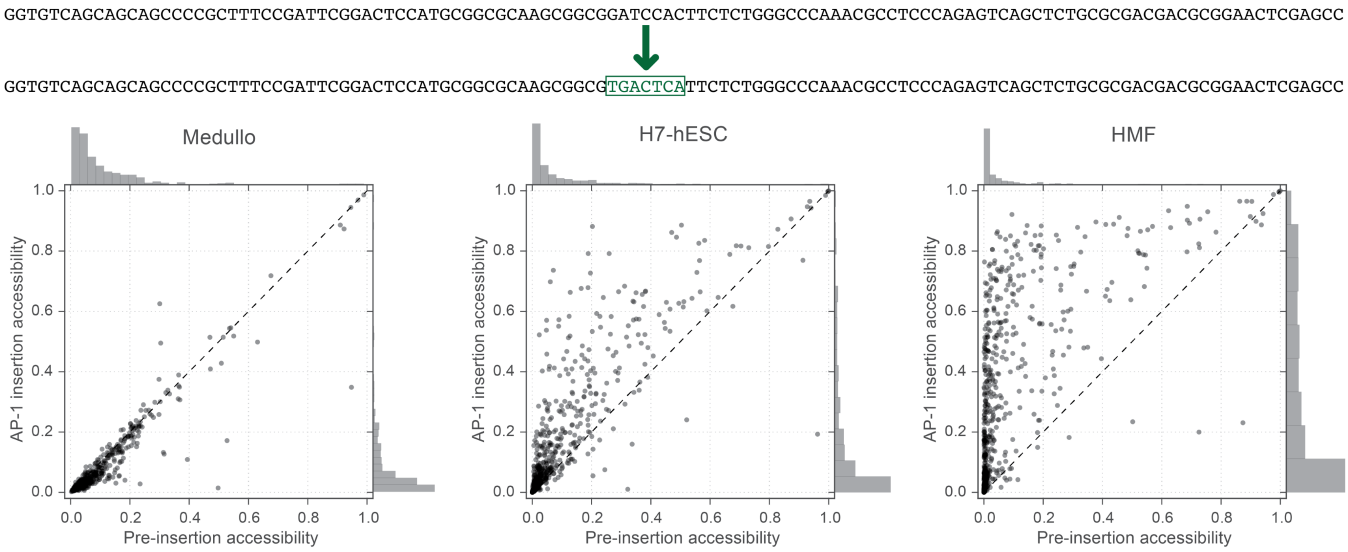
Supplementary Figure 4

We annotated sites as promoter, intragenic, or intergenic using GENCODE v18 gene annotations. In each scatter plot, we compare the AUC for the 164 cell types for test sequences assigned to the labeled annotations. Predictive accuracy was consistent between annotations. Basset predicted promoter accessibility with slightly greater accuracy than other sites and intergenic site accessibility with slightly greater accuracy than intragenic sites. Intragenic accessibility is likely more challenging to predict because of complex interaction with the transcription machinery.



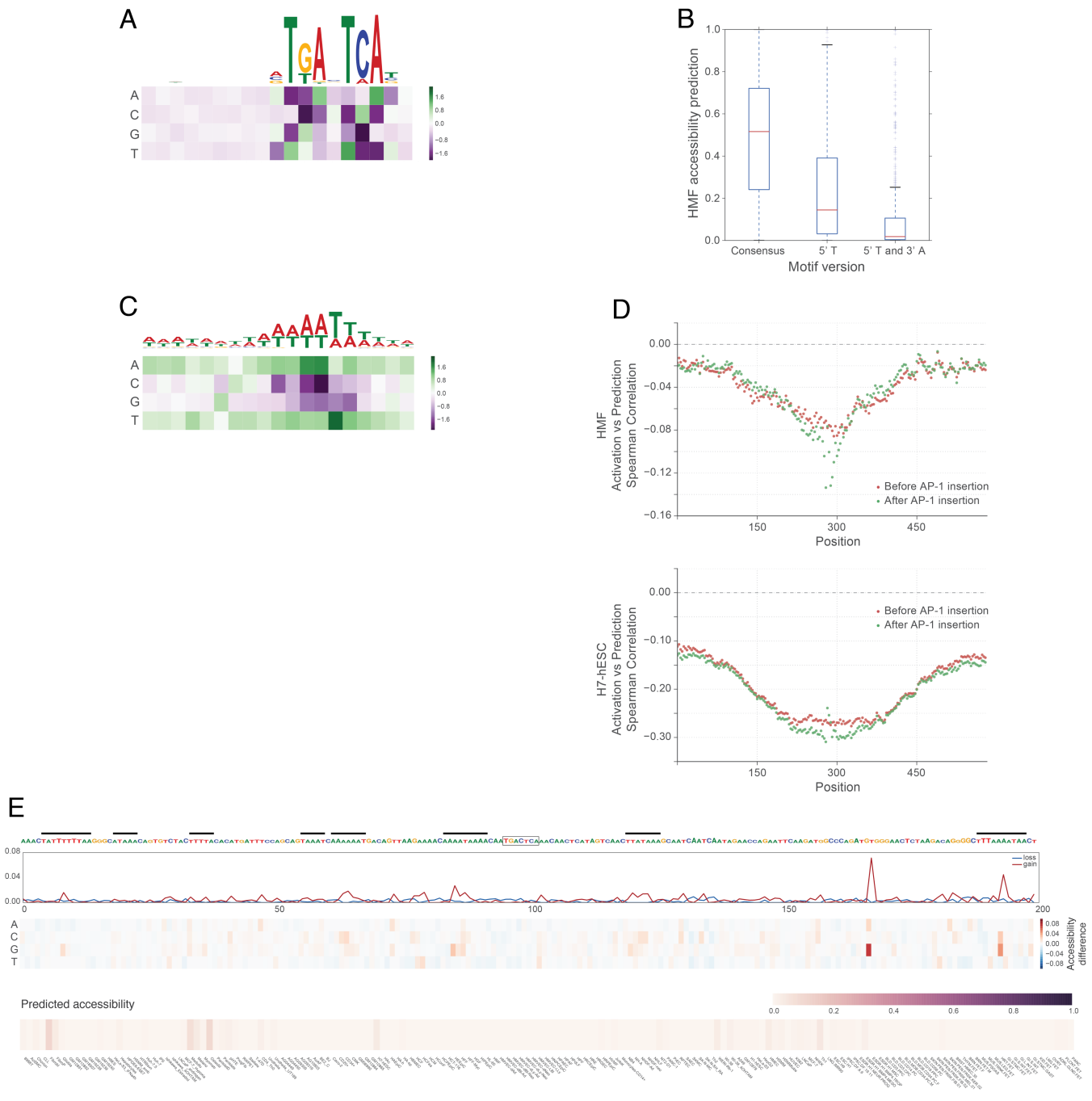
Supplementary Figure 5

On the left, 12 first-layer convolution filters aligned significantly to the CTCF CIS-BP motif, of which the 7 depicted here aligned to the motif in the forward direction. Each filter focuses on different aspects of this complex binding motif. On the right, the CIS-BP motif for NR1H2 comprises two similar pieces. Basset learns each piece individually and re-uses the filters for both halves. It also learns filters that span the junction, e.g. filter0 and filter265.



Supplementary Figure 6

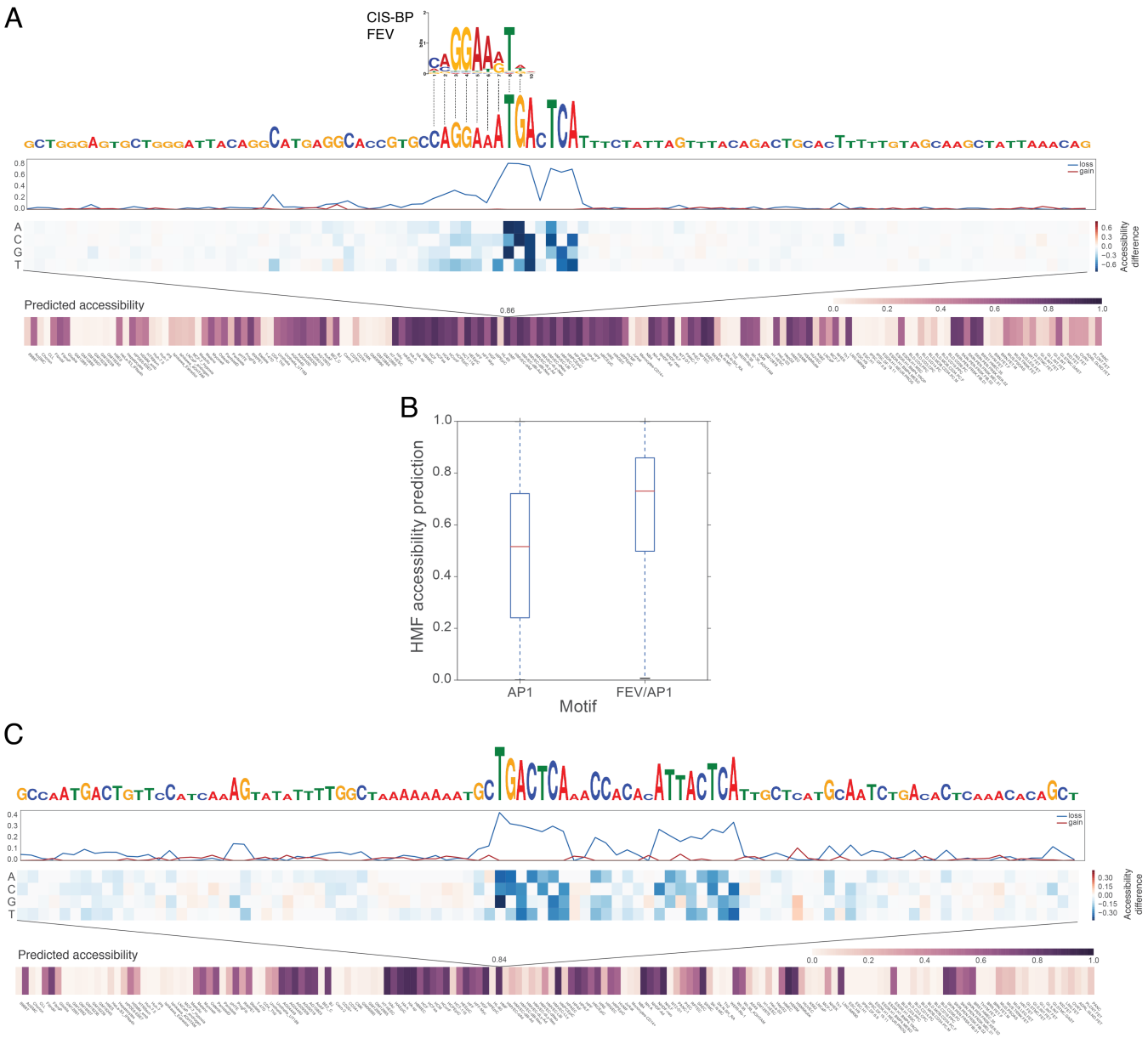
We introduced the AP-1 consensus TGACTCA motif into the center of random test sequences. The 164 cell types span a range of response magnitudes; Medullo, H7-hESC, and HMF cells exemplify a low, medium, and high response respectively to the AP-1 motif. The scatter plots show the accessibility prediction of the sequences before and after inserting the motif. In the few cases where the prediction decreases, the AP-1 motif insertion disrupts a pre-existing functional motif.



Supplementary Figure 7

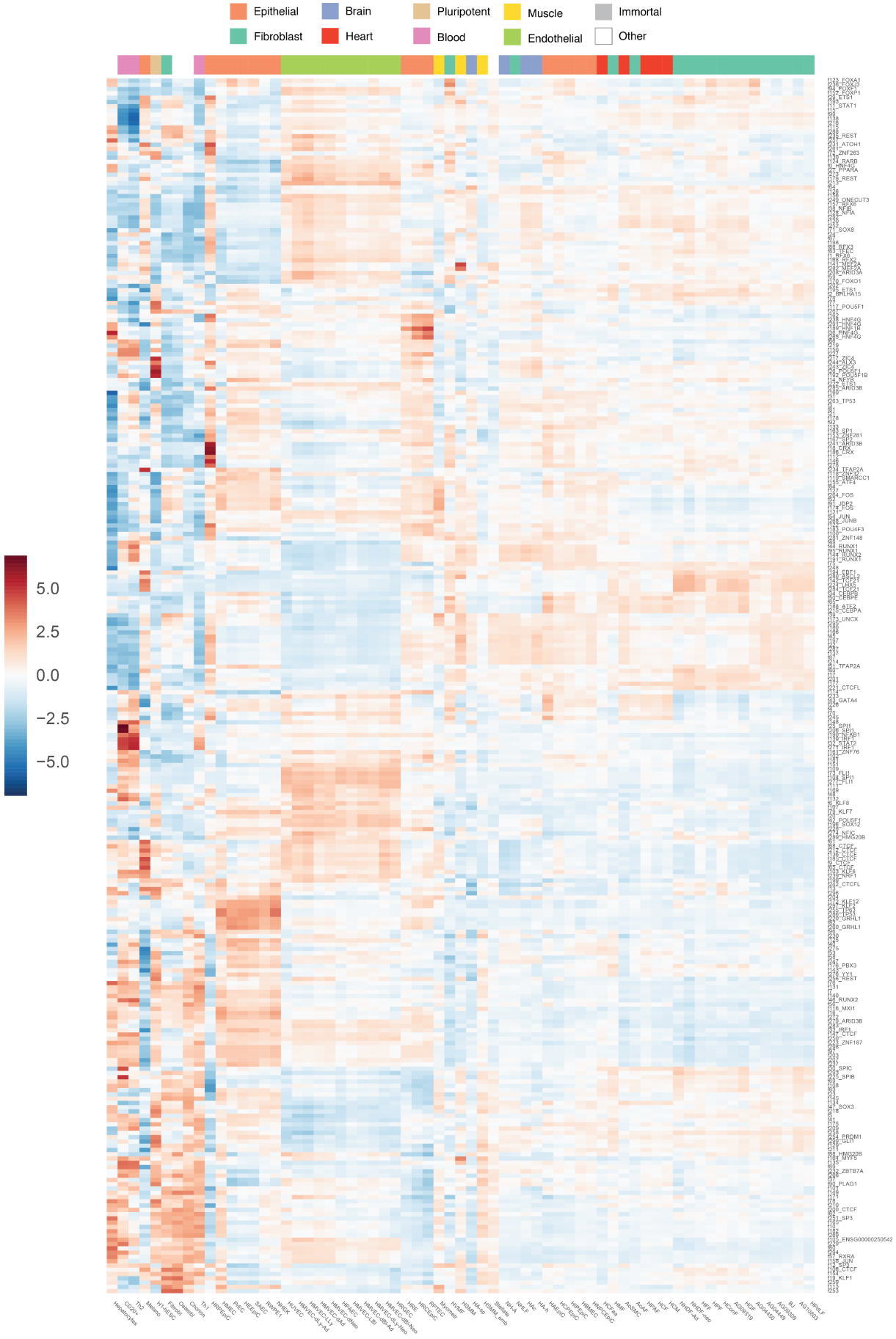
Basset captures various discriminating features of the sequence flanking the primary binding motif. (A) The model learned to strongly prefer TGACTCA motifs that avoid a 5' T and 3' A, which we can see in the weight matrix for filter 91 from the first convolution layer. (B) In HMF cells, presence of the 5' T drastically decreases the distribution of predicted accessibility from mean 0.49 for the full consensus motif to 0.25 with a 5' T. Addition of a 3' A further reduces the predictions to a mean 0.10. (C) One filter captured AT-richness, especially poly-AT stretches. (D) Activation of this filter at positions across the

sequences had significant negative correlation with the ultimate prediction, exemplified by HMF and H7-hES cells. This correlation varied by position, peaking near the center. AP-1 insertion altered this position-dependence, suggesting that the flanking 50-100 nt around the motif were specially considered. (E) A site at Chr6:167500200-167500800 illustrates this phenomenon whereby poly-AT stretches result in low predicted accessibility across cells. We describe the mutagenesis heat maps in Results section “In silico saturation mutagenesis pinpoints nucleotides driving accessibility”.



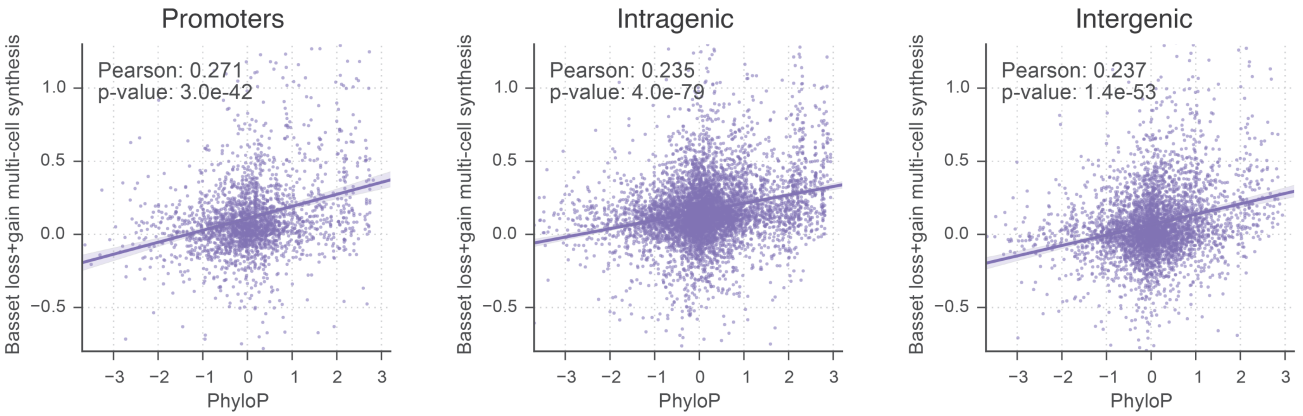
Supplementary Figure 8

Basset captures interactions with additional motifs. We describe the mutagenesis heat maps in Results section “In silico saturation mutagenesis pinpoints nucleotides driving accessibility”. (A) Basset predicts a very high probability of accessibility for a site at Chr20:10763795-10764395 where a motif that matches multiple ETS family TFs, including FEV, occurs directly adjacent to and overlapping the AP-1 motif. (B) Inserting the joint FEV/AP-1 motif shifted the predicted accessibility distribution in HMF cells to mean 0.66 from mean 0.49 with only the AP-1 motif. (C) A site at Chr3:157775355-15777595 illustrates a case where a non-consensus AP-1 motif TTASTCA enhances the accessibility prediction.


Epithelial
Brain
Pluripotent
Muscle
Immortal
Fibroblast
Heart
Blood
Endothelial
Other
5.0
2.5
0.0
−2.5
−5.0

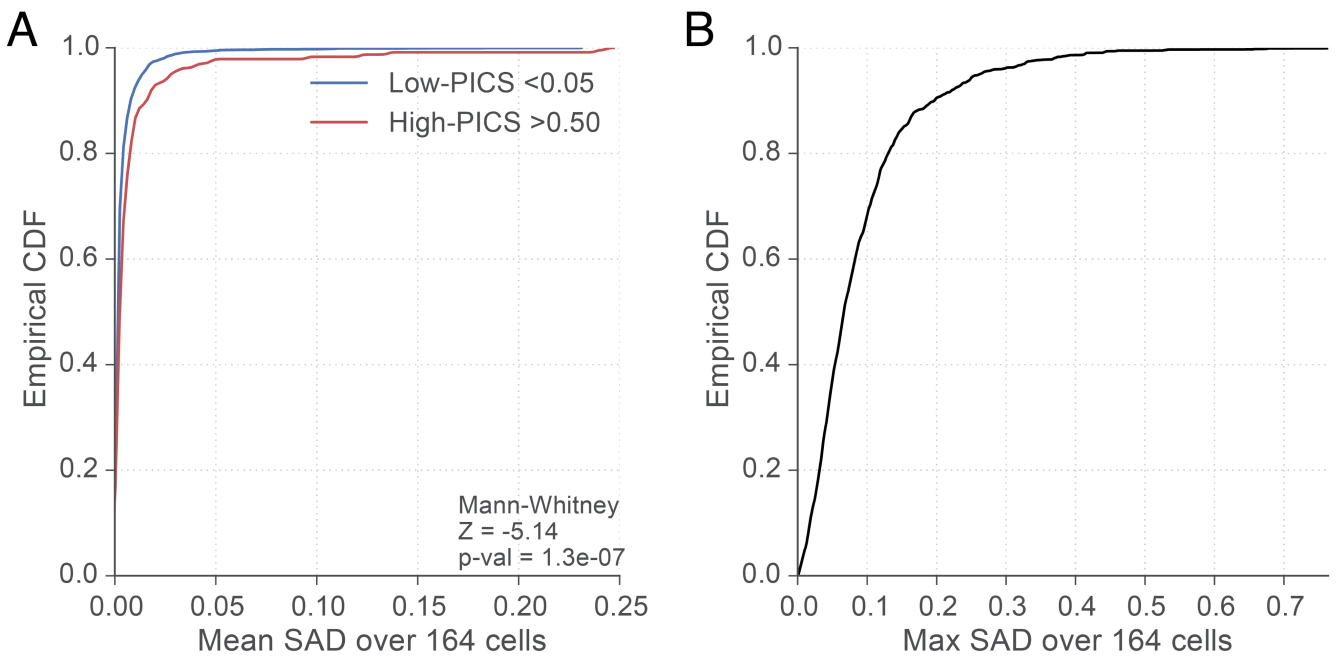
Supplementary Figure 9

The heat map plots a cell-specific influence for all pairs of filters and cells, measured as a normalized change in accessibility in that cell type after setting all output from the filter to its mean. We plotted only primary ENCODE cells for easier visualization and annotated the cells using ENCODE descriptions. We performed average linkage hierarchical clustering using the cosine distance in order to favor direction over magnitude. We annotated each filter with its most significant database motif using Tomtom.



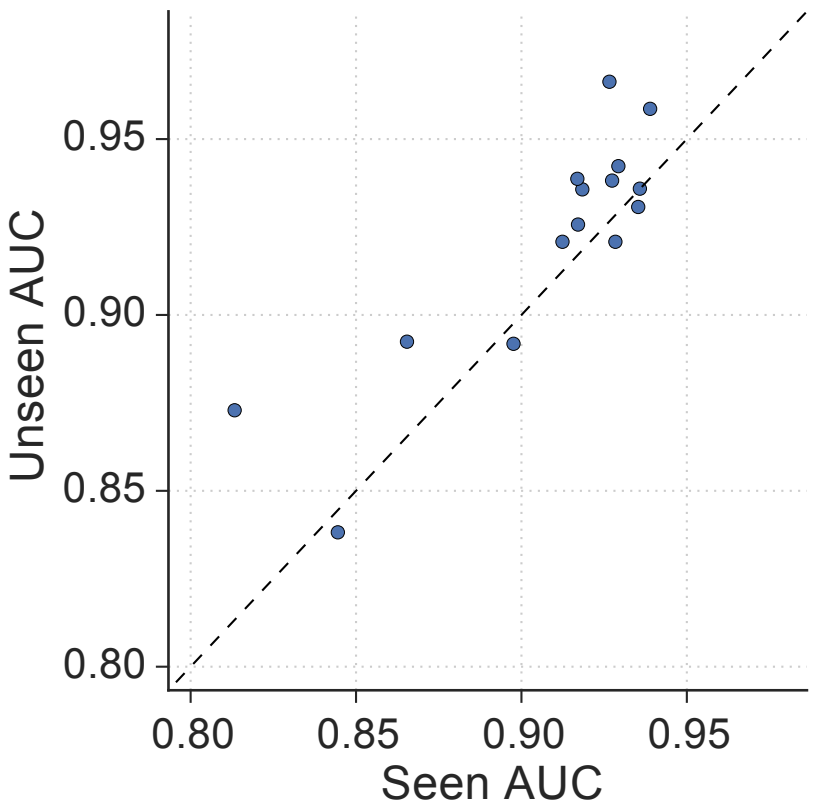
Supplementary Figure 10

Similar to Figure 4D, we plotted PhyloP conservation statistics versus Basset's regressed multi-cell synthesis of loss and gain scores, separated by promoters, intragenic, and intergenic sites. The strong genome-wide correlation was consistent across annotation.



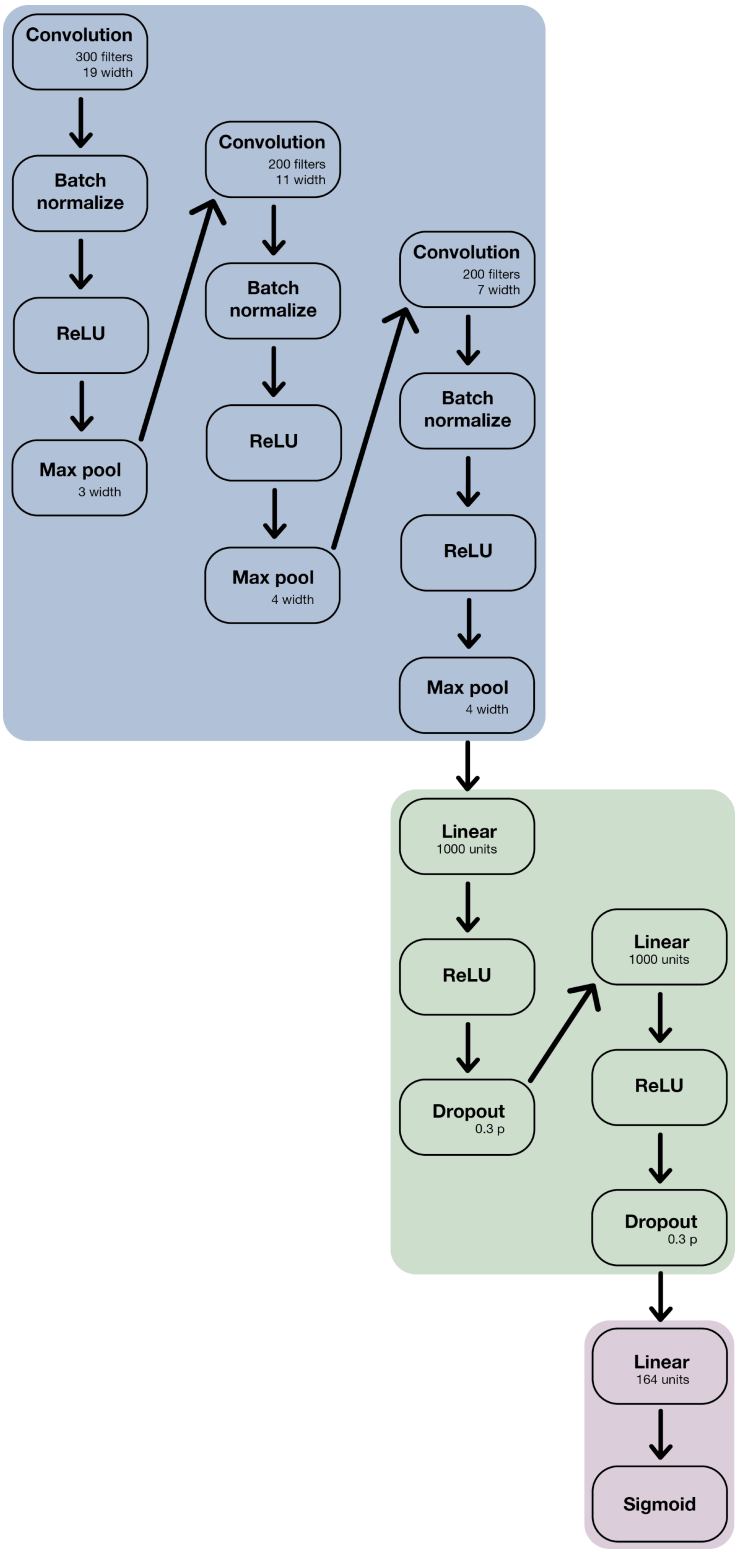
Supplementary Figure 11

(A) On the left, we plot the empirical cumulative distribution function (CDF) of SNPs' mean SNP accessibility difference (SAD) score over the 164 cell types studied for the sets of high-PICS SNPs that are likely causal and low-PICS SNPs that are likely not. High-PICS SNPs have significantly greater evidence of modulating genomic accessibility. (B) On the right, we plot the empirical CDF of the maximum SAD score achieved by any SNP in linkage disequilibrium with the indexed SNP. That is, one can read statistics such as 31% of index SNPs have a SNP in LD for which Basset predicts a >0.1 SAD in some cell type.



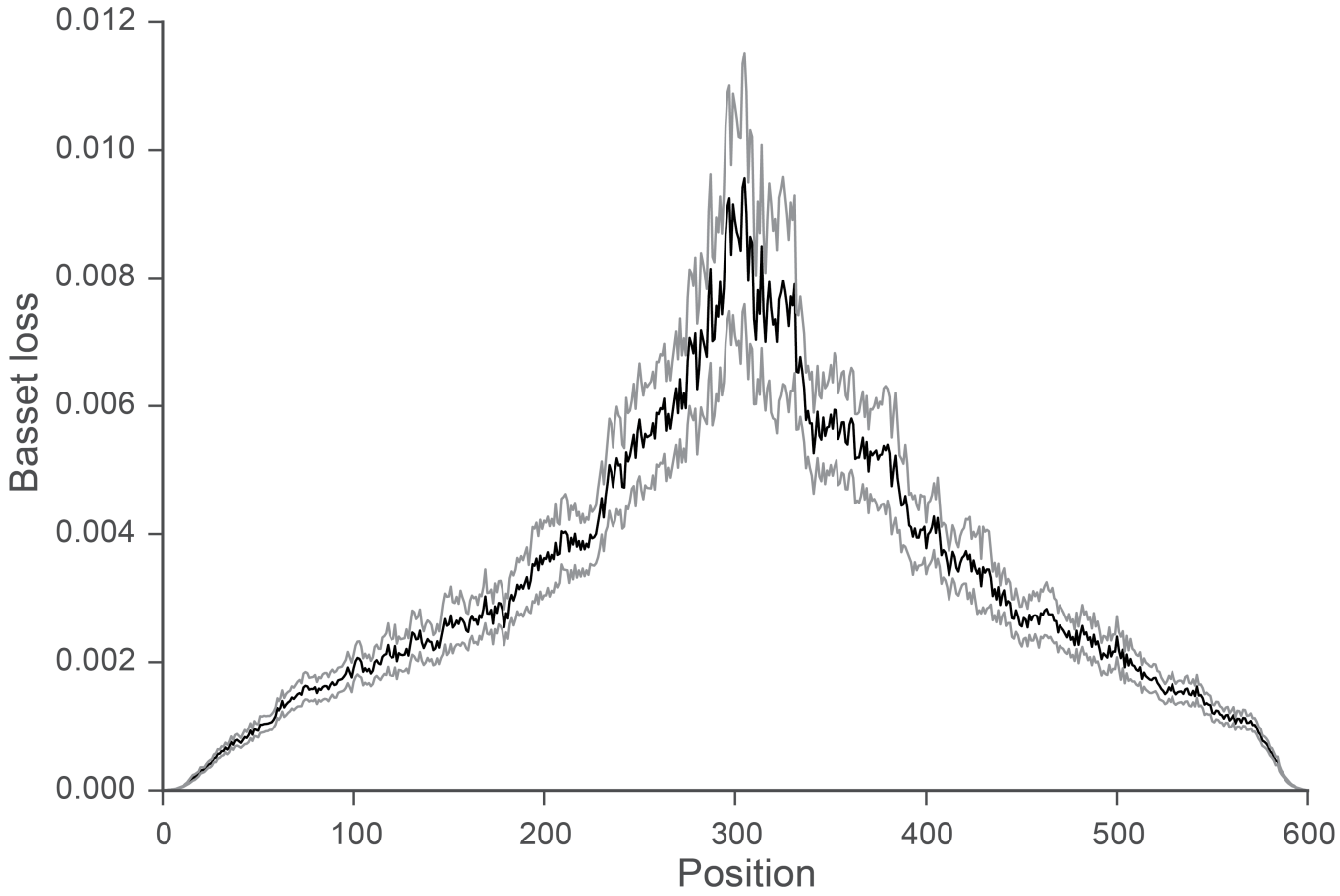
Supplementary Figure 12

For each of the 15 datasets trained individually with the seeded model, we computed AUC separately for accessible sites “seen” before because they overlap one in the “public” dataset upon which the seed model was pre-trained on. AUC is similarly high for novel “unseen” sites and “seen” sites.



Supplementary Figure 13

We depict a detailed view of a top performing model architecture that we studied throughout the manuscript.



Supplementary Figure 14

The line plot shows the average loss score at each position in a set of 600 bp test sequences provided as input to the model. The loss score measures the extent of the most damaging mutation to a nucleotide in a sequence; thus, the average loss score for each position reflects the influence of that position on the predictions. The grey lines indicate 2 standard deviations up and down for the estimate of the mean at each position.