

Domain adaptive neural networks improve cross-species prediction of transcription factor binding

Kelly Cochran,^{1,3} Divyanshi Srivastava,^{1,2} Avanti Shrikumar,³ Akshay Balsubramani,⁴ Ross C. Hardison,^{1,2} Anshul Kundaje,^{3,4,*} and Shaun Mahony^{1,2,*}

¹ *Center for Eukaryotic Gene Regulation, Pennsylvania State University, University Park, PA, USA*

² *Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA*

³ *Department of Computer Science, Stanford University, Stanford, CA, USA*

⁴ *Department of Genetics, Stanford University, Stanford, CA, USA*

**To whom correspondence should be addressed (akundaje@stanford.edu; mahony@psu.edu).*

¹ Abstract

² The intrinsic DNA sequence preferences and cell-type specific cooperative partners of transcription factors (TFs) are typically highly conserved. Hence, despite the rapid evolutionary turnover of individual TF binding sites, predictive sequence models of cell-type specific genomic occupancy of a TF in one species should generalize to closely matched cell types in a related species. To assess the viability of cross-species TF binding prediction, we train neural networks to discriminate ChIP-seq peak locations from genomic background and evaluate their performance within and across species. Cross-species predictive performance is consistently worse than within-species performance, which we show is caused in part by species-specific repeats. To account for this domain shift, we use an augmented network architecture to automatically discourage learning of training species-specific sequence features. This domain adaptation approach corrects for prediction errors on species-specific repeats and improves overall cross-species model performance. Our results demonstrate that cross-species TF binding prediction is feasible when models account for domain shifts driven by species-specific repeats.

14 **Introduction**

15 Characterizing where transcription factors (TFs) bind to the genome, and which genes they regulate, is key
16 to understanding the regulatory networks that establish and maintain cell identity. A TF's genomic occu-
17 pancy depends not only on its intrinsic DNA sequence preferences, but also on several cell-specific factors,
18 including local TF concentration, chromatin state, and cooperative binding schemes with other regulators
19 (Siggers and Gordân 2014; Slattery et al. 2014; Srivastava and Mahony 2020). Experimental assays such as
20 ChIP-seq can profile a TF's genome-wide occupancy within a given cell type, but such experiments remain
21 costly, rely on relatively large numbers of cells, and require either high-quality TF-specific antibodies or
22 epitope tagging strategies (Park 2009; Savic et al. 2015). Accurate predictive models of TF binding could
23 circumvent the need to perform costly experiments across all cell types and all species of interest.

24 Cross-species TF binding prediction is complicated by the rapid evolutionary turnover of individual
25 TF binding sites across mammalian genomes, even within cell types that have conserved phenotypes. For
26 example, only 12-14% of binding sites for the key liver regulators CEBPA and HNF4A are shared across
27 orthologous genomic locations in mouse and human livers (Schmidt et al. 2010). On the other hand, the
28 general features of tissue-specific regulatory networks appear to be strongly conserved across mammalian
29 species. The amino acid sequences of TF proteins, their DNA-binding domains, and intrinsic DNA sequence
30 preferences are typically highly conserved (e.g., both CEBPA and HNF4A have at least 93% whole protein
31 sequence identity between human and mouse). Further, the same cohorts of orthologous TFs appear to
32 drive regulatory activities in homologous tissues. Thus, while genome sequence conservation information
33 is not sufficient to accurately predict TF binding sites across species, it may still be possible to develop
34 predictive models that learn the sequence determinants of cell-type specific TF binding and generalize
35 across species. Indeed, several recent studies have demonstrated the feasibility of cross-species prediction
36 of regulatory profiles using machine learning approaches (Chen et al. 2018; Kelley 2020; Schreiber et al.
37 2020; Huh et al. 2018).

38 Here, we evaluate different training strategies on the generalizability of neural network models of cell-
39 type specific TF occupancy across species. We train our model using genome-wide TF ChIP-seq data in a
40 given cell type in one species, and then assess its performance in predicting genome-wide binding of the
41 same TF in a closely matched cell type in a different species. Specifically, we focus on predicting binding of
42 four TFs (CTCF, CEBPA, HNF4A, and RXRA) in liver due to the existence of high quality ChIP-seq data in

43 both mouse and human. We proceed to investigate gaps in performance between within-species and cross-
44 species models, with the aim of identifying specific genomic patterns that are associated with systematic
45 misprediction specifically across species.

46 We further evaluate the model performance improvement gained from integrating an unsupervised
47 domain adaptation approach into model training. This domain adaptation strategy involves a neural net-
48 work architecture with two sub-networks that share an underlying convolutional layer. We train the two
49 sub-networks in parallel on different tasks. One subnetwork is trained with standard backpropagation
50 to optimize classification of TF bound and unbound sequences in one species (the source domain). The
51 other subnetwork attempts to predict species labels from sequences drawn randomly from two species (the
52 source and target domain), but training is subject to a gradient reversal layer (GRL) (Ganin et al. 2016).
53 While backpropagation typically has the effect of giving higher weights to discriminative features, a GRL
54 reverses this effect, and discriminative features are down-weighted. Thus, our network aims to encourage
55 features in the shared convolutional layer that discriminate between bound and unbound sites, while si-
56 multaneously discouraging features that are species-specific. Importantly, this approach does not use TF
57 binding labels from the target species at any stage in training. We conclude by assessing the effectiveness
58 of domain adaptation in terms of reducing systematic mispredictions.

59 Results

60 **Conventionally trained neural network models of TF binding show reduced predictive performance 61 across species**

62 First, we set out to evaluate the ability of neural networks to predict TF binding in a previously unseen
63 species. We chose neural networks due to their ability to learn arbitrarily complex predictive sequence
64 patterns (Avsec et al. 2021a; Avsec et al. 2021b; Fudenberg et al. 2020; Kelley 2018; Koo et al. 2021).
65 In particular, hybrid convolutional and recurrent network architectures have successfully been applied to
66 accurately predict TF binding in diverse applications (Quang and Xie 2016; Quang and Xie 2019; Srivastava
67 et al. 2020). The motivation behind these architectures is that convolutional filters can encode binding site
68 motifs and other contiguous sequence features, while the recurrent layers can model flexible, higher-order
69 spatial organization of these features. Our baseline neural network is designed in line with these state-of-
70 the-art hybrid architectures (Figure 1).

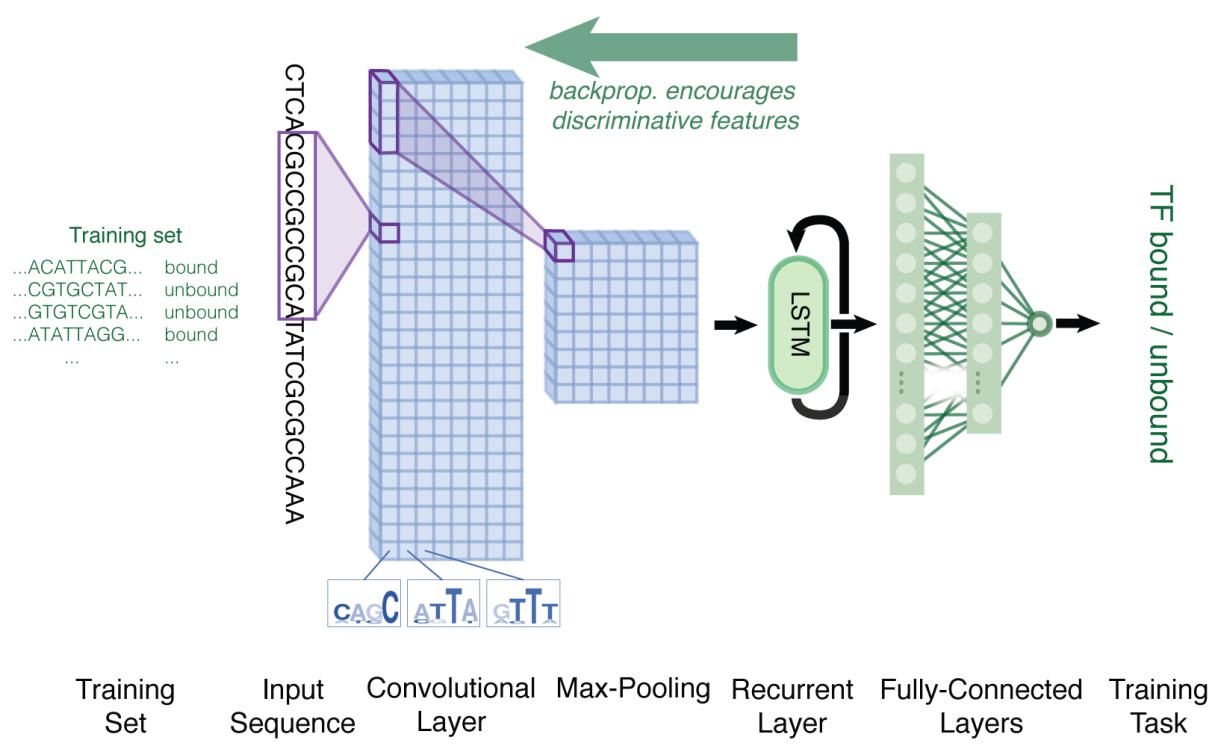


Figure 1: Conventional network architecture. Convolutional filters scan the 500-bp input DNA sequence for TF binding features. The convolutional layer is followed by a recurrent layer (LSTM) and two fully connected layers. A final sigmoid-activated neuron predicts if a ChIP-seq peak falls within the input window.

71 Using this architecture, named the “conventional model,” we trained the network to predict whether
72 a given input sequence contained a ChIP-seq peak or not, using training data from a single source species,
73 and then assessed the model’s predictive performance on entire held-out chromosomes in both the source
74 species and a target (previously unseen) species. We chose mouse and human as our species of interest due
75 to the availability of high-quality TF ChIP-seq datasets in liver from both species and the high conservation
76 of key regulator TFs present in both species. For four different TFs, we trained two sets of models: one with
77 mouse as the source species, and the other with human as the source species. To monitor reproducibility,
78 model training was repeated 5 times for each TF and source species.

79 As models trained for 15 epochs, we monitored source-species and target-species performance on
80 held-out validation sets (Figure 2). Performance was measured using the area under the precision-recall
81 curve (auPRC) which is sensitive to the extreme class imbalance of labels in our TF binding prediction
82 task. We observed that over the course of model training, improvements in source-species auPRC from
83 epoch to epoch did not always translate to improved auPRC in the target species. Generally, cross-species
84 auPRCs showed greater variability across epochs and model replicates compared to source-species auPRCs.
85 For HNF4A in particular, the mouse-trained models’ performance on the human validation set appeared
86 to split part way through training – based on cross-species auPRC, some model-replicates appeared to
87 become trapped in a suboptimal state relative to other models (see divergence in red lines in left column
88 of Figure 2); meanwhile, the training-species auPRC did not show a similar trend. Evidently, validation
89 set performance in the source species is not an ideal surrogate for validation set performance in the target
90 species.

91 Nevertheless, the epochs where models had highest source-species auPRCs were often epochs where
92 models had near-best cross-species auPRC. Thus, we selected models saved at the point in training when
93 source-species auPRC was maximized for downstream analysis. We next evaluated performance on held-
94 out test datasets (distinct from the validation datasets) from each species (Figure 3).

95 We observe across all TFs that for a given target species, the models trained in that species always
96 outperformed or matched the performance of the models trained in the other species. We refer to this
97 within-species vs. cross-species auPRC difference as a cross-species performance gap, while noting that
98 models trained in either species were still relatively effective at cross-species prediction. Because we ob-
99 serve a wider cross-species gap for mouse-trained models predicting in human than for human-trained
100 models predicting in mouse, subsequent analysis focuses on addressing the mouse-to-human gap.

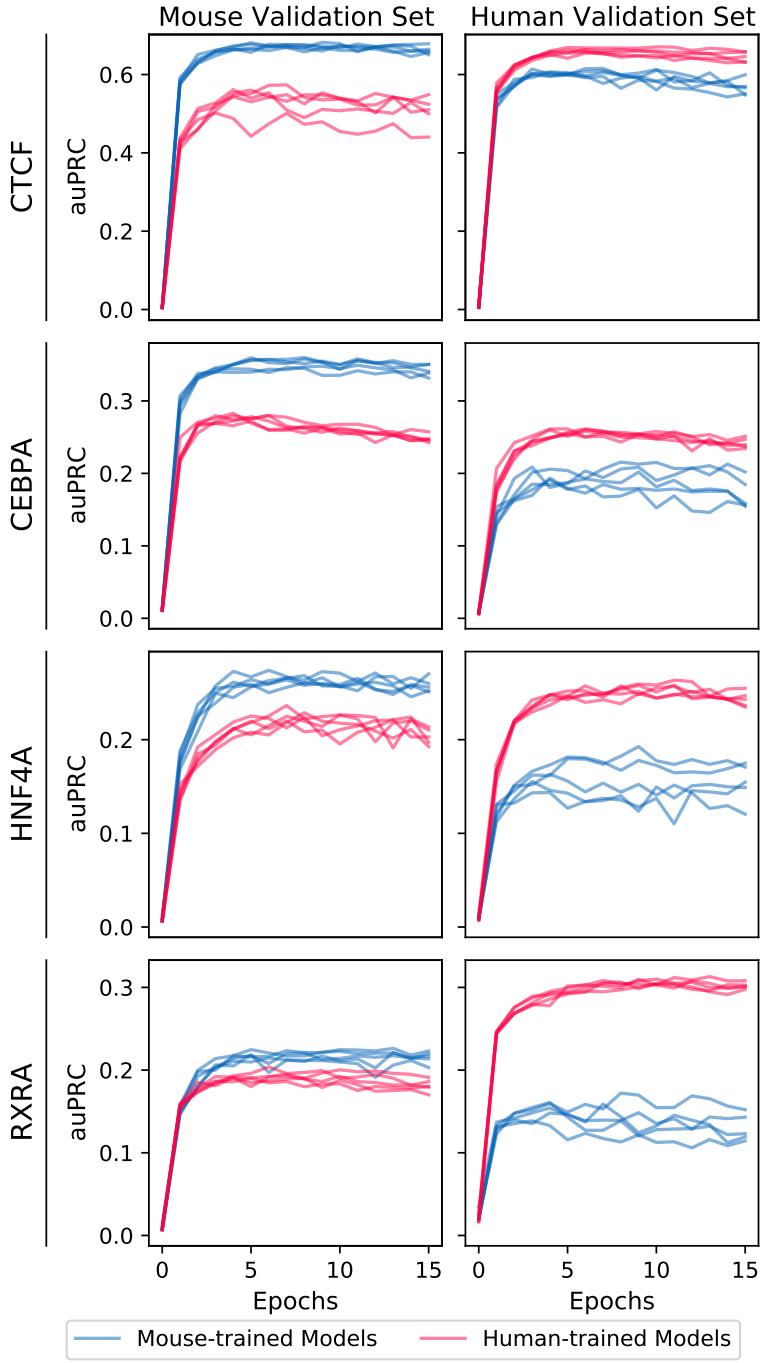


Figure 2: Model performance over the course of training, evaluated on held-out validation data from mouse (left) and human (right) Chromosome 1. Five models were independently trained for each TF and source species (mouse-trained models in blue, human-trained models in red). Values at epoch 0 are evaluations of models after weight initialization but before training (akin to a random baseline). Note that auPRCs are not directly comparable between different validation sets because ground truth labels are derived from a different experiment for each dataset; the auPRC will depend on the fraction of sites labeled bound as well as model prediction correctness.

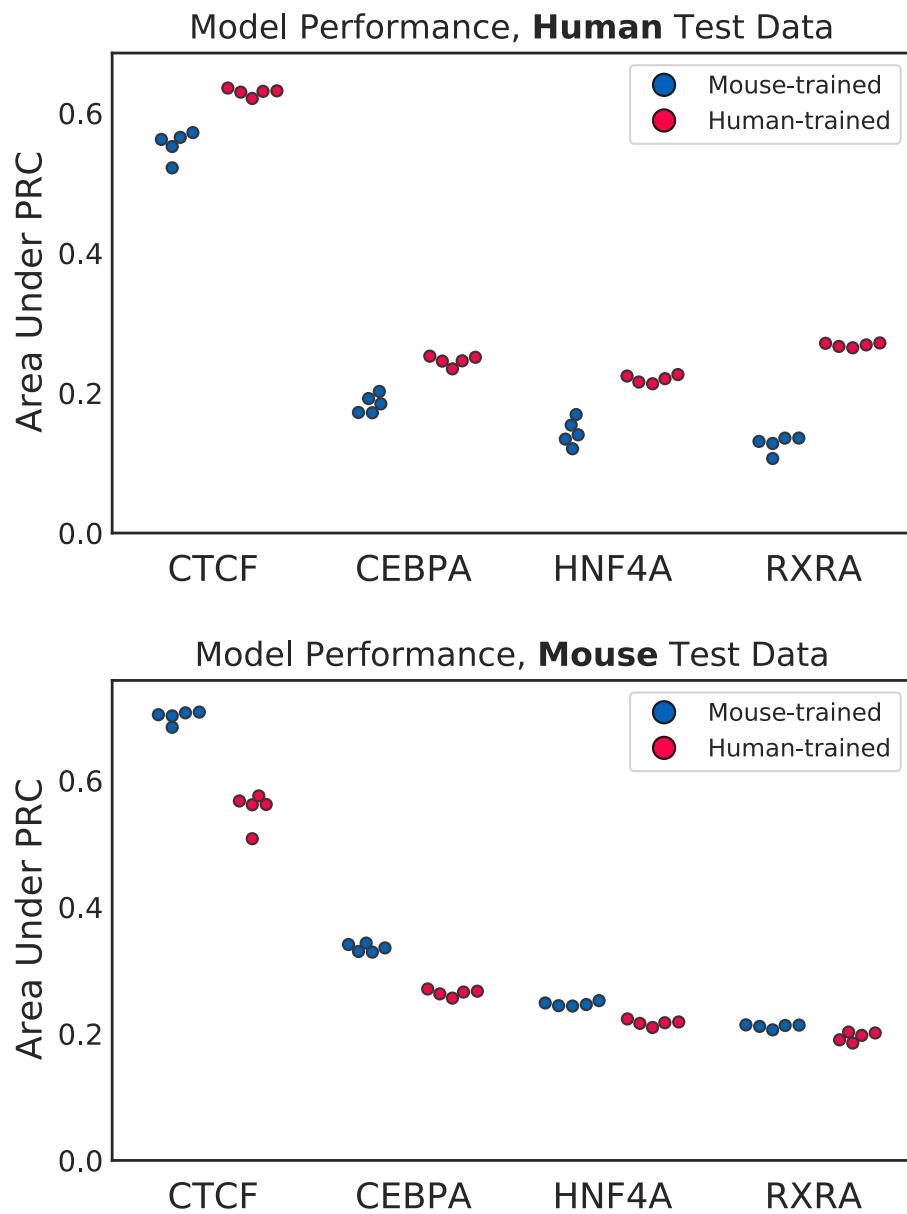


Figure 3: Model performance evaluated on held-out test data: Chromosome 2 from human (top) and mouse (bottom). Five models were independently trained for each TF and source species.

101 To get a sense of how specific to our model design or training strategy this cross-species gap might be,
102 we applied multiple sufficiently different machine learning approaches to the same problem and datasets
103 and assessed whether the cross-species gap persists. First, we trained gapped k -mer support vector ma-
104 chines, or gkSVMs, to classify a balanced sample of bound vs. unbound windows for each TF and species
105 (Ghandi et al. 2014; Lee 2016). We then evaluated those models on the set of non-overlapping windows
106 in each test dataset (Supplemental Fig. S1). We observe that the cross-species gap persists, although it
107 shrinks in absolute magnitude, presumably due to the drastically lower auPRC values across the board.
108 These auPRCs also demonstrate that our neural network approach can indeed outperform related methods
109 on this task.

110 Next, we sought to assess the cross-species performance of another state-of-the-art deep learning
111 model trained on a related TF binding prediction task, distinct from our binary classification setup. We ap-
112 plied a BPNet-like profile model, which predicts the distribution of the raw, base-resolution ChIP-seq read
113 profile at a given genomic window rather than a 0-1 binary label, to both our mouse and human datasets
114 across our four TFs (Avsec et al. 2021b). The profile models were trained using a peak-enriched subset of
115 the training data used by the binary models, and performance was evaluated on the same test datasets (see
116 Methods).

117 First, we investigated how well individual profile predictions transfer across species (Supplemental
118 Fig. S2, bottom). We observe that overall, within-species profile models are usually able to predict both the
119 location and the shape of peaks accurately. Cross-species profile models tend to predict the peak location
120 nearly as well as within-species models, but for some TFs, there is a clear discrepancy between the predicted
121 and true profile shape. Specifically, there are apparent non-biological differences in experimental protocol
122 or quality between our matched datasets across species; this can cause profile models that learned how
123 reads typically distribute around binding sites from one experiment to appear to generalize imperfectly to
124 other datasets with different read distributions about binding sites.

125 Next, we quantified the performance of the profile models, using the predicted total number of reads
126 across a genomic window as a proxy for binary label prediction (Supplemental Fig. S2, top). We again
127 observe cross-species performance gaps for most datasets. We also note that the auPRC values attained by
128 the profile models are comparable to those attained by our conventional model in most cases, so we decided
129 to focus on understanding the cross-species gap in the context of the conventional model in the remainder
130 of the study.

131 **The mouse-to-human cross-species gap originates from misprediction of both bound and unbound sites**

132 Since the target-species model consistently outperforms the source-species model (on target-species valida-
133 tion), there must be some set of differentially predicted sites that the target-species model predicts correctly,
134 but the source-species model does not. By comparing the distribution of source-model and target-model
135 predictions over all target-species genomic windows, we can potentially identify trends of systematic errors
136 unique to the source-species model. Whether these differentially predicted sites are primarily false posi-
137 tives (unbound sites incorrectly predicted to be bound), false negatives (bound sites incorrectly predicted
138 as unbound), or a combination of both can provide useful insight into the performance gap between the
139 source and target models.

140 For each TF, we generated predictions over the genomic windows in the human test dataset from both
141 our mouse-trained and human-trained models. Then, we plotted all of the human-genome test sites using
142 the average mouse model prediction (over 5 independent training runs) and the average human model
143 prediction as the x- and y-axis, respectively (Figure 4). Bound and unbound sites are segregated into
144 separate plots for clarity.

145 For three of the four TFs, the unbound site plots show a large set of windows given low scores by
146 the human model but mid-range to high scores by the mouse model – these are false positives unique to
147 cross-species prediction (Figure 4 right column, bottom/bottom-right region of each plot). These sites are
148 distinct from false positives mistakenly predicted highly by both models, as those common false positives
149 would not contribute significantly to the auPRC gap. Even for CTCF, the exception to the pattern, there
150 is an enrichment of unbound sites that can be characterized as mispredictions specific to mouse models.
151 Additionally, in the bound site plots of all TFs except CEBPA, we see some bound sites that are scored
152 high by the human model but are given mid-range to low scores by the mouse model – these are cross-
153 species-unique false negatives (Figure 4 left column, top left region of each plot). Hence, our cross-species
154 models are committing prediction errors in both directions on separate sets of site, although the errors in
155 the unbound sites appear more prevalent than the errors in the bound sites.

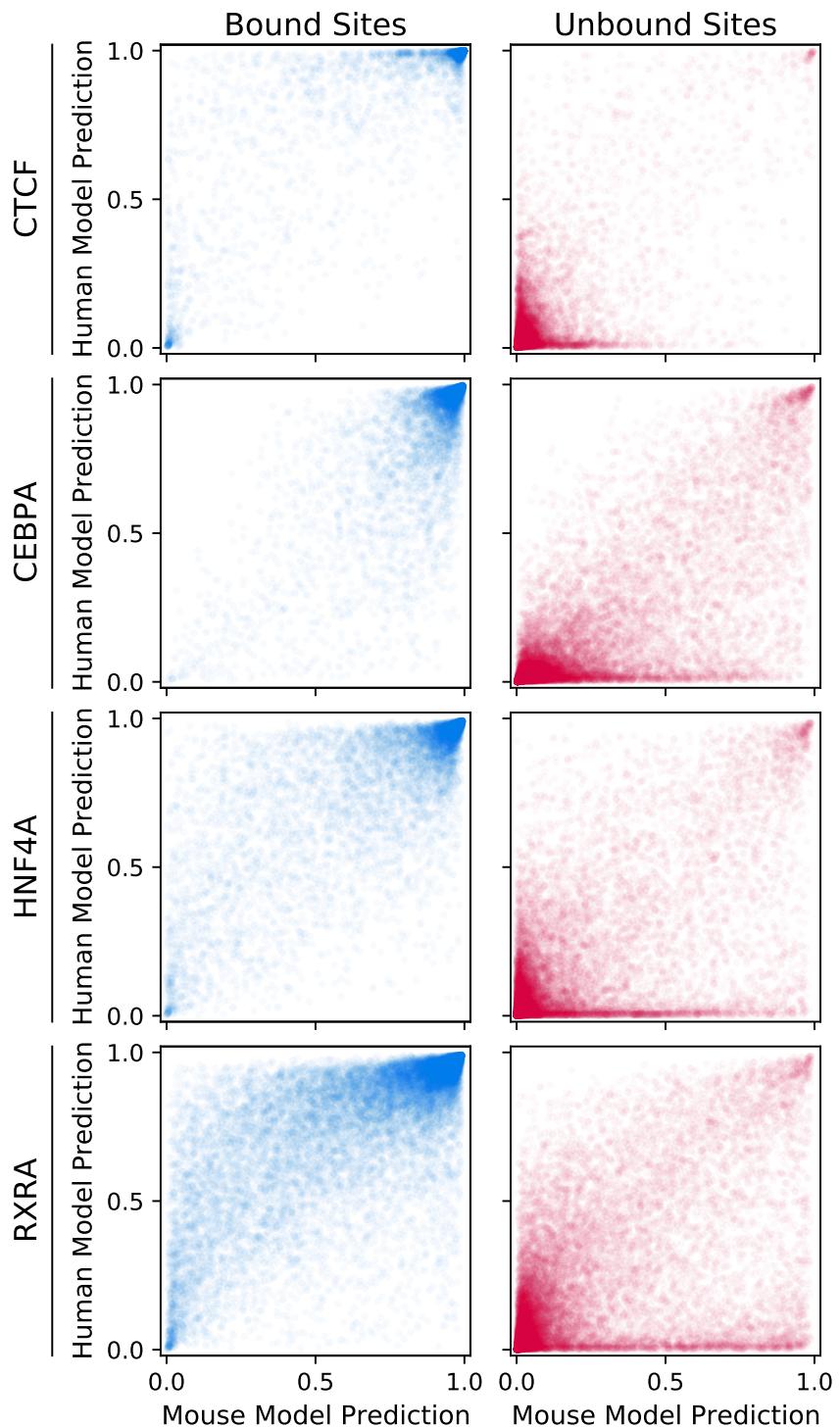


Figure 4: Both bound and unbound sites from human Chromosome 2 show evidence of differential binding predictions by human-trained (y-axis) vs. mouse-trained (x-axis) models. For visual clarity, only 25% of bound sites and 5% of unbound sites are shown (sampled systematically).

156 **Motif-like sequence features discriminate between true-positive and false-negative mouse model pre-
157 dictions**

158 Since the only input to our models is DNA sequence, sequence features must be responsible for differential
159 prediction of certain sites across source and target models. Other potential culprits, such as chromatin
160 accessibility changes or co-factor binding, may contribute to TF binding divergence across species without
161 changes to sequence; but without an association between those factors and sequence, the human-trained
162 model would not be able to gain an advantage over the mouse-trained model by training on sequence input
163 alone. Thus, we focused on genomic sequence to understand differential site prediction.

164 To begin, we searched for sequence features associated with differential prediction of bound sites from
165 the human genome – specifically, we compared bound sequences that both the human-trained and mouse-
166 trained models correctly predicted (true positives) to bound sequences the human-trained model correctly
167 predicted but the mouse-trained model did not (mouse-specific false negatives). We used SeqUnwinder, a
168 tool for deconvolving discriminative sequence features between sets of genomic sequences, to extract motifs
169 that can discriminate between the two groups of sequences and quantitatively assess how distinguishable
170 the sequence groups are (Kakumanu et al. 2017). SeqUnwinder was able to distinguish mouse-specific
171 false negatives from true positives and randomly selected background genomic sequences with area under
172 the ROC curve (auROC) of 0.78, 0.79, 0.80, and 0.87 for CTCF, CEBPA, HNF4A, and RXRA, respectively.
173 Supplemental Fig. S3 shows the breakdown of sequence features that are able to distinguish between
174 mouse-specific false negatives and true positives for each TF. Thus, we were able to identify TF-specific
175 motifs that were enriched (or depleted) at mouse-specific false negatives. However, we did not observe
176 systemic sequence features that unanimously contributed to the performance gap across all TFs studied,
177 beyond a poly-A/poly-T motif.

178 **Primate-unique SINEs are a dominant source of the mouse-to-human cross-species gap**

179 One potential source of sequences that could confuse a cross-species model are repeat elements found
180 in the genome of the target species but not the source species. *Alu* elements, a type of SINE, cover a
181 large portion (10%) of the human genome and are found only in primates (Batzer and Deininger 2002).
182 Several other factors make *Alus* even more likely candidates for confounding mouse-to-human TF binding
183 predictions: they are enriched in gene-rich, GC-rich areas of the genome and contain 33% of the genome's

184 CpG dinucleotides (a marker for promoter regions); they may play a role in gene regulation; and in silico
185 studies have previously found putative TF binding sites within *Alu* sequences (Batzer and Deininger 2002;
186 Schmid 1998; Ferrari et al. 2019; Polak and Domany 2006).

187 Figure 5 shows only the unbound human-genome windows that overlap annotated *Alu* elements. Ta-
188 ble 1 provides corresponding quantification of *Alu* enrichment. Note that while *Alu* elements are typically
189 poorly mappable, and it is thus often difficult to assign them as bound or unbound in ChIP-seq experiments,
190 we focus analyses here only on highly mappable *Alu* instances (see Methods). Across all four TFs, we see
191 that *Alus* are substantially enriched in the unbound windows mispredicted only by the mouse model. On
192 average, 89% of these false positives unique to the mouse model overlap with an *Alu* element, compared
193 to the average overlap rate of 21% for unbound sites overall, or 18% for unbound sites incorrectly pre-
194 dicted by both models. In contrast, *Alus* on average only overlap 6% of false negatives unique to the mouse
195 model, which is less than the overlap fraction for bound sites overall (15%) and for false negatives common
196 to both models (11%). We repeated this analysis using other repeat classes, including LINEs and LTRs,
197 and confirmed that no other major repeat family shows an enrichment of comparable strength with either
198 the false positives or false negatives unique to the mouse model (Supplemental Table S1). Investigating
199 the enrichment of individual *Alu* subfamilies in mouse-model-unique false positives showed that this phe-
200 nomenon is not restricted to a single subtype of *Alu*, but that subfamilies are enriched at different levels in
201 a manner that is TF-specific and varies particularly between the *AluJ*, *AluS*, and *AluY* subfamily groupings
202 (Supplemental Fig. S4).

203 Thus, the vast majority of the false positives from the human genome mispredicted only by mouse
204 models can be directly attributed to one type of primate-unique repeat element. We did not observe any
205 similar direct associations between primate-unique elements and the false negatives unique to the mouse
206 model, besides the expected depletion of *Alu* elements.

207 **Model interpretation reveals sequence features driving divergent mouse and human model predictions**

208 To understand why mouse and human models make divergent predictions at some sites, we compared base-
209 pair resolution importance scores from both models at selected example sites. Specifically, we implemented
210 a strategy similar to in silico mutagenesis (ISM) where a base's score was determined by the differential
211 model output between the original sequence and the sequence with 5bp centered on that base replaced
212 with bases from a dinucleotide-shuffled reference (Alipanahi et al. 2015). We observed that this strategy

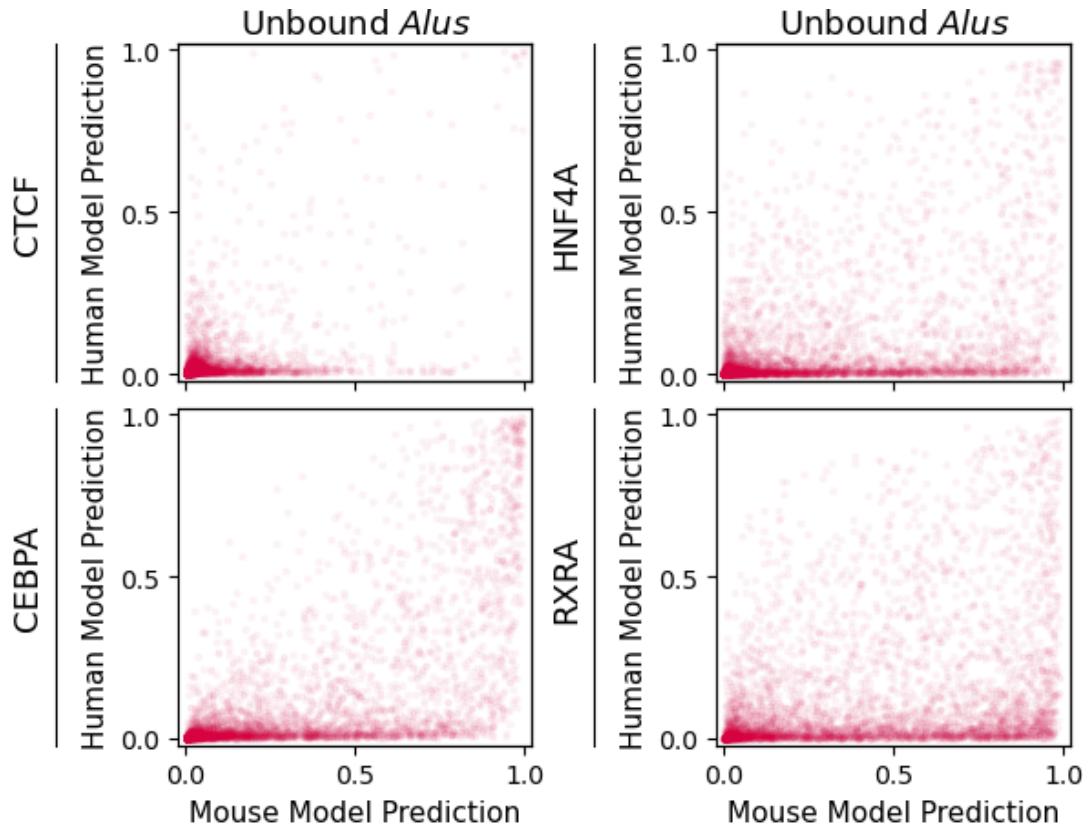


Figure 5: Most unbound sites from the human genome mispredicted by mouse-trained models (x-axis), but not by human-trained (y-axis) models, contain *Alu* repeats. For visual clarity, only 5% of windows are shown.

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	12.6%	12.8%	9.9%	21.3%	10.0%	78.6%
CEBPA	18.3%	11.1%	0.0%	21.3%	22.9%	84.8%
HNF4A	13.6%	10.4%	8.0%	21.3%	16.9%	95.1%
RXRA	13.7%	10.6%	5.5%	21.4%	20.3%	97.4%

Table 1: Percent of windows overlapping an *Alu* element, for various categories of genomic windows from the held-out test set. *Alu* elements dominate the false positives unique to the mouse models. FPs: false positives. FNs: false negatives. See Methods for more details on site categorization.

213 outperformed backpropagation-based scoring methods, potentially by avoiding gradient instability.

214 First, we compared importance scores between the mouse and human models at example bound sites
 215 that both models predicted correctly (Supplemental Fig. S5). If the two models learned to use similar
 216 logic to make binding predictions, we would expect to see similar sequence features highlighted in the
 217 importance scores. Overall, we observe that the scores generated by the mouse and human models are
 218 reasonably concordant, although the extent of agreement varies noticeably across TFs. CTCF and CEBPA
 219 show the greatest tendency for agreement in importance scores across models. HNF4A showed a slightly
 220 weaker trend of score agreement, while RXRA importance scores were the most likely to disagree across
 221 models, including instances where motifs are highlighted by high scores from one model but given near-
 222 zero scores by the other model. However, across all TFs, instances of the primary cognate motif for the
 223 appropriate TF are common in the sequences marked by higher importance scores from either model.

224 Next, we repeated the analysis on example unbound windows classified as mouse-model-unique false
 225 positives (Supplemental Fig. S6). At these sites, the mouse model’s prediction scores overshoot those of the
 226 human model by at least 0.5. Importance scores in this set of sites show much greater disagreement between
 227 the two models. Commonly across all four TFs, we observed two trends: first, the mouse models often
 228 assigned high importance to motif-sized contiguous stretches of bases which were not similarly recognized
 229 by the human models. These pseudo-motifs can superficially resemble approximate matches to the TF’s
 230 cognate motif. Second, the human models commonly showed apparent sensitivity to specific, often sparse
 231 features which received negative scores of moderate to high magnitude. These observations imply that the
 232 human model has learned to ignore certain sequence features that the mouse model’s scores suggest are
 233 favorable for binding. Furthermore, the human model may be adopting that strategy based on whether or
 234 not there are nearby sequence contexts that indicate that the sequence is not a binding site.

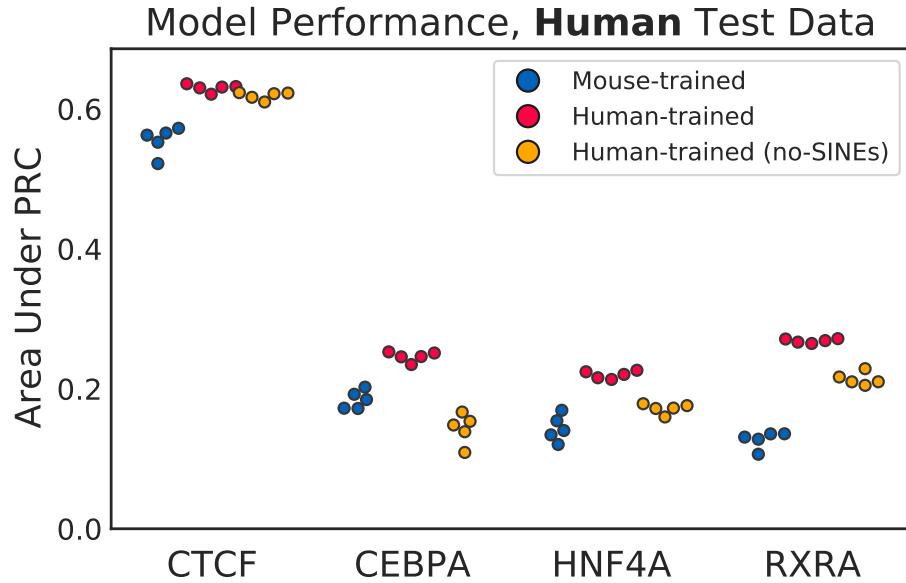


Figure 6: Performance of models that are mouse-trained (blue), human-trained with SINE examples (red), and human-trained without SINE examples (yellow), evaluated on the held-out human Chromosome 2. Five models were independently trained for each TF and training species.

235 **Human models trained without SINE examples behave like hybrid mouse-human models**

236 To further characterize how *Alu* elements are influencing cross-species model performance, we trained
 237 additional models on the human dataset after removing all windows from the training dataset that overlap
 238 with any SINEs (Figure 6). We filtered out all SINEs, including the primate-specific *FLAM* and *FRAM*
 239 repeats as well as *Alus*, to avoid keeping examples that shared any sequence homology with *Alus*. The no-
 240 SINE models were evaluated on the same held-out chromosome test data used previously (which includes
 241 SINEs). For all TFs except CTCF, the no-SINE models perform substantially worse than models trained
 242 using the complete human training sets.

243 Site-distribution plots show that, for unbound sites, no-SINE human-trained models make mispre-
 244 dictions in a pattern similar to mouse-trained models; there is a similarly-sized subset of unbound sites
 245 mispredicted by the no-SINE human-trained models but not by the standard human-trained models (Fig-
 246 ure 7). Plotting only the sites that overlap with *Alus* confirms that the false positives unique to the no-SINEs
 247 model are predominantly *Alu* elements (Supplemental Fig. S7). For bound sites, on the other hand, no-
 248 SINE human-trained models make predictions that generally agree with predictions from standard human-

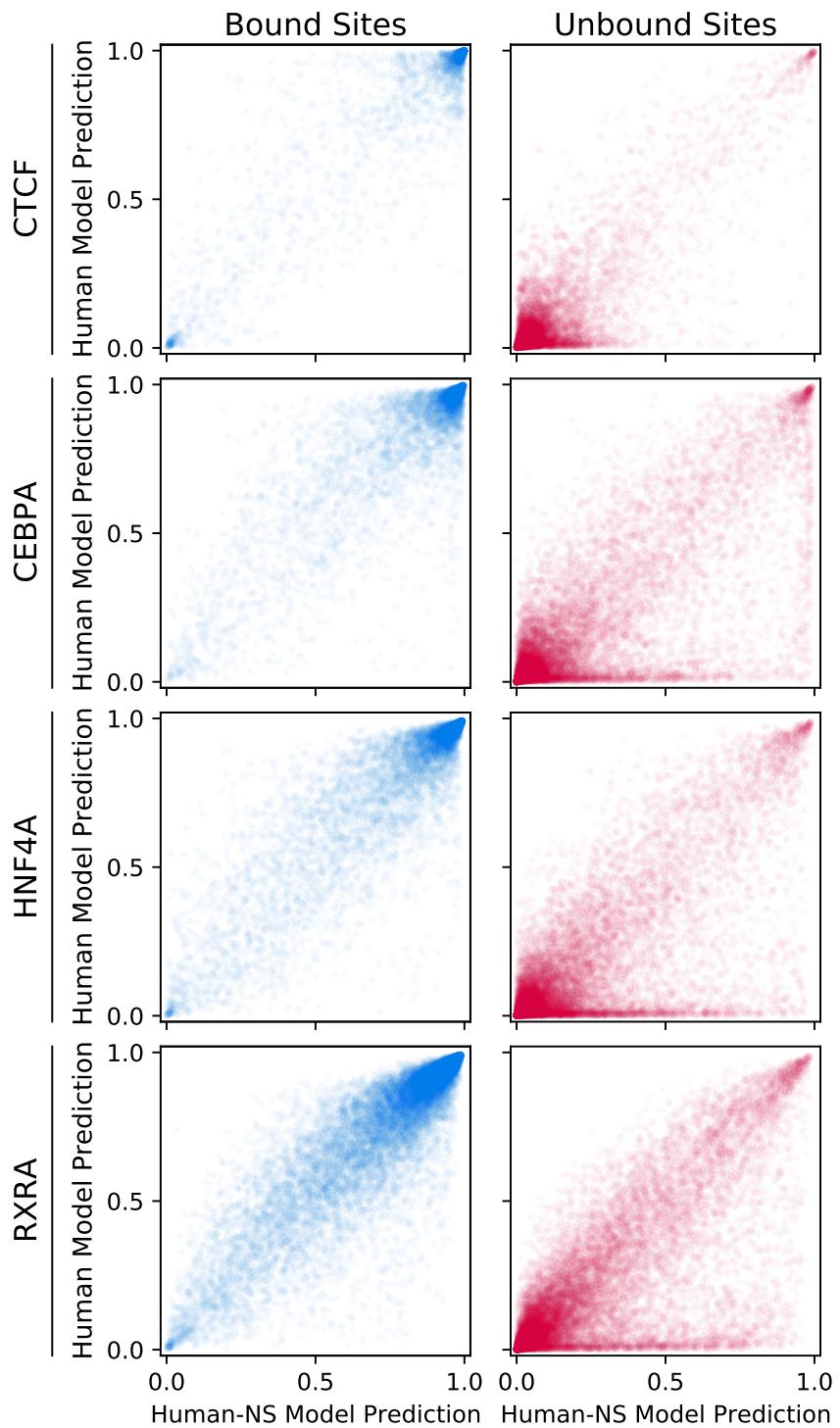


Figure 7: Differential human Chromosome 2 site predictions between models trained on human data with or without any examples of SINE windows. Human-NS: models trained on human data with no SINE examples. Similar to mouse-trained models, no-SINE human-trained models systematically mispredict some unbound sites.

249 trained models.

250 This suggests that the *Alu* false positives unique to the mouse-trained model may simply be due to the
251 fact that mouse models are not exposed to *Alus* during training (i.e., *Alu* elements are “out of distribution”).
252 In addition, the reduction in model-unique false negatives observed when the no-SINE human-trained
253 model is compared to the normal human-trained model suggests that those mispredictions are unrelated
254 to *Alus*.

255 **Domain-adaptive mouse models can improve cross-species performance**

256 Having observed an apparent “domain shift” across species, partially attributable to species-unique re-
257 peats, our next step is to ask how we might bridge this gap and reduce the difference in cross-species model
258 performance. Our problem is analogous to one encountered in some image classification tasks, where the
259 test data is differently distributed from the training data to the extent that the model performs well on
260 training data but much worse on test data (for example, the training images were taken during the day
261 but the test images were taken at sunset). In these situations, various techniques for explicitly forcing the
262 model to adapt across different image “domains” have been shown to improve performance at test time
263 (e.g., Long et al. 2015; Bousmalis et al. 2016; Sun et al. 2016).

264 One unsupervised domain adaptation method utilizes a gradient reversal layer to encourage the “fea-
265 ture generator” portion of a neural network to be domain-generic (Ganin et al. 2016). The gradient reversal
266 layer’s effect is to backpropagate a loss to the feature generator that prevents any domain-unique features
267 from being learned. We chose to test the effectiveness of this version of domain adaptation for our cross-
268 species TF binding prediction problem because we have observed evidence that domain-unique features
269 (species-unique repeat elements) were a major component of the cross-species domain shift.

270 We modified our existing model architecture to perform training-integrated domain adaptation across
271 species (Figure 8). A gradient reversal layer (GRL) was added in parallel with the LSTM, taking in the result
272 of the max-pooling step (after the convolutional layer) as input. During standard feed-forward prediction,
273 the GRL merely computes the identity of its input, but as the loss gradient backpropagates through the
274 GRL, it is reversed. The output of the GRL then passes through two fully connected layers before reaching a
275 new, secondary output neuron. This secondary output, a “species discriminator,” is tasked with predicting
276 whether the model’s input genomic window is from the source or target species. The model training process

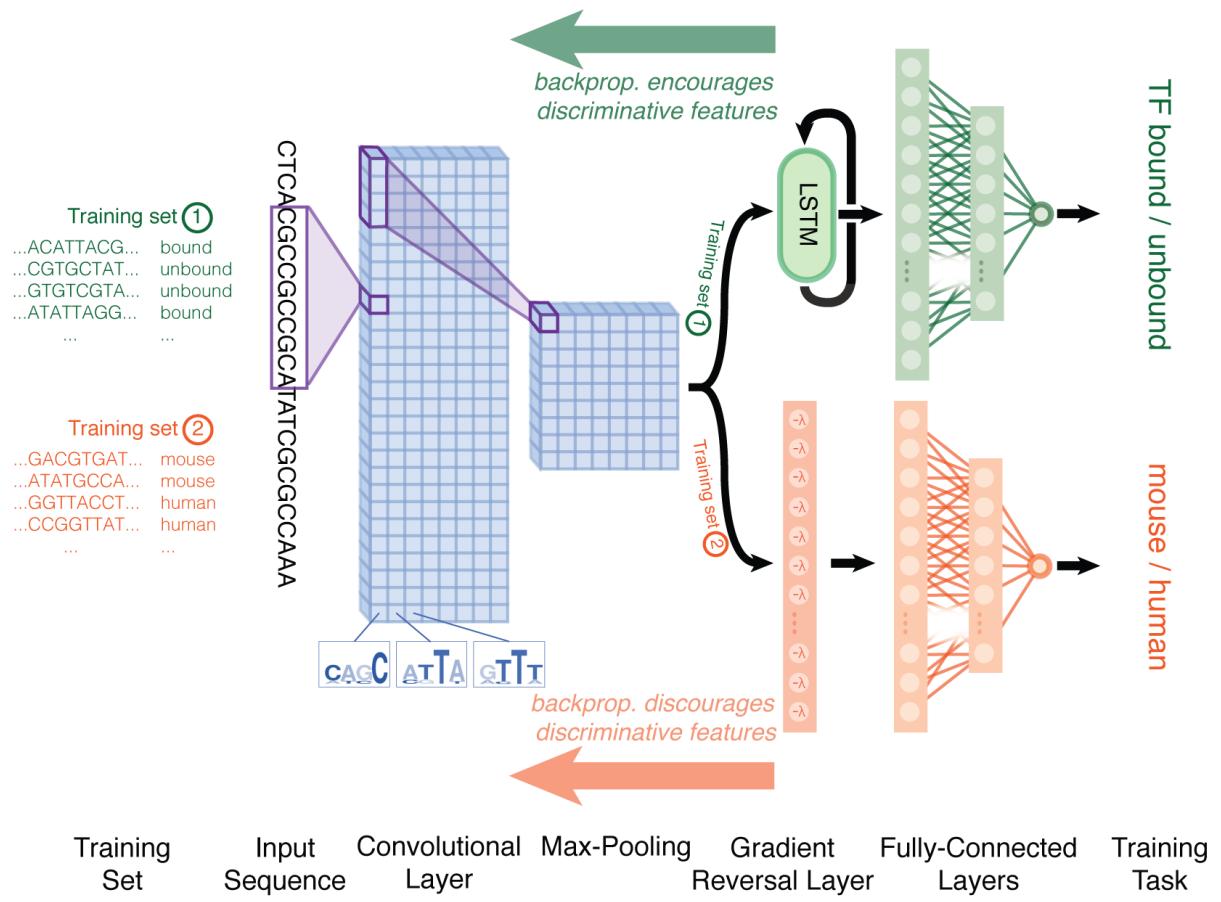


Figure 8: Domain-adaptive network architecture. The top network output predicts TF binding, as before, while the bottom network output predicts the species of origin of the input sequence window. The gradient reversal layer has the effect of discouraging the underlying convolutional filters from learning sequence features relevant to the species prediction task.

277 is modified so that the model is exposed to sequences from both species, but only the binding labels of
278 the source species (see Methods). Without the GRL, adding the species discrimination task to the model
279 would encourage the convolutional filters to learn sequence features that best differentiate between the
280 two species – features like species-unique repeats – but with the GRL included, the convolutional filters
281 are instead *discouraged* from learning these features. We hypothesize that this domain-adaptive model will
282 outperform our basic model architecture by reducing mispredictions on species-unique repeats.

283 We trained domain-adaptive models using the same binding training datasets as before and evaluated
284 performance with the same held-out datasets. We observe that the auPRC for our domain-adaptive models
285 on cross-species test data is moderately higher than the auPRC for the basic mouse models for all TFs except
286 CTCF, where auPRCs are merely equal (Figure 9, top, blue/left vs. green/middle dots). The domain-
287 adaptive models' auPRCs on mouse test data, meanwhile, is comparable to the auPRCs of basic models
288 (Figure 9, bottom, blue/left vs. green/middle). While the auPRC improvement is promising, it is also
289 modest in comparison to the full cross-species gap; the domain-adaptive models still do not achieve a level
290 of performance comparable to same-species models (Figure 9, top, green/middle vs. red/right).

291 **Domain-adaptive mouse models reduce over-prediction on *Alu* elements**

292 Next, we repeated our site-distribution analysis to determine what constituted the domain-adaptive mod-
293 els' improved performance. The unbound site plots in Figure 10 compare human genome predictions
294 between domain-adaptive mouse models and the original human models. *Alu* elements are highlighted in
295 Figure 11, with quantification in Supplemental Table S2.

296 Compared to Figure 4, the mouse-model-specific false positives have diminished for all TFs. This
297 suggests that the domain-adaptive models are able to correct the problem of false positive predictions from
298 *Alus* by scoring unbound sites overlapping *Alus* lower than the basic model did. This effect is even present
299 for CTCF, even though there was no noticeable auPRC difference for CTCF between domain-adaptive and
300 basic mouse models – likely because the initial *Alu* enrichment in CTCF mouse-model false positives was
301 lower than for other TFs.

302 In contrast, the site-distribution plots for bound sites demonstrate no noticeable difference from the
303 original plots for the basic model architecture. We applied the same SeqUnwinder analysis to look for
304 sequence features that discriminate between mouse-model false negatives and true positives and discov-

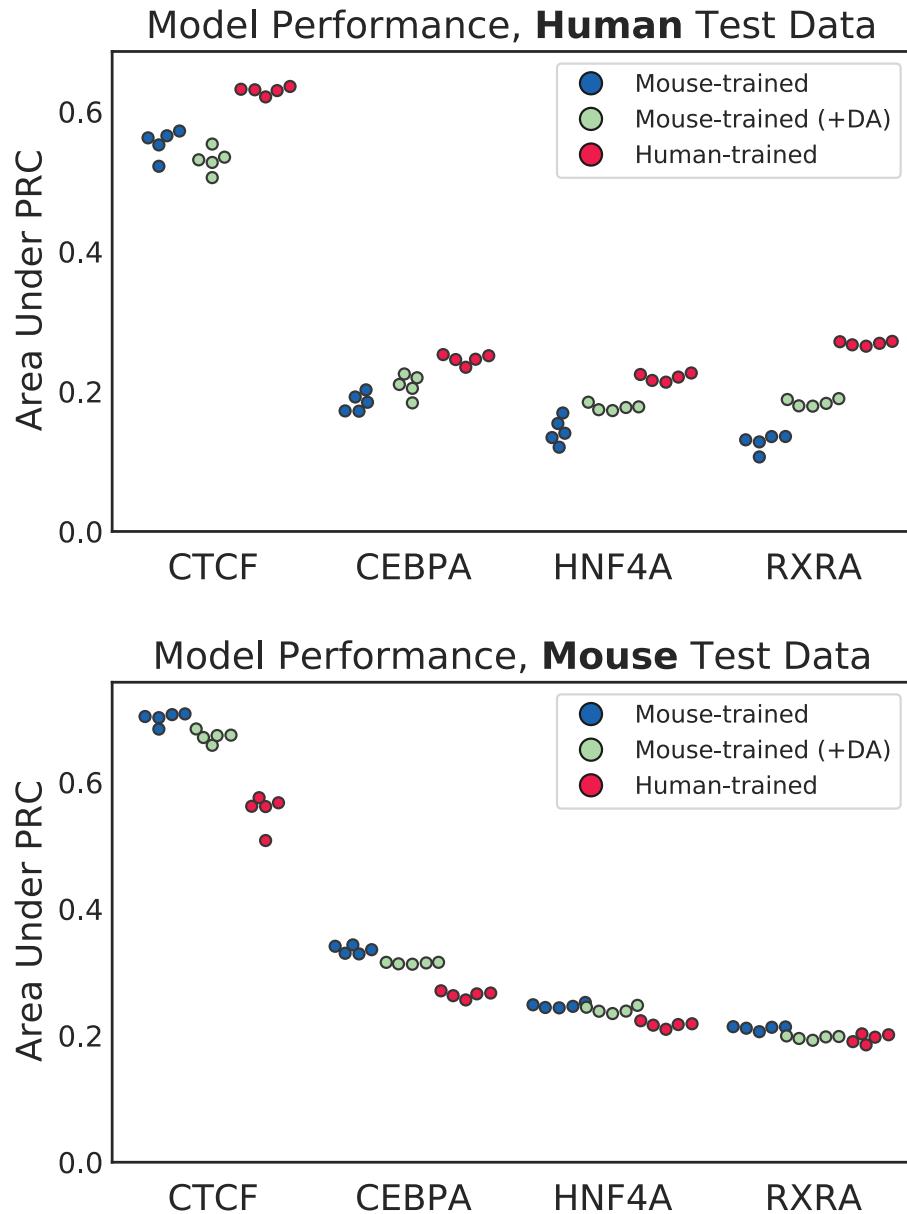


Figure 9: Performance of mouse-trained generic (blue), mouse-trained domain-adaptive (green), and human-trained (red) models, evaluated on human (top) and mouse (bottom) Chromosome 2. Five models were independently trained and evaluated for each TF and training species.

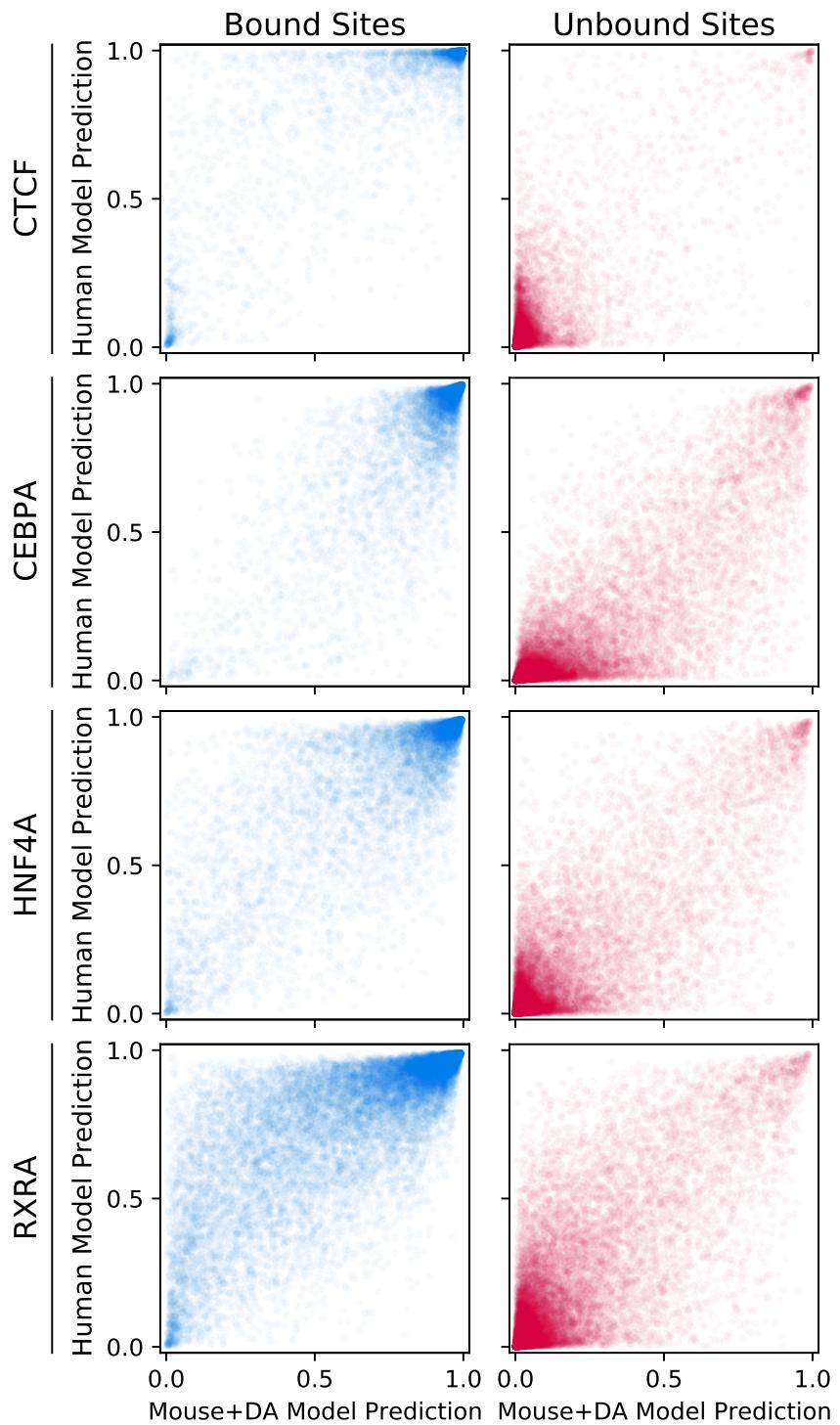


Figure 10: Differential predictions of human genome sites between human-trained and domain-adaptive mouse-trained models. Domain-adaptive mouse models, unlike the original mouse models, do not show species-specific systematic misprediction of unbound sites.

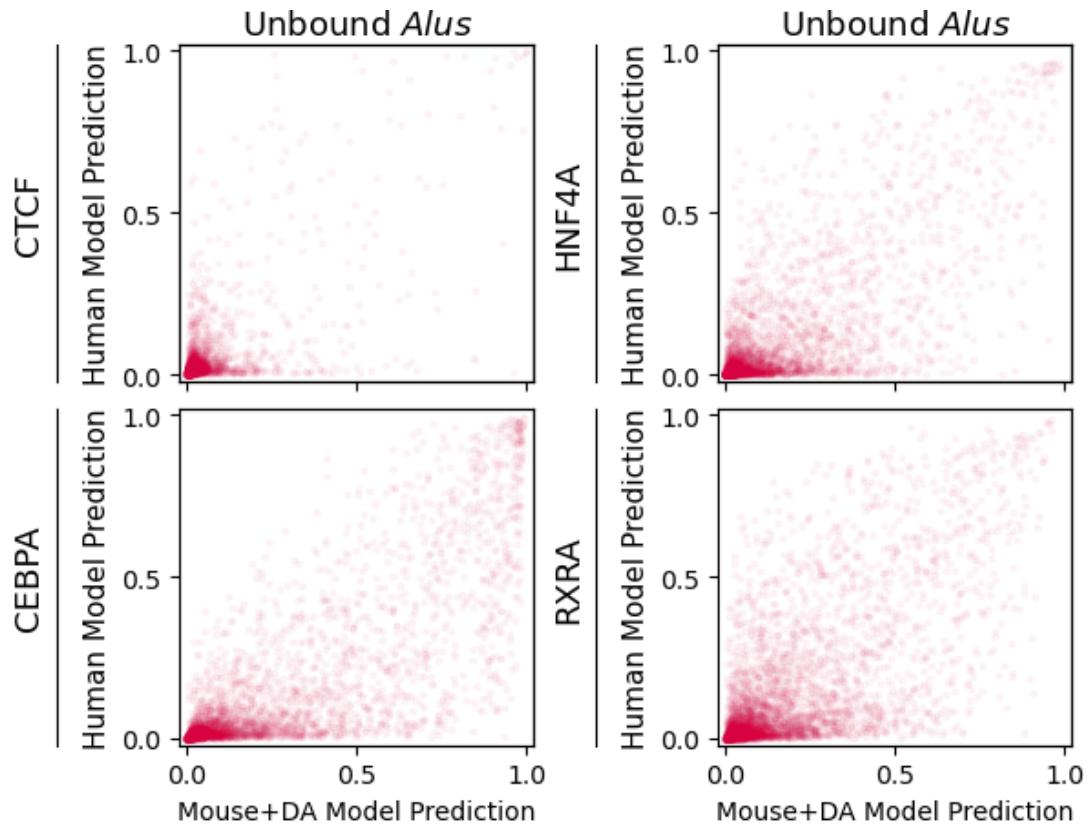


Figure 11: Differential predictions of unbound sites containing *Alu* elements between domain-adaptive mouse-trained models and human-trained models. Unlike the original mouse models, domain-adaptive mouse models do not show systematic overprediction of *Alu* repeats.

305 ered similar, but not identical, motif-like short sequence patterns as we did previously (Supplemental Fig.
306 S8). Thus, our domain adaptation approach does not appear to have any major influence on bound site
307 predictions.

308 ***Alus* commonly drive mouse-model false positives across diverse cell types**

309 Finally, we asked whether the observed over-prediction of species-specific repeats is a general issue of
310 concern in cross-species TF binding prediction, or whether it is particular to the examined liver TFs. We
311 thus widened our analyses to 53 additional pairs of ChIP-seq datasets targeting orthologous TFs across 8
312 additional equivalent human and mouse cell types (see Methods). One caveat is that the expanded set of
313 paired datasets typically focus on cell lines and cell types that are more difficult to closely match across
314 species than liver samples. Thus, the additional experiments examined here may not be as comparable
315 across species as the previously examined liver datasets.

316 Our expanded analyses confirm that the cross-species performance gap is present in most tested TFs
317 and cell types (Supplemental Table S3). A large portion of mouse-to-human false positive predictions is
318 attributable to *Alu* elements. In 43 of the 53 additional examined datasets, *Alu* elements overlap a third or
319 more of the mouse-model-unique false positive predictions (Supplemental Table S4). Our domain adap-
320 tation procedure is successful in reducing *Alu*-related false positive predictions in 46 of the 53 additional
321 examined datasets (Figure 12; Supplemental Table S4). However, in megakaryocyte and hematopoietic
322 progenitor datasets, we generally see a smaller percentage of mouse-model-unique false positives being
323 attributable to *Alus*. The false positive predictions that do overlap *Alus* are also generally less likely to be
324 corrected by our domain adaptation approach in these cell types (Figure 12). Therefore, our observations
325 may not apply uniformly to all cell types.

326 **Discussion**

327 Enabling effective cross-species TF binding imputation strategies would be transformative for studying
328 mammalian regulatory systems. For instance, TF binding information could be transferred from model
329 organisms in cell types and developmental stages that are difficult or unethical to assay in humans. Simi-
330 larly, one could annotate regulatory sites in non-model species of agricultural or evolutionary interest by
331 leveraging the substantial investment that has been made to profile TF binding sites in human, mouse, and

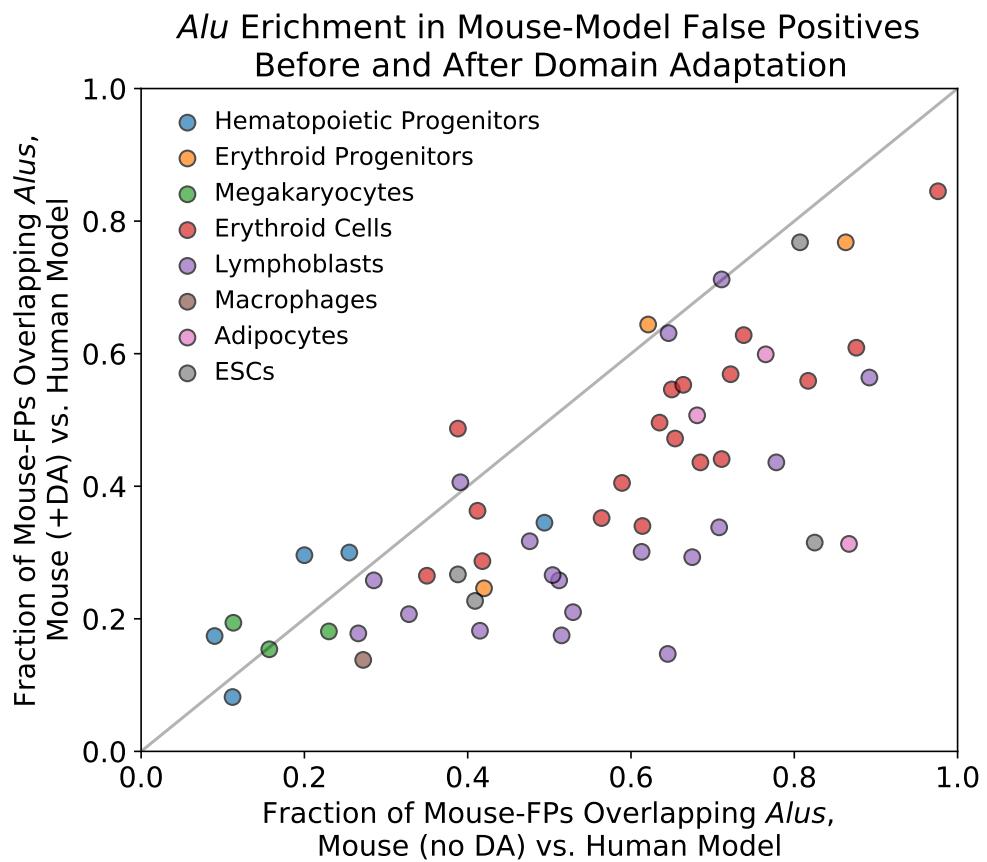


Figure 12: The fraction of mouse-model-unique false positives that overlap *Alus* when either the basic mouse model (x-axis) or the domain-adaptive mouse model (y-axis) are compared against the human model, across our additional paired datasets. The black diagonal line shows $y = x$; points below the line represent TFs where the fraction of *Alus* in mouse-model-unique false positives decreased with our domain adaptation strategy.

332 other model organisms (The ENCODE Project Consortium 2012; Yue et al. 2014; Roadmap Epigenomics
333 Consortium et al. 2015).

334 Our results suggest that cross-species TF binding imputation is feasible, but we also find a pervasive
335 performance gap between within-species and cross-species prediction tasks. One set of culprits for this
336 cross-species performance gap are species-specific transposable elements. For example, models trained
337 using mouse TF binding data have never seen an *Alu* SINE element during training, and often falsely
338 predict that these elements are bound by the relevant TF. Since *Alu* elements appear at high frequency in
339 the human genome, their misprediction constitutes a large proportion of the cross-species false positive
340 predictions, and thereby substantially affect the genome-wide performance metrics of the model. It should
341 be noted that *Alus* and other transposable elements can serve as true regulatory elements (Bourque et al.
342 2008; Sundaram et al. 2014), and thus we don't assume that all transposable elements should be labeled as
343 TF "unbound". Indeed, we minimized the potential mislabeling of truly bound transposable elements as
344 "unbound" by focusing all our analyses on regions of the genome that have a high degree of mappability
345 (and are thereby less likely to be subject to mappability-related false negative labeling issues in the TF
346 ChIP-seq data).

347 We demonstrated that a simple domain adaptation approach is sufficient to correct the systematic
348 mispredictions of *Alu* elements as TF bound. Training a parallel task (discriminating between species) but
349 with gradient reversal employed during backpropagation has the effect of discouraging species-specific
350 features being learned by the shared convolutional layers of the network. This approach is straightforward
351 to implement and has the advantage that TF binding labels need only be known in the training species.
352 Our approach accounts for domain shifts in the underlying genome sequence composition, assuming that
353 the general features of TF binding sites are conserved within the same cell types across species.

354 We note that the underlying assumption of cross-species TF binding prediction - i.e., that the overall
355 features of cell-specific TF binding sites are conserved - may not hold true in all cases. For some TFs,
356 concordant importance scores between mouse and human models across true-positive bound sites suggests
357 that both models learned similar representations of the TF's cognate motif. However, for other TFs, the
358 same analysis suggests that the models' representations of the sequences important for binding may not
359 completely agree. We also observe, particularly for those TFs with less concordant importance scores across
360 species, that there are sequence features in bound sites that discriminate between correct and incorrect
361 predictions specific to cross-species models. Therefore, cross-species false negative prediction errors could

362 be the result of differential TF activity across the two species. Such differential activities could result from
363 gain or loss of TF expression patterns, non-conserved cooperative binding capabilities, or evolved sequence
364 preferences of the TF itself. Our sequence composition domain adaptation approach is unlikely to address
365 situations where TF binding logic is not fully conserved across species.

366 Other recent work has also demonstrated the feasibility of cross-species regulatory imputation. For
367 example, Chen, et al. assessed the abilities of support vector machines (SVMs) and CNNs to predict po-
368 tential enhancers (defined by combinations of histone marks) when trained and tested across species of
369 varying evolutionary distances (Chen et al. 2018). They observed that while CNNs outperform SVMs in
370 within-species enhancer prediction tasks, they are worse at generalizing across species. Our work suggests
371 a possible reason for, and a solution to, this generalization gap. Two other recent manuscripts have ap-
372 plied more complex neural network architectures to impute TF binding and other regulatory signals across
373 species (Kelley 2020; Schreiber et al. 2020). Those studies focus on models that are trained jointly across
374 thousands of mouse and human regulatory genomic datasets. They thus assume that substantial amounts
375 of regulatory information has already been characterized in the target species, which may not be true in
376 some desired cross-species imputation settings. In general, however, joint modeling approaches are also
377 likely to benefit from domain adaptation strategies that account for species-specific differences in sequence
378 composition, and our results are thus complementary to these recent reports.

379 In summary, our work suggests that cross-species TF binding prediction approaches should beware of
380 systematic differences between the compositions of training and test species genomes, including species-
381 specific repetitive elements. Our contribution also suggests that domain adaptation is a promising strategy
382 for addressing such differences and thereby making cross-species predictions more robust. Further work is
383 needed to characterize additional sources of the cross-species performance gap and to generalize domain
384 adaptation approaches to scenarios where training data is available from multiple species.

385 Methods

386 Data processing

387 Datasets were constructed by splitting the mouse (mm10) and human (hg38) genomes, excluding sex chro-
388 mosomes, into 500 bp windows, offset by 50 bp. Any windows overlapping ENCODE blacklist regions

389 were removed (Amemiya et al. 2019). We then calculated the fraction of each window that was uniquely
390 mappable by 36 bp sequencing reads and retained only the windows that were at least 80% uniquely map-
391 pable (Karimzadeh et al. 2018). Mappability filtering was performed to remove potential peak-calling false
392 negatives; otherwise, any genomic window too unmappable for confident peak-calling would be a potential
393 false negative.

394 ChIP-seq experiments and corresponding controls (where available) were collected from ENCODE,
395 GEO, and ArrayExpress. Database accession IDs for all data used in this study are listed in Supplemen-
396 tal Tables S5, S6, and S7. We chose to focus our initial analyses on liver, as several previous studies have
397 provided matched ChIP-seq experiments characterizing orthologous TF binding across mammalian liver
398 samples (Schmidt et al. 2010; Odom et al. 2007). Our expanded analyses use erythroid, lymphoblast, and
399 ES cell line experiments that were previously compared across species by Denas, et al. (Denas et al. 2015).
400 We also analyzed matched adipocyte datasets that were performed on adipocyte cell lines within the same
401 labs (Schmidt et al. 2011; Mikkelsen et al. 2010). Additional datasets were sourced by searching the lit-
402 erature for ChIP-seq data targeting orthologous TFs in erythroid progenitor, megakaryocyte, macrophage,
403 and hematopoietic progenitor cell types (Tijssen et al. 2011; Hu et al. 2011; Pham et al. 2012; Pencovich et
404 al. 2013; Kaikkonen et al. 2013; Beck et al. 2013; Yue et al. 2014; Huang et al. 2016; Goode et al. 2016).

405 For cell types where all data was sourced from the mouse and human ENCODE projects (i.e., erythroid,
406 lymphoblast, and ES cell lines), we downloaded ChIP-seq narrow peak calls from the ENCODE portal. For
407 liver and all other cell types, we first aligned the FASTQ files to the mm10 and hg38 reference genomes
408 using Bowtie (version 1.3.0) (Langmead and Salzberg 2012). We then called ChIP-seq peaks using MultiGPS
409 v0.74 with default parameters, excluding ENCODE blacklist regions (Mahony et al. 2014; Amemiya et al.
410 2019). Corresponding control experiments were utilized during peak calling when available. Peak calls
411 were converted to binary labels for each window in a genome: “bound” (1) if any peak center fell within
412 the window, “unbound” (0) otherwise. Supplemental Table S5 shows the numbers of peaks called for liver
413 datasets, as well as the number of bound windows retained after filtering and the fraction of all retained
414 windows that are bound; Supplemental Tables S6 and S7 show the same information for all other datasets.
415 Candidate datasets were discarded from the analysis if the numbers of called peaks was less than 1000 in
416 mouse or human.

417 **Dataset splits for training and testing**

418 Chromosomes 1 and 2 of both species were held out from all training datasets. For computational effi-
419 ciency, one million randomly selected windows from Chromosome 1 were used as the validation set for
420 each species (for hyperparameter tuning). All windows from Chromosome 2 were used as the test sets.
421 Chromosomes X and Y were not used to avoid confounding because our matched datasets across species
422 did not always match in sex.

423 TF binding task training data was constructed identically for all model architectures. Since binary
424 classifier neural networks often perform best when the classes are balanced in the training data, the binding
425 task training dataset consisted of all bound examples and an equal number of randomly sampled (without
426 replacement) unbound examples, excluding examples from Chromosomes 1 and 2. To increase the diversity
427 of examples seen by the network across training, in each epoch a distinct random set of unbound examples
428 was used, with no repeated unbound examples across epochs.

429 Domain-adaptive models also require an additional “species-background” training set from both species
430 for the species discrimination task. Species-background data consisted of randomly selected (without re-
431 placement) examples from all chromosomes except 1, 2, X, and Y. Binding labels were **not** used in the
432 construction of these training sets. In each batch, the species-background examples were balanced, with
433 50% human and 50% mouse examples, and labeled according to their species of origin (not by binding).
434 The total number of species-background examples in each batch was double the number of binding exam-
435 ples.

436 **Basic model architecture**

437 The network takes in a one-hot encoded 500 bp window of DNA sequence and passes it through a convolu-
438 tional layer with 240 20-bp filters, followed by a ReLU activation and max-pooling (pool window and stride
439 of 15 bp). After the convolutional layer is an LSTM with 32 internal nodes, followed by a 1024-neuron
440 fully-connected layer with ReLU activation, followed by a 50% Dropout layer, followed by a 512-neuron
441 fully-connected layer with sigmoid activation. The final layer is a single sigmoid-activated neuron.

442 **Domain-adaptive model architecture**

443 The domain-adaptive network builds upon the basic model described above by adding a new “species
444 discriminator” task. The network splits into two output halves following max-pooling after the convolu-
445 tional layer. The max-pooling output feeds into a gradient reversal layer (GRL) – the GRL merely outputs
446 the identity of its input during the feed-forward step of model training, but during backpropagation, it
447 multiplies the gradient of the loss by -1 . The GRL is followed by a Flatten layer, a ReLU-activated fully
448 connected layer with 1024 neurons, a sigmoid-activated fully connected layer of 512 neurons, and finally a
449 single-neuron layer with sigmoid activation.

450 **Model training**

451 All models were trained with Keras v2.3.1 using the Adam optimizer with default parameters (Chollet
452 2015; Kingma and Ba 2014). Training ran for 15 epochs, with models saved after each epoch. After train-
453 ing, we selected models for downstream analysis by choosing the saved model with highest auPRC on the
454 training-species validation set.

455 The basic models were trained by standard procedure with a batch size of 400 (see Section 2.1.2 for
456 training dataset construction). The domain-adaptive models, on the other hand, required a more complex
457 batching setup. Because domain-adaptive models predict two tasks – binding and the species of origin of
458 the input sequence – they require two stages of dataset input per batch. The first stage is identical to a basic
459 model training batch, but with $\lfloor 400/3 \rfloor = 133$ binding examples from the source species. The second stage
460 uses $\lceil 400 * 2/3 \rceil = 267$ examples each from the source species’ and target species’ “species-background”
461 datasets.

462 Crucially, the stages differ in how task labels are masked. For each stage, only one of the two output
463 halves of the network trains (the loss backpropagates from one output only). In the first stage, we mask
464 the species discriminator task, so that only the binding task half of the model trains on binding examples
465 from the training species. In the second stage, we mask the binding task, so only the species discriminator
466 task half trains. Thus, the binding task only trains on examples from the source species, while the species
467 discriminator task doesn’t see binding labels from either species.

468 Meanwhile, the weights of the shared convolutional layer are influenced by both tasks. Because these
469 stages occur within a single batch and not in alternating batches, they concurrently influence the weights

470 of the convolutional filters; there is no oscillating “back-and-forth” between the two tasks from batch to
471 batch.

472 Model performance evaluations were computed with the sci-kit-learn v0.23 implementation of the
473 average_precision_score function, which closely approximates the area under the precision-recall curve
474 (auPRC).

475 **Differentially-predicted site categorization**

476 To quantify site enrichment within discrete categories such as “false positives” and “false negatives”, it was
477 necessary to define the boundaries for these labels. In particular, when comparing prediction distributions
478 between models, we needed to define what constitutes, for instance, a “false positive unique to model A .”
479 We constructed the following rules for site categorization: 1) unbound sites must have predictions above 0.5
480 to be labeled false positives, and bound sites must have predictions below 0.5 to be labeled false negatives;
481 2) a site is considered to be differentially predicted between two source species A and B if $|P_A - P_B| > 0.5$,
482 where P_A and P_B are the predictions from models trained on data from species A and species B , respectively;
483 3) only sites meeting this differential prediction threshold are labeled as a false positive or negative unique
484 to one model. Thus, if we are comparing models from species A and B , and a site is labeled a false positive
485 unique to model A , then $P_A > 0.5$ and $P_B < 0.5$. To reduce noise in these categorizations, rather than letting
486 P_A and P_B equal the predictions from single models, we trained 5 independent replicate models for each TF
487 and source species, and then let P_A be the average prediction across the 5 replicate models trained on data
488 from species A for a given TF.

489 **Bound site discriminative motif discovery**

490 SeqUnwinder (v. 0.1.3) (Kakumanu et al. 2017) was used to find motifs that discriminate between true
491 positive predictions and mouse-model-specific false negative predictions using the following command-
492 line settings: “--threads 10 --makerandregs --makerandregs --win 500 --mink 4 --maxk 5 --r 10 --x 3 --a 400
493 --hillsthresh 0.1 --memesearchwin 16”, and using MEME v. 5.1.0 (Machanick and Bailey 2011) internally.

494 **Repeat analysis**

495 All repeat analysis used the RepeatMasker track from the UCSC Genome Browser (Smit et al. 1996).
496 Genome windows were labeled as containing an *Alu* element if there was any overlap (1 or more bp) with
497 any *Alu* annotation. For Supplemental Table S1, repeat classes were excluded if fewer than 500 examples
498 of that class were annotated in the test chromosome (before mappability filtering).

499 **Gapped *k*-mer SVMs**

500 The gkmtrain and gkmpredict utilities from the lsgkm package were used for gkmSVMs gkm training and
501 prediction generation, respectively (Lee 2016). For training, 50000 examples each were selected randomly
502 from the set of all bound windows and unbound windows in the original neural network model train-
503 ing sets. Every 10th example from the original test set (in other words, sampling windows such that all
504 selected windows were non-overlapping) was considered in evaluation for computational efficiency. All
505 default parameters were used in running lsgkm (center-weighted + truncated *l*-mer kernel, word length
506 11, maximum 3 mismatches).

507 **Profile models**

508 Our profile model consists of a dilated convolutional residual model architecture that closely resembles the
509 BPNet architecture (Avsec et al. 2021b), with the following modifications: 1) 21bp-long filters in the first
510 convolutional layer, rather than 25bp; 2) 8 dilated convolutional layers, rather than 9; 3) a learning rate
511 of 0.001; 4) 2114 bases of sequence input. The first three hyperparameters were selected by tuning on the
512 source-species validation set loss; the sequence input length was chosen based on what would produce a
513 1000bp-long profile prediction given the 8-layer architecture's receptive field.

514 The profile models were trained using the same task and loss scheme as in Avsec et al. 2021b, with
515 the loss function value of λ set to 10. Training lasted 30 epochs, with early stopping used to select the best
516 model according to the source-species validation set profile (multinomial) loss. The training data used was
517 sampled from regions in the training set used by the binary models: specifically, each epoch the profile
518 model saw a 3:1 ratio of windows centered on peaks from training set chromosomes, with up to 200bp
519 jitter, and windows not overlapping peaks with a GC-content distribution that matched the set of peak-
520 centered windows. Hyperparameter tuning was performed using a combination of the BPNet multinomial

521 loss for the profile task, calculated on peaks from Chromosome 1, and auPRCs calculated using the same
522 validation set of 1 million random windows from Chromosome 1 that the binary models used. Final model
523 evaluation was performed on the full original test sets from Chromosome 2 used by the binary models.

524 **Importance scoring**

525 For a given 500bp window and model, importance scores were generated using a method similar to in
526 silico mutagenesis, which measures the change in model prediction when a given base and the region
527 immediately around it are ablated. First, ten independent dinucleotide-shuffled versions of the original
528 sequence were generated to serve as reference sequences unlikely to contain motifs. Next, the 5bp region
529 centered at a particular base was replaced with the corresponding 5bp region from one of the ten shuffled
530 sequences, and the post-sigmoid difference in model output for this ablated sequence was recorded. This
531 was repeated for all ten shuffled sequences, with the average model prediction differential reported as
532 the score for the base that the ablated region centered on. This process was repeated for all bases in the
533 sequence being scored.

534 **Software availability**

535 Open source code (MIT license) is provided in a Supplemental Code file and is also available from:
536 <https://github.com/seqcode/cross-species-domain-adaptation>

537 **Acknowledgements**

538 The authors thank the members of the Center for Eukaryotic Gene Regulation at Penn State and Jacob
539 Schreiber for helpful feedback and discussion. The authors also thank Daniela Uribe, Edgar Roman, and
540 Yishu Chen for their work replicating the findings pertaining to profile models in Supplemental Fig. S2.

541 **Funding**

542 This work was supported by NIH NIGMS grant R01GM121613 and NSF CAREER 2045500 (both to SM),
543 NIH NIGMS grant DP2GM123485 (to AK), and the Stanford Graduate Fellowship (to KC). RCH is sup-
544 ported by NIH NIDDK grant R24DK106766. The funders had no role in study design, data collection and
545 analysis, decision to publish, or preparation of the manuscript.

546 **Competing interest statement**

547 A.K. is a scientific co-founder of Ravel Biotechnology Inc, consultant with Illumina Inc., and on SAB of
548 OpenTargets, SerImmune and PatchBio. The other authors have no potential conflicts of interest to declare.

549 **References**

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

552 Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions
553 of the Genome. *Sci Rep* **9**: 9354.

554 Avsec Ž, Agarwal V, Visentin D, Ledsam J, Barwinska AG, Taylor K, Assael Y, Jumper J, Kohli P, Kelley D.
555 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat
556 Methods*. doi: <https://doi.org/10.1038/s41592-021-01252-x>.

557 Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J,
558 Kundaje A, et al. 2021. Base-resolution models of transcription factor binding reveal soft motif syntax.
559 *Nat Genet* **53**: 354–366.

560 Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370–379.

561 Beck D, Thoms JAI, Perera D, Schütte J, Unnikrishnan A, Knezevic K, Kinston SJ, Wilson NK, O'Brien TA,
562 Göttgens B, et al. 2013. Genome-wide analysis of transcriptional regulators in human HSPCs reveals a
563 densely interconnected network of coding and noncoding genes. *Blood* **122**: e12–22.

564 Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, et al.
565 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements.
566 *Genome Res* **18**: 1752–1762.

567 Bousmalis K, Silberman N, Dohan D, Erhan D, Krishnan D. 2017. Unsupervised Pixel-Level Domain Adap-
568 tation with Generative Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern*
569 *Recognition (CVPR)*, 95–104.

570 Chen L, Fish AE, Capra JA. 2018. Prediction of gene regulatory enhancers across species reveals evolution-
571 arily conserved sequence properties. *PLoS Comput Biol* **14**: e1006484.

572 Chollet F et al. 2015. Keras. <https://keras.io>.

573 Denas O, Sandstrom R, Cheng Y, Beal K, Herrero J, Hardison RC, Taylor J. 2015. Genome-wide comparative
574 analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics* **16**: 87.

575 Ferrari R, Llobet Cucalon LI de, Di Vona C, Le Dilly F, Vidal E, Lioutas A, Oliete JQ, Jochem L, Cutts E,
576 Dieci G, et al. 2019. TFIIIC Binding to Alu Elements Controls Gene Expression via Chromatin Looping
577 and Histone Acetylation. *Mol Cell* **77**: 475–487.

578 Fudenberg G, Kelley DR, Pollard KS. 2020. Predicting 3D genome folding from DNA sequence with Akita.
579 *Nat Methods* **17**: 1111–1117.

580 Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V. 2016.
581 Domain-Adversarial Training of Neural Networks. *J Mach Learn Res* **17**: 1–35.

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

584 Goode DK, Obier N, Vijayabaskar MS, Lie-A-Ling M, Lilly AJ, Hannah R, Lichtinger M, Batta K, Florkowska
585 M, Patel R, et al. 2016. Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and
586 Differentiation. *Dev Cell* **36**: 572–587.

587 Hu G, Schones DE, Cui K, Ybarra R, Northrup D, Tang Q, Gattinoni L, Restifo NP, Huang S, Zhao K. 2011.
588 Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by
589 BRG1. *Genome Res* **21**: 1650–1658.

590 Huang J, Liu X, Li D, Shao Z, Cao H, Zhang Y, Trompouki E, Bowman T, Zon LI, Yuan GC, et al. 2016. Dy-
591 namic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis.
592 *Dev Cell* **36**: 9–23.

593 Huh I, Mendizabal I, Park T, Yi SV. 2018. Functional conservation of sequence determinants at rapidly
594 evolving regulatory regions across mammals. *PLoS Comput Biol* **14**: e1006451.

595 Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha R,
596 Benner CK, et al. 2013. Remodeling of the enhancer landscape during macrophage activation is coupled
597 to enhancer transcription. *Mol Cell* **51**: 310–325.

598 Kakumanu A, Velasco S, Mazzoni E, Mahony S. 2017. Deconvolving sequence features that discriminate
599 between overlapping regulatory annotations. *PLoS Comput Biol* **13**: e1005795.

600 Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. 2018. Umap and Bismap: quantifying genome and
601 methylome mappability. *Nucleic Acids Res* **46**: e120.

602 Kelley DR. 2020. Cross-species regulatory sequence activity prediction. *PLoS Comput Biol* **16**: e1008050.

603 Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. 2018. Sequential regulatory activity
604 prediction across chromosomes with convolutional neural networks. *Genome Res* **28**: 739–750.

605 Kingma DP, Ba J. 2014. Adam: A Method for Stochastic Optimization. arXiv: 1412.6980 [cs.LG].

606 Koo PK, Majdandzic A, Ploenzke M, Anand P, Paul SB. 2021. Global importance analysis: An interpretabil-
607 ity method to quantify importance of genomic features in deep neural networks. *PLoS Comput Biol* **17**:
608 e1008925.

609 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

610 Lee D. 2016. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**: 2196–2198.

611 Long M, Cao Y, Wang J, Jordan MI. 2015. Learning Transferable Features with Deep Adaptation Networks.
612 *2015 International Conference of Machine Learning (ICML)*, 97–105.

613 Machanick P, Bailey TL. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**: 1696–
614 1697.

615 Mahony S, Edwards MD, Mazzoni EO, Sherwood RI, Kakumanu A, Morrison CA, Wichterle H, Gifford DK.
616 2014. An Integrated Model of Multiple-Condition ChIP-Seq Data Reveals Predeterminants of Cdx2
617 Binding. *PLoS Comput Biol* **10**: e1003501.

618 Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, Rosen ED. 2010. Comparative epigenomic
619 analysis of murine and human adipogenesis. *Cell* **143**: 156–69.

620 Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford
621 DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between hu-
622 man and mouse. *Nat Genet* **39**: 730–732.

623 Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**: 669–680.

624 Pencovich N, Jaschek R, Dicken J, Amit A, Lotem J, Tanay A, Groner Y. 2013. Cell-autonomous function of
625 Runx1 transcriptionally regulates mouse megakaryocytic maturation. *PLoS One* **8**: e64248.

626 Pham TH, Benner C, Lichtinger M, Schwarzfischer L, Hu Y, Andreesen R, Chen W, Rehli M. 2012. Dynamic
627 epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentia-
628 tion states. *Blood* **24**: e161–171.

629 Polak P, Domany E. 2006. Alu elements contain many binding sites for transcription factors and may play
630 a role in regulation of developmental processes. *BMC Genomics* **7**: 133.

631 Quang D, Xie X. 2016. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying
632 the function of DNA sequences. *Nucleic Acids Res* **44**: e107.

633 — 2019. FactorNet: A deep learning framework for predicting cell type specific transcription factor bind-
634 ing from nucleotide-resolution sequential data. *Methods* **166**: 40–47.

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

638 Savic D, Partridge EC, Newberry KM, Smith SB, Meadows SK, Roberts BS, Mackiewicz M, Mendenhall EM,
639 Myers RM. 2015. CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. *Genome Res*
640 **25**: 1581–1589.

641 Schmid CW. 1998. Does SINE evolution preclude Alu function? *Nucleic Acids Res* **26**: 4541–4550.

642 Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-
643 Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of tran-
644 scription factor binding. *Science* **328**: 1036–1040.

645 Schmidt SF, Jørgensen M, Chen Y, Nielsen R, Sandelin A, Mandrup S. 2011. Cross species comparison
646 of C/EBP α and PPAR γ profiles in mouse and human adipocytes reveals interdependent retention of
647 binding sites. *BMC Genomics* **12**: 152.

648 Schreiber J, Hegde D, Noble W. 2020. Zero-shot imputations across species are enabled through joint mod-
649eling of human and mouse epigenomics. *ACM International Conference on Bioinformatics, Computational*
650 *Biology and Health Informatics*. doi: 10.1145/3388440.3412412.

651 Siggers T, Gordan R. 2014. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res*
652 **42**: 2099–2111.

653 Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordan R, Rohs R. 2014. Absence of a simple code: how
654 transcription factors read the genome. *Trends Biochem Sci* **39**: 381–399.

655 Smit A, Hubley R, Green P. 1996-2010. RepeatMasker Open-3.0. <http://www.repeatmasker.org>.

656 Srivastava D, Aydin B, Mazzoni EO, Mahony S. 2020. An interpretable bimodal neural network character-
657 izes the sequence and preexisting chromatin predictors of induced TF binding. *Genome Biol* **22**: 20.

658 Srivastava D, Mahony S. 2020. Sequence and chromatin determinants of transcription factor binding and
659 the establishment of cell type-specific binding patterns. *BBA Gene Regul Mech* **1863**: 194443.

660 Sun B, Feng J, Saenko K. 2016. Correlation Alignment for Unsupervised Domain Adaptation. arXiv: 1612.
661 01939 [cs].

662 Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of
663 transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**: 1963–1976.

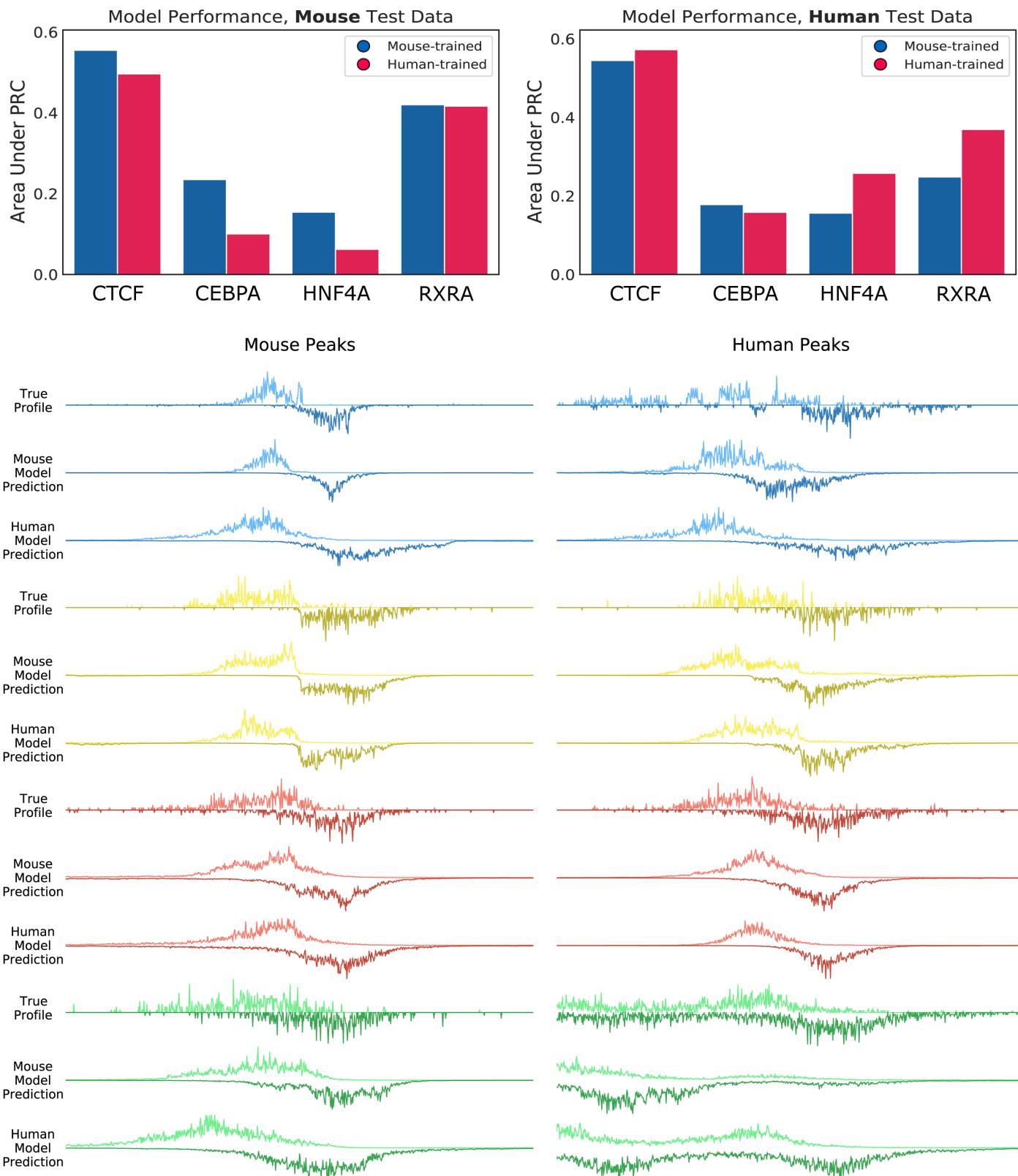
664 The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human
665 genome. *Nature* **489**: 57–74.

666 The Mouse ENCODE Consortium, Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma
667 Z, Davis C, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**:
668 355–364.

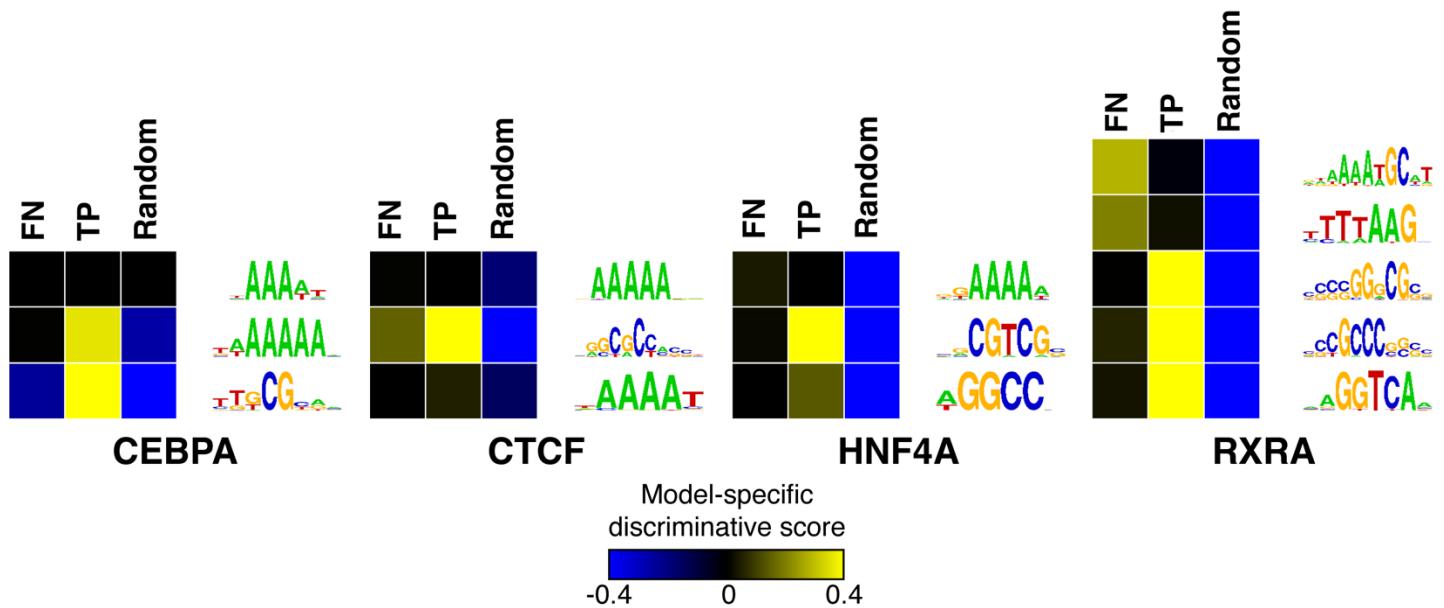
669 Tijssen M, Cvejic A, Joshi A, Hannah R, Ferreira R, Forrai A, Bellissimo D, Oram SH, Smethurst P, Wilson
670 N, et al. 2011. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in
671 megakaryocytes identifies hematopoietic regulators. *Dev Cell* **20**: 597–609.



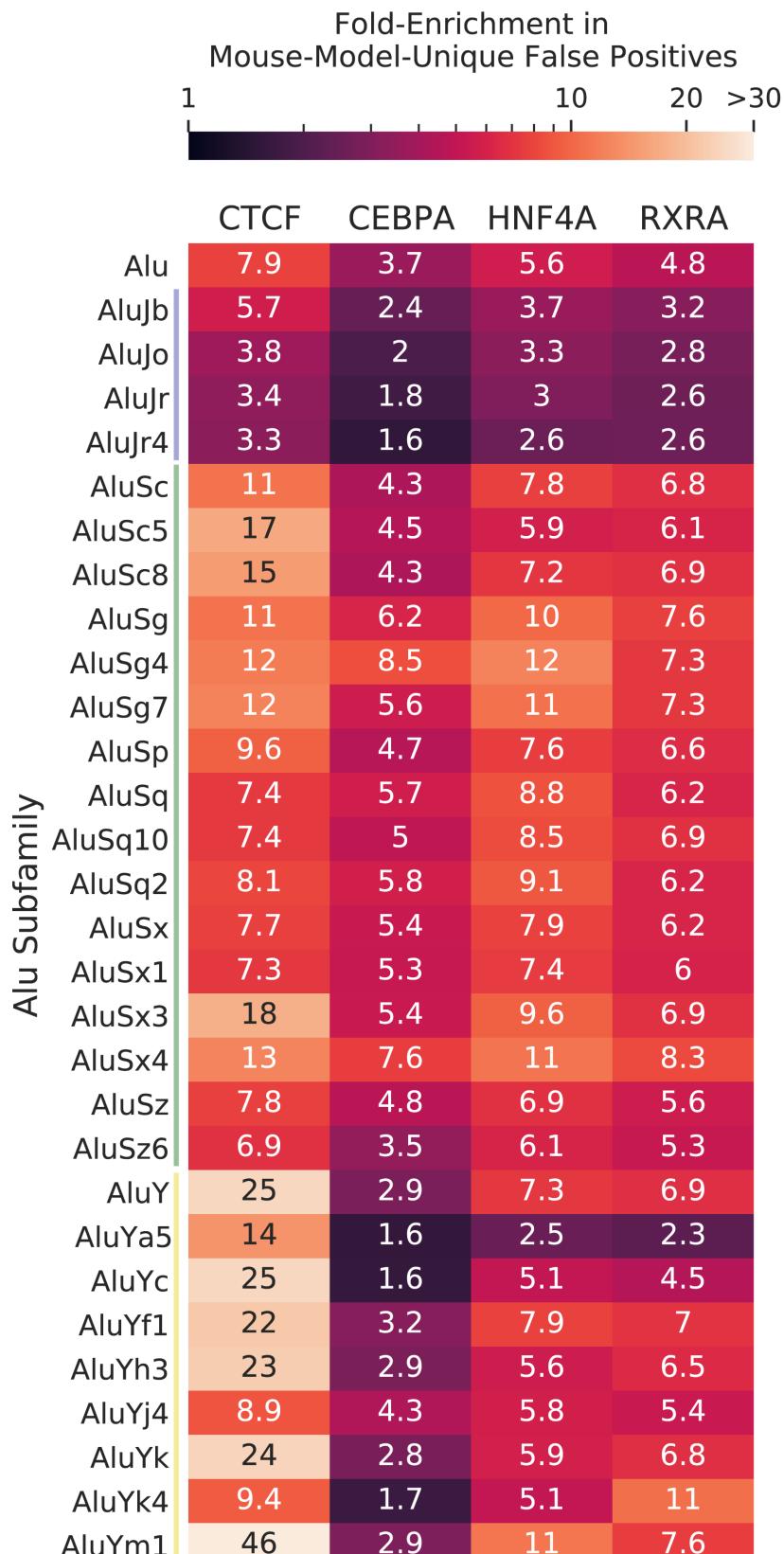
Supplemental Figure 1: Results of evaluating the performance of mouse-trained (blue) and human-trained (red) gapped k -mer SVM models on non-overlapping windows from the mouse (left) and human (right) test datasets (Chromosome 2). For each TF and species, an SVM was trained using a balanced set of bound and unbound windows from the original training set.



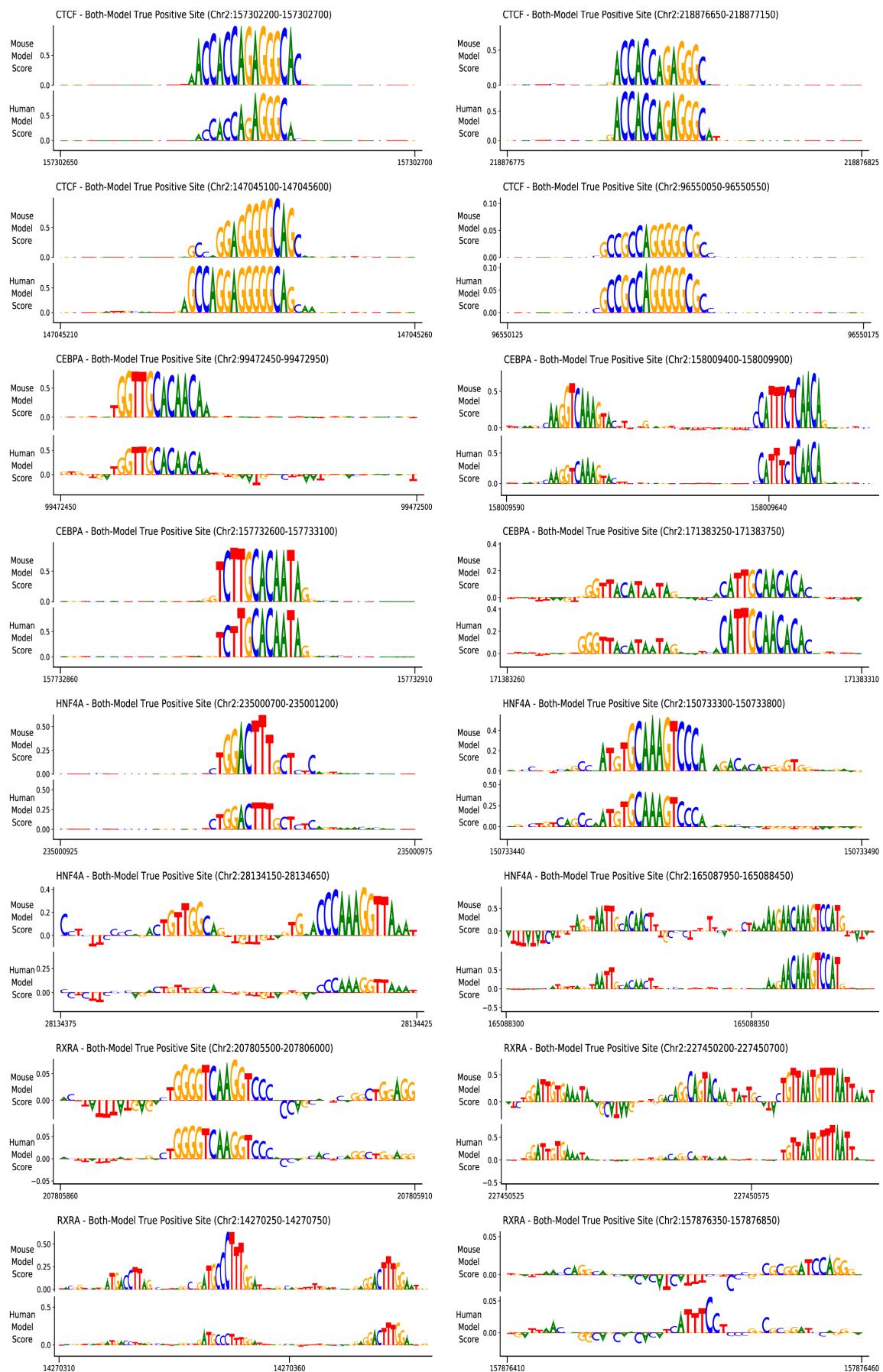
Supplemental Figure 2: Top: results from evaluating the performance of mouse-trained (blue/left) and human-trained (red/right) profile models on non-overlapping windows from the human (left panel) and mouse (right panel) test datasets (Chromosome 2). Bottom: example predicted vs. true ChIP-seq read distribution profiles from mouse (left) and human (right) test datasets. See Methods for profile model details.



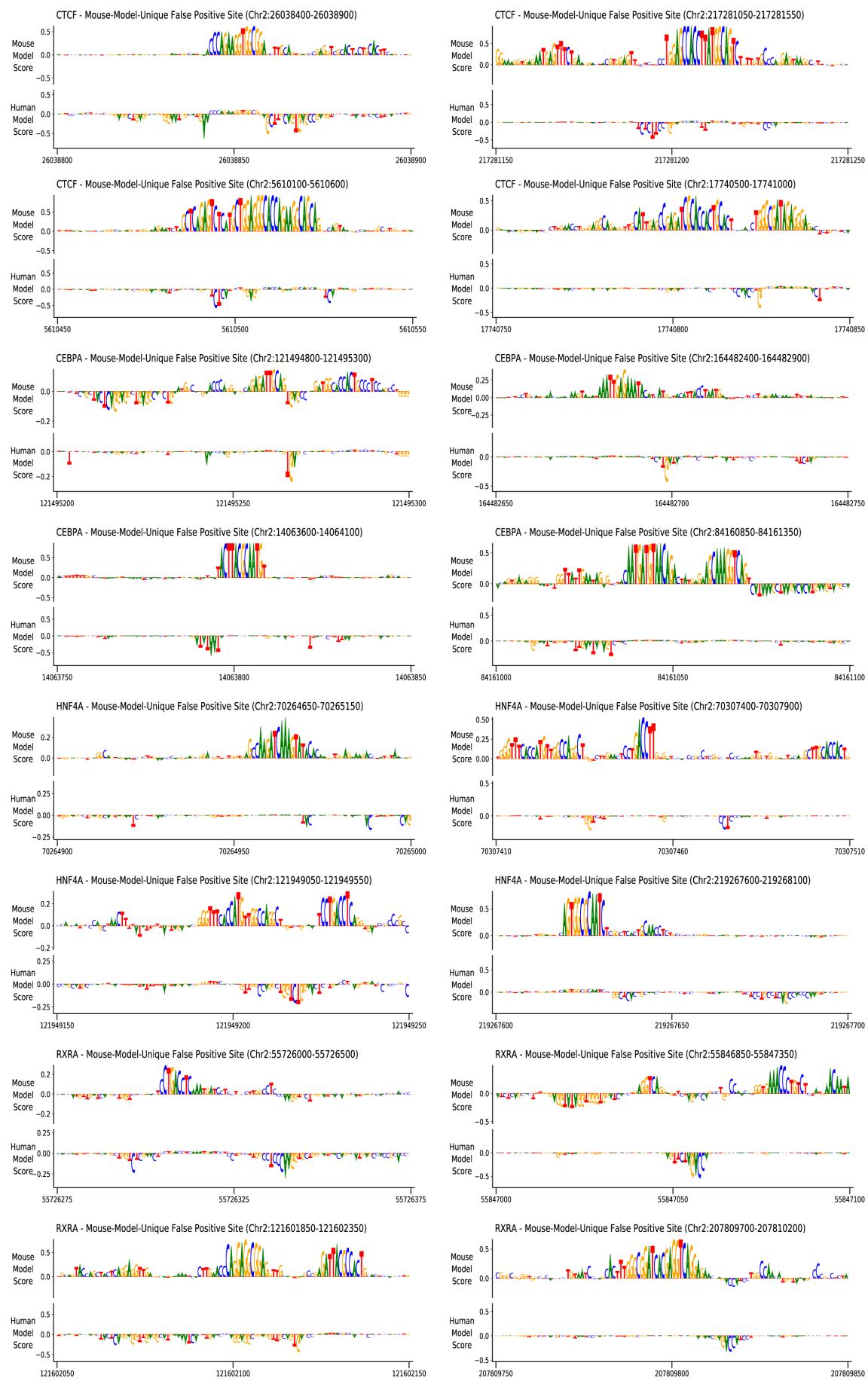
Supplemental Figure 3: Motif-like sequence features can discriminate between human-genome bound sites correctly predicted by mouse-trained and human-trained models (true positives or TP) and bound sites correctly predicted only by human-trained models (mouse-specific false negatives or FN) for each TF. See Methods for site categorization details.



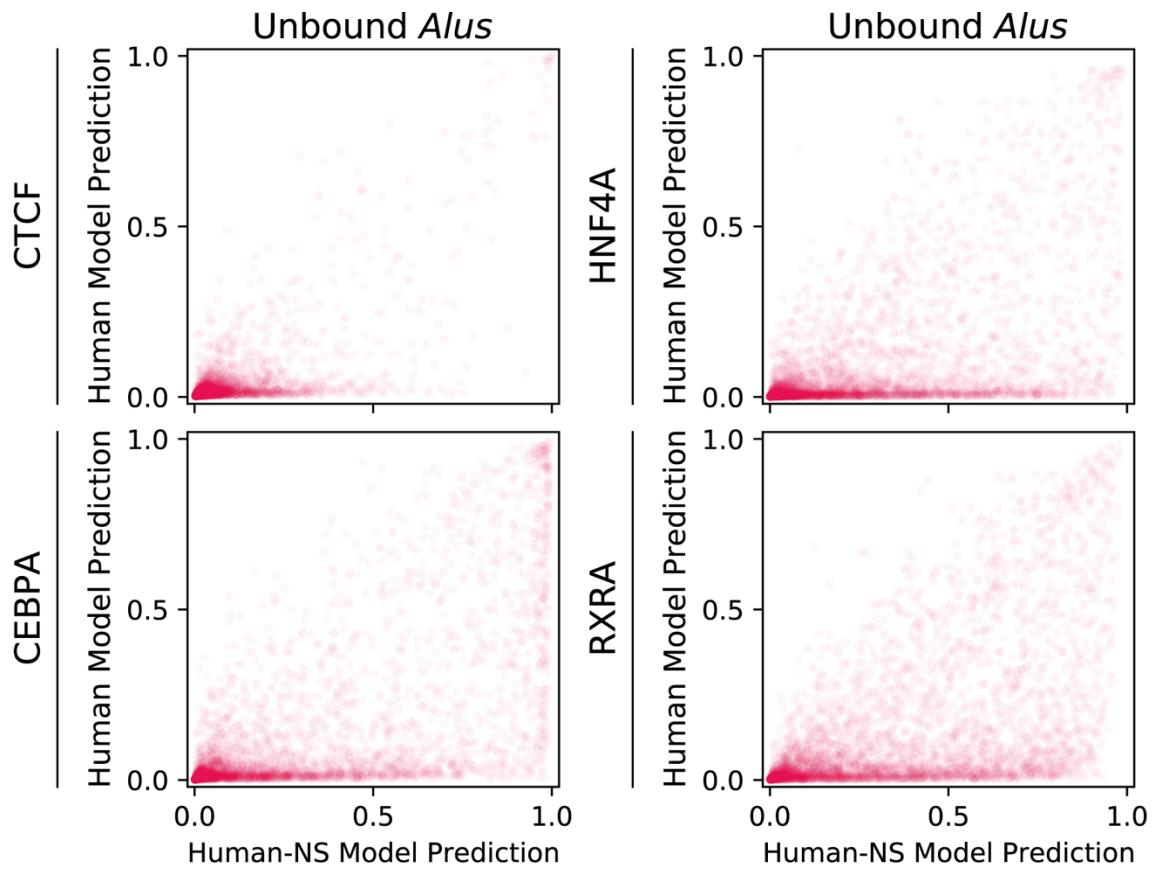
Supplemental Figure 4: Enrichment of specific *Alu* subfamilies within the set of false positives unique to the mouse-trained model, relative to false positives common to both mouse-trained and human-trained models. For a given TF and *Alu* subfamily, the fraction of windows overlapping any RepeatMasker-annotated instance of that repeat type were calculated for both classes of false positives. The values in the figure show the ratio between *Alu* fractions in mouse-model-unique false positives and both-model false positives. Only *Alu* subfamilies with annotations overlapping at least 500 examples in the test dataset (Chromosome 2) and covering at least 0.01% of the both-model false positive category are included.



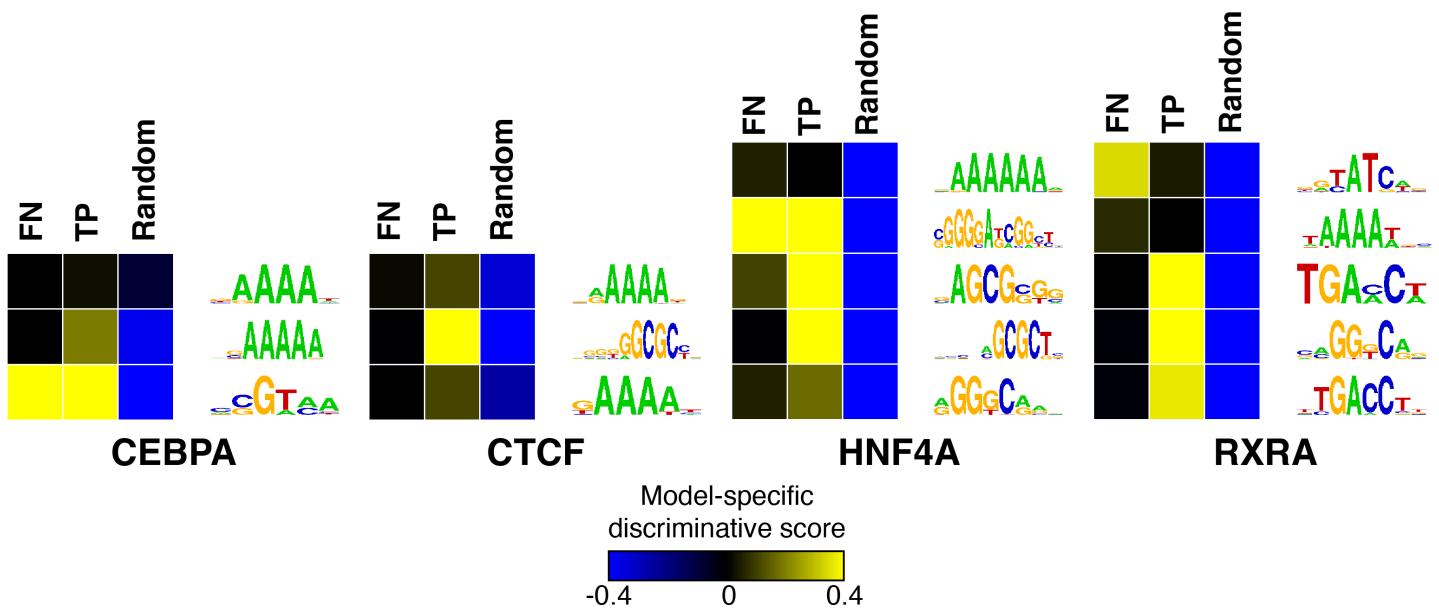
Supplemental Figure 5: Importance scores for example both-model true positive sites for the four TFs. Bases were scored using a modified ISM algorithm (see Methods). The 500bp example sites have been enlarged and cropped around motif-like instances for readability.



Supplemental Figure 6: Importance scores for example false positive sites mispredicted only by the mouse model. Bases were scored using a modified ISM algorithm (see Methods).



Supplemental Figure 7: Comparison between the predictions of human-trained models that were trained without examples overlapping SINEs (x-axis) to the predictions of standard human-trained models (y-axis). Unbound *Alu* repeats make up a large part of the false positives unique to the no-SINEs model. For visual clarity, only 5% of windows are shown.



Supplemental Figure 8: False negative predictions unique to mouse-trained models trained with domain adaptation, compared to human-trained models, can be distinguished from true positive predictions through motif-like sequence features. See Methods for site categorization details.

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
DNA						
CTCF	10.1%	11.4%	7.3%	11.4%	8.9%	9.0%
CEBPA	12.3%	10.4%	8.3%	11.3%	13.0%	9.2%
HNF4A	10.7%	12.0%	11.3%	11.4%	9.5%	9.0%
RXRA	10.1%	11.7%	8.8%	11.4%	10.0%	9.4%
LINE						
CTCF	18.3%	22.5%	21.3%	37.6%	17.8%	31.6%
CEBPA	25.6%	26.3%	25.0%	37.6%	29.0%	32.3%
HNF4A	21.0%	25.3%	26.3%	37.6%	21.5%	30.5%
RXRA	21.0%	27.9%	22.1%	37.8%	22.1%	33.2%
Low Complexity						
CTCF	2.5%	1.0%	2.6%	1.9%	4.0%	1.5%
CEBPA	1.5%	1.4%	0.0%	1.9%	1.6%	1.4%
HNF4A	2.0%	1.1%	1.5%	1.9%	2.4%	1.3%
RXRA	2.1%	1.3%	2.0%	1.9%	2.2%	1.5%
LTR						
CTCF	8.6%	12.6%	8.2%	17.6%	16.4%	15.5%
CEBPA	12.8%	12.5%	41.7%	17.6%	19.4%	14.0%
HNF4A	13.3%	15.0%	16.0%	17.6%	19.4%	12.6%
RXRA	12.2%	14.8%	9.8%	17.6%	18.2%	11.5%
Simple Repeat						
CTCF	13.4%	10.1%	10.9%	11.5%	15.8%	12.0%
CEBPA	9.4%	8.6%	25.0%	11.6%	9.8%	11.9%
HNF4A	12.3%	11.4%	9.1%	11.5%	11.6%	12.8%
RXRA	11.5%	9.4%	13.3%	11.5%	11.1%	13.4%
SINE						
CTCF	23.9%	23.2%	20.6%	31.2%	18.4%	81.1%
CEBPA	30.9%	22.6%	41.7%	31.2%	35.2%	86.9%
HNF4A	27.1%	23.6%	22.6%	31.2%	28.9%	95.5%
RXRA	27.1%	25.3%	18.9%	31.3%	32.2%	97.6%
Unknown						
CTCF	0.2%	0.1%	0.0%	0.2%	0.1%	0.0%
CEBPA	0.3%	0.5%	0.0%	0.1%	0.2%	0.0%
HNF4A	0.2%	0.1%	0.0%	0.2%	0.2%	0.0%
RXRA	0.2%	0.3%	0.0%	0.1%	0.2%	0.0%

Table S1: Percent of windows overlapping various RepeatMasker-defined repeat elements, for different categories of genomic windows from the held-out test set. Only RepeatMasker repeat classes with at least 500 distinct annotations within the testset are shown. FPs: false positives. FNs: false negatives. Mouse Only: specific to mouse-trained models. See Methods for more details on site categorization.

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	12.6%	13.5%	13.7%	21.3%	9.0%	28.8%
CEBPA	18.3%	16.8%	0.0%	21.3%	21.9%	49.5%
HNF4A	13.6%	14.8%	13.7%	21.3%	14.0%	34.3%
RXRA	13.7%	17.7%	10.7%	21.4%	15.8%	58.7%

Table S2: Percent of windows overlapping an *Alu* element when domain-adaptive mouse models are compared to human models (compare to Table 1). The fraction of mouse-model-unique false positives overlapping *Alu* elements (right-most column) have decreased notably for all TFs. FPs: false positives. FNs: false negatives.

TF	auPRC, Mouse Test Set			auPRC, Human Test Set		
	Mouse(Basic)	Mouse(+DA)	Human	Mouse(Basic)	Mouse(+DA)	Human
Adipocytes						
CEBPA	0.18	0.18	0.17	0.18	0.21	0.35
CTCF	0.67	0.66	0.55	0.56	0.56	0.62
PPARG	0.08	0.07	0.08	0.07	0.06	0.22
Erythroid Cells						
BHLHE40	0.09	0.09	0.07	0.13	0.13	0.19
CTCF	0.71	0.68	0.62	0.60	0.58	0.67
E2F4	0.10	0.07	0.09	0.17	0.17	0.23
ELF1	0.28	0.28	0.27	0.26	0.26	0.34
ETS1	0.16	0.16	0.05	0.11	0.10	0.21
GATA1	0.20	0.19	0.11	0.09	0.09	0.10
JUND	0.05	0.03	0.02	0.10	0.09	0.26
MAFK	0.14	0.12	0.14	0.17	0.16	0.39
MAX	0.18	0.18	0.14	0.19	0.20	0.26
MAZ	0.15	0.15	0.14	0.21	0.22	0.32
MEF2A	0.03	0.01	0.02	0.02	0.01	0.04
MXI1	0.20	0.21	0.16	0.09	0.10	0.10
MYC	0.14	0.14	0.09	0.17	0.18	0.23
NRF1	0.33	0.32	0.22	0.33	0.35	0.36
TAL1	0.14	0.14	0.11	0.14	0.14	0.19
UBTF	0.15	0.15	0.15	0.19	0.19	0.23
USF1	0.21	0.18	0.16	0.17	0.16	0.25
USF2	0.12	0.11	0.09	0.14	0.16	0.13
Erythroid Progenitors						
CTCF	0.69	0.67	0.57	0.60	0.59	0.67
GATA1	0.09	0.09	0.08	0.10	0.08	0.16
TAL1	0.06	0.04	0.07	0.08	0.07	0.21
ESCs						
CTCF	0.78	0.76	0.71	0.53	0.54	0.66
MAFK	0.43	0.40	0.40	0.31	0.28	0.34
NANOG	0.14	0.12	0.05	0.05	0.05	0.08
POU5F1	0.11	0.10	0.09	0.07	0.06	0.09
Hematopoietic Progenitors						
FLI1	0.21	0.16	0.09	0.06	0.06	0.17
LMO2	0.06	0.04	0.00	0.00	0.00	0.01
RUNX1	0.06	0.04	0.05	0.05	0.05	0.20
SPI1	0.32	0.28	0.32	0.38	0.38	0.62

Continued on next page.

TF	auPRC, Mouse Test Set			auPRC, Human Test Set		
	Mouse(Basic)	Mouse(+DA)	Human	Mouse(Basic)	Mouse(+DA)	Human
Lymphoblasts						
BHLHE40	0.23	0.21	0.15	0.13	0.14	0.17
CTCF	0.70	0.69	0.58	0.63	0.61	0.65
E2F4	0.12	0.09	0.12	0.12	0.11	0.13
ELF1	0.32	0.30	0.27	0.34	0.34	0.34
ETS1	0.16	0.15	0.05	0.05	0.05	0.19
IRF4	0.23	0.22	0.14	0.11	0.10	0.14
JUND	0.09	0.07	0.05	0.04	0.04	0.07
MAX	0.17	0.17	0.13	0.17	0.18	0.19
MAZ	0.13	0.12	0.12	0.20	0.20	0.24
MEF2A	0.16	0.14	0.09	0.06	0.06	0.11
MXI1	0.19	0.20	0.18	0.14	0.15	0.16
MYC	0.14	0.14	0.07	0.08	0.10	0.11
NRF1	0.32	0.30	0.25	0.38	0.34	0.45
TBP	0.16	0.15	0.14	0.09	0.09	0.11
TCF12	0.24	0.23	0.17	0.12	0.11	0.14
USF1	0.22	0.20	0.17	0.20	0.20	0.19
USF2	0.16	0.15	0.12	0.10	0.10	0.09
Macrophages						
SPI1	0.41	0.41	0.33	0.29	0.30	0.46
Megakaryocytes						
FLI1	0.26	0.15	0.22	0.15	0.07	0.16
GATA1	0.09	0.08	0.02	0.03	0.02	0.04
RUNX1	0.08	0.06	0.04	0.13	0.12	0.28

Table S3: Average auPRC values from evaluating the basic mouse models, domain-adaptive mouse models, and basic human models on the mouse (left columns) and human (right columns) test sets, across all additional datasets beyond the primary liver TFs. The auPRCs shown are the average across three replicate model trainings for basic mouse-trained and human-trained models and across two replicate model trainings for domain-adaptive mouse models. Note that because the auPRC metric depends on the sparsity of the positive class (bound sites), these values are not comparable across test sets, across TFs, or across cell types.

TF	Basic Mouse Models			Domain-adaptive Mouse Models		
	Unbound	FPs (Both)	FPs (Mouse Only)	Unbound	FPs (Both)	FPs (Mouse Only)
Adipocytes						
CEBPA	21.3%	20.5%	76.5%	21.3%	19.5%	59.9%
CTCF	21.3%	11.4%	86.7%	21.3%	9.8%	31.3%
PPARG	21.4%	15.9%	68.1%	21.4%	14.8%	50.7%
Erythroid Progenitors						
CTCF	21.3%	9.7%	42.0%	21.3%	9.8%	24.6%
GATA1	21.3%	15.1%	62.1%	21.3%	14.5%	64.4%
TAL1	21.3%	17.1%	86.3%	21.3%	14.0%	76.8%
Erythroid Cells						
BHLHE40	21.3%	17.2%	63.5%	21.3%	16.0%	49.6%
CTCF	21.3%	13.6%	61.4%	21.3%	11.6%	34.0%
E2F4	21.3%	8.8%	71.1%	21.3%	8.2%	44.1%
ELF1	21.3%	13.7%	58.9%	21.3%	12.8%	40.5%
ETS1	21.3%	9.7%	35.0%	21.3%	8.9%	26.5%
GATA1	21.3%	14.1%	41.8%	21.3%	13.7%	28.7%
JUND	21.3%	20.6%	65.0%	21.3%	17.6%	54.6%
MAFK	21.3%	15.7%	56.4%	21.3%	14.2%	35.2%
MAX	21.3%	14.0%	72.2%	21.3%	13.0%	56.9%
MAZ	21.3%	12.3%	87.6%	21.3%	11.2%	60.9%
MEF2A	21.3%	12.5%	41.2%	21.3%	13.0%	36.3%
MXI1	21.3%	13.2%	81.7%	21.3%	11.9%	55.9%
MYC	21.3%	15.3%	68.5%	21.3%	14.0%	43.6%
NRF1	21.3%	8.9%	65.4%	21.3%	8.8%	47.2%
TAL1	21.3%	16.1%	38.8%	21.3%	15.4%	48.7%
UBTF	21.3%	10.6%	97.6%	21.3%	9.4%	84.5%
USF1	21.3%	14.8%	66.4%	21.3%	14.2%	55.3%
USF2	21.3%	11.9%	73.8%	21.3%	11.5%	62.8%
ESCs						
CTCF	21.4%	14.9%	82.5%	21.4%	12.2%	31.5%
MAFK	21.3%	13.5%	80.7%	21.3%	13.4%	76.8%
NANOG	21.3%	10.3%	38.8%	21.3%	9.4%	26.7%
POU5F1	21.3%	11.0%	40.9%	21.3%	9.8%	22.7%
Hematopoietic Progenitors						
FLI1	21.3%	6.9%	49.4%	21.3%	6.6%	34.5%
LMO2	21.3%	6.0%	20.0%	21.3%	6.4%	29.6%
RUNX1	21.3%	8.4%	25.5%	21.3%	7.9%	30.0%
SPI1	21.5%	13.6%	9.0%	21.5%	13.0%	17.4%

Continued on next page.

TF	Basic Mouse Models			Domain-adaptive Mouse Models		
	Unbound	FPs (Both)	FPs (Mouse Only)	Unbound	FPs (Both)	FPs (Mouse Only)
Lymphoblasts						
BHLHE40	21.3%	15.3%	67.5%	21.3%	14.0%	29.3%
CTCF	21.3%	11.4%	70.8%	21.3%	10.8%	33.8%
E2F4	21.3%	6.8%	51.2%	21.3%	6.6%	25.8%
ELF1	21.3%	11.8%	41.5%	21.3%	10.4%	18.2%
ETS1	21.3%	10.4%	32.8%	21.3%	9.7%	20.7%
IRF4	21.3%	13.3%	52.9%	21.3%	12.2%	21.0%
JUND	21.3%	12.6%	28.5%	21.3%	12.1%	25.8%
MAX	21.3%	8.6%	61.3%	21.3%	8.2%	30.1%
MAZ	21.3%	8.5%	50.4%	21.3%	8.2%	26.6%
MEF2A	21.3%	17.3%	64.5%	21.3%	13.8%	14.7%
MXI1	21.3%	11.6%	77.8%	21.3%	10.6%	43.6%
MYC	21.3%	6.3%	47.6%	21.3%	5.8%	31.7%
NRF1	21.3%	5.6%	26.6%	21.3%	5.6%	17.8%
TBP	21.3%	14.3%	89.2%	21.3%	12.4%	56.4%
TCF12	21.3%	11.6%	51.5%	21.3%	10.4%	17.5%
USF1	21.3%	11.0%	64.6%	21.3%	10.5%	63.1%
USF2	21.3%	11.4%	71.1%	21.3%	11.0%	71.2%
Macrophages						
SPI1	21.4%	15.5%	27.2%	21.4%	14.2%	13.8%
Megakaryocytes						
FLI1	21.3%	8.4%	23.0%	21.3%	7.6%	18.1%
GATA1	21.3%	7.6%	11.3%	21.3%	8.2%	19.4%
RUNX1	21.3%	7.8%	15.7%	21.3%	8.2%	15.4%

Table S4: The percent of sites overlapping an *Alu* element without (left column set) or with domain adaptation (right column set), for each of the additional datasets included in Figure 12. FPs: false positives; either the set of unbound sites mispredicted as bound by both the mouse model and the human model, or false positives mispredicted by the mouse model only. See Methods for site categorization details.

TF	Species	Raw Peaks	Filtered Peaks	Bound Windows	Frac. Bound	Accession ID
CTCF	Mouse	32006	28943	296117	0.71%	ENCSR000CBU
	Human	29067	26477	270100	0.55%	ENCSR911GFJ
CEBPA	Mouse	62636	48812	566945	1.35%	E-TABM-722
	Human	32243	28545	298066	0.61%	E-TABM-722
HNF4A	Mouse	44800	36540	415846	0.99%	E-TABM-722
	Human	42766	34714	387077	0.79%	E-TABM-722
RXRA	Mouse	46443	33751	404284	0.97%	GSM1299600
	Human	95085	71032	854289	1.75%	ENCSR098XMN

Table S5: For the primary experimental data used in this study, the following quantities are listed: the number of peaks called across the entire genome; the number of called peaks within the filtered window set, merged if within 500 bp of each other; the number of windows in the filtered window set labeled bound due to peak overlap; the fraction of the filtered window set labeled bound; and the database accession ID (ENCODE, GEO, or ArrayExpress). The size of the filtered window sets for the mouse and human genomes were 41883806 and 48742577, respectively.

TF	Cell Type	Peaks	Bound Windows	Frac. Bound	Accession ID
CEBPA	Adipocyte	15458	117291	0.26%	GSE27450
CTCF	Adipocyte	53354	431411	0.97%	GSE20752
PPARG	Adipocyte	11465	93351	0.21%	GSE20752
BHLHE40	Erythroid	16967	142473	0.32%	ENCSR000ESH
CTCF	Erythroid	47297	397943	0.90%	ENCSR000ETQ
E2F4	Erythroid	4925	39272	0.09%	ENCSR000ETY
ELF1	Erythroid	19518	162146	0.37%	ENCSR033OWC
ETS1	Erythroid	40913	337420	0.76%	ENCSR000ETB
GATA1	Erythroid	46171	384871	0.87%	ENCSR000EUG
JUND	Erythroid	7277	60316	0.14%	ENCSR000ETZ
MAFK	Erythroid	9846	81190	0.18%	ENCSR000ETK
MAX	Erythroid	28616	228588	0.52%	ENCSR000ETX
MAZ	Erythroid	20486	162023	0.37%	ENCSR000ESL
MEF2A	Erythroid	4827	40776	0.09%	ENCSR867SDZ
MXI1	Erythroid	39255	298761	0.67%	ENCSR000ETN
MYC	Erythroid	29362	230225	0.52%	ENCSR000EUA
NRF1	Erythroid	11191	93732	0.21%	ENCSR135SWH
TAL1	Erythroid	18775	156858	0.35%	ENCSR000DIA
UBTF	Erythroid	6549	51131	0.12%	ENCSR000ESJ
USF1	Erythroid	19958	167188	0.38%	ENCSR705HGT
USF2	Erythroid	4186	35607	0.08%	ENCSR000ETF
CTCF	Erythroid progenitor	30998	256447	0.58%	GSE36029
GATA1	Erythroid progenitor	11162	93403	0.21%	GSE36029
TAL1	Erythroid progenitor	4356	34771	0.08%	GSE36029
CTCF	ESC	44581	371955	0.84%	ENCSR362VNF
MAFK	ESC	17123	142736	0.32%	ENCSR604XDL
NANOG	ESC	16037	130587	0.29%	ENCSR779CZG
POU5F1	ESC	4198	34197	0.08%	ENCSR392DGA
FLI1	Hematopoietic progenitor	9838	83763	0.19%	GSE69099
LMO2	Hematopoietic progenitor	4618	39374	0.09%	GSE69099
RUNX1	Hematopoietic progenitor	2888	24860	0.06%	GSE69099
SPI1	Hematopoietic progenitor	17273	145004	0.33%	GSE69099
BHLHE40	Lymphoblast	46376	382831	0.86%	ENCSR000ERC
CTCF	Lymphoblast	63983	535673	1.21%	ENCSR000ERM
E2F4	Lymphoblast	6263	50365	0.11%	ENCSR000ERU
ELF1	Lymphoblast	28901	237091	0.54%	ENCSR293WTN
ETS1	Lymphoblast	30839	252203	0.57%	ENCSR000ERA
IRF4	Lymphoblast	43458	357105	0.81%	ENCSR743ZJL
JUND	Lymphoblast	15998	131797	0.30%	ENCSR000ERR
MAX	Lymphoblast	31198	250415	0.57%	ENCSR000ERL
MAZ	Lymphoblast	21334	169037	0.38%	ENCSR000EQT
MEF2A	Lymphoblast	30789	250229	0.57%	ENCSR806JZK
MXI1	Lymphoblast	32224	244761	0.55%	ENCSR000ERE
MYC	Lymphoblast	30247	241587	0.55%	ENCSR000ERN
NRF1	Lymphoblast	17314	142918	0.32%	ENCSR980YXJ
TBP	Lymphoblast	23408	183549	0.41%	ENCSR000ERP
TCF12	Lymphoblast	36815	294533	0.67%	ENCSR906QEK
USF1	Lymphoblast	8158	69811	0.16%	ENCSR973SOG
USF2	Lymphoblast	5322	45624	0.10%	ENCSR000ERJ
SPI1	Macrophage	65942	525645	1.19%	GSE48759
FLI1	Megakaryocyte	3224	27394	0.06%	GSE36029
GATA1	Megakaryocyte	4154	35204	0.08%	GSE36029
RUNX1	Megakaryocyte	10700	90597	0.20%	GSE45372

Table S6: Summary statistics for all additional mouse datasets. The mouse genome filtered window set consisted of 41883806 windows in total.

TF	Cell Type	Peaks	Bound Windows	Frac. Bound	Accession ID
CEBPA	Adipocyte	53157	396024	0.77%	GSE27450
CTCF	Adipocyte	48914	376510	0.73%	GSE20752
PPARG	Adipocyte	58757	462122	0.90%	GSE20752
BHLHE40	Erythroid	27808	217471	0.42%	ENCSR000EGV
CTCF	Erythroid	59803	476076	0.92%	ENCSR000DMA
E2F4	Erythroid	9109	68965	0.13%	ENCSR000EWL
ELF1	Erythroid	32683	258940	0.50%	ENCSR000BMD
ETS1	Erythroid	13775	101997	0.20%	ENCSR000BKQ
GATA1	Erythroid	14676	113735	0.22%	ENCSR000EFT
JUND	Erythroid	47180	367973	0.71%	ENCSR000EGN
MAFK	Erythroid	27213	213251	0.41%	ENCSR000EGX
MAX	Erythroid	37342	286474	0.56%	ENCSR000EFV
MAZ	Erythroid	40398	308748	0.60%	ENCSR000EFX
MEF2A	Erythroid	6407	49536	0.10%	ENCSR000BNV
MXI1	Erythroid	9081	70132	0.14%	ENCSR000EGZ
MYC	Erythroid	31378	233216	0.45%	ENCSR000EGJ
NRF1	Erythroid	4436	36511	0.07%	ENCSR000EHH
TAL1	Erythroid	29476	229424	0.45%	ENCSR000EHB
UBTF	Erythroid	19228	139064	0.27%	ENCSR000EFZ
USF1	Erythroid	22382	177524	0.34%	ENCSR000BKT
USF2	Erythroid	3621	29702	0.06%	ENCSR000EHG
CTCF	Erythroid progenitor	36729	292844	0.57%	GSE26501
GATA1	Erythroid progenitor	25710	198358	0.38%	GSE26501
TAL1	Erythroid progenitor	38152	285562	0.55%	GSE26501
CTCF	ESC	57384	466110	0.90%	ENCSR000BNH
MAFK	ESC	13422	109310	0.21%	ENCSR000EBS
NANOG	ESC	8905	72332	0.14%	ENCSR000BMT
POU5F1	ESC	5029	41330	0.08%	ENCSR000BMU
FLI1	Hematopoietic progenitor	38760	310707	0.60%	GSE45144
LMO2	Hematopoietic progenitor	2037	16312	0.03%	GSE45144
RUNX1	Hematopoietic progenitor	29950	241749	0.47%	GSE45144
SPI1	Hematopoietic progenitor	167273	1283083	2.49%	GSE70660
BHLHE40	Lymphoblast	28651	227674	0.44%	ENCSR000DZJ
CTCF	Lymphoblast	41765	339466	0.66%	ENCSR000DZN
E2F4	Lymphoblast	4375	35071	0.07%	ENCSR000DYY
ELF1	Lymphoblast	27369	212273	0.41%	ENCSR000BMB
ETS1	Lymphoblast	12912	103978	0.20%	ENCSR000BKA
IRF4	Lymphoblast	23043	182227	0.35%	ENCSR000BGY
JUND	Lymphoblast	7602	61307	0.12%	ENCSR000DYS
MAX	Lymphoblast	13605	104721	0.20%	ENCSR000DZF
MAZ	Lymphoblast	23166	175906	0.34%	ENCSR000DZA
MEF2A	Lymphoblast	22588	180702	0.35%	ENCSR000BKB
MXI1	Lymphoblast	21737	164076	0.32%	ENCSR000DZI
MYC	Lymphoblast	4950	37375	0.07%	ENCSR000DKU
NRF1	Lymphoblast	3363	27933	0.05%	ENCSR000DZO
TBP	Lymphoblast	19535	147978	0.29%	ENCSR000DZZ
TCF12	Lymphoblast	25023	201436	0.39%	ENCSR000BGZ
USF1	Lymphoblast	8461	69700	0.14%	ENCSR000BGI
USF2	Lymphoblast	4450	36621	0.07%	ENCSR000DZU
SPI1	Macrophage	88793	693731	1.35%	GSE31621
FLI1	Megakaryocyte	4649	38182	0.07%	GSE24674
GATA1	Megakaryocyte	4147	33052	0.06%	GSE24674
RUNX1	Megakaryocyte	58757	209261	0.41%	GSE24674

Table S7: Summary statistics for all additional human datasets. The human genome filtered window set consisted of 51548966 windows in total.