

# Learning from mistakes: Accurate prediction of cell type-specific transcription factor binding

Jens Keilwagen<sup>1</sup>, Stefan Posch<sup>2</sup>, and Jan Grau<sup>2</sup>

<sup>1</sup>Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, D-06484, Germany

<sup>2</sup>Institute of Computer Science, Martin Luther University Halle–Wittenberg, Halle (Saale), D-06120, Germany.

Computational prediction of cell type-specific, *in-vivo* transcription factor binding sites is still one of the central challenges in regulatory genomics, and a variety of approaches has been proposed for this purpose.

Here, we present our approach that earned a shared first rank in the “ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge” in 2017. This approach employs an extensive set of features derived from chromatin accessibility, binding motifs, gene expression, sequence and annotation to train classifiers using a supervised, discriminative learning principle. Two further key aspects of this approach is learning classifier parameters in an iterative training procedure that successively adds additional negative examples to the training set, and creating an ensemble prediction by averaging over classifiers obtained for different training cell types.

In post-challenge analyses, we benchmark the influence of different feature sets and find that chromatin accessibility and binding motifs are sufficient to yield state-of-the-art performance for *in-vivo* binding site predictions. We also show that the iterative training procedure and the ensemble prediction are pivotal for the final prediction performance.

To make predictions of this approach readily accessible, we predict 682 peak lists for a total of 31 transcription factors in 22 primary cell types and tissues, which are available for download at <https://www.synapse.org/#!Synapse:syn11526239>, and we demonstrate that these predictions may help to yield biological conclusions.

**Contact:** [grau@informatik.uni-halle.de](mailto:grau@informatik.uni-halle.de)

## 35 1 Introduction

36 Activation or repression of transcription is one of the fundamental levels of gene regu-  
37 lation. Transcriptional regulation depends on transcription factors (TFs), which specif-  
38 ically bind to sites in promoters or enhancers of regulated genes or bind indirectly via  
39 other, sequence specific TFs. Modeling binding specificities, typically represented as se-  
40 quence motifs, has been an important topic of bioinformatics since its infancy (Staden,  
41 1984; Berg and von Hippel, 1987). However, it soon became evident that *in-silico* binding  
42 site predictions based on sequence motifs alone are insufficient to explain *in-vivo* bind-  
43 ing of TFs and that determinants beyond sequence specificity likely play an important  
44 role (Stormo and Fields, 1998; Bulyk, 2003).

45 The emergence of high-throughput techniques like ChIP-chip (Wu *et al.*, 2006) or  
46 ChIP-seq (Johnson *et al.*, 2007) allowed for experimentally determining *in-vivo* TF bind-  
47 ing regions on a genome-wide scale. While especially ChIP-seq and derived techniques  
48 have the potential to measure TF and cell type-specific binding, the experimental effort  
49 and costs currently preclude ChIP-seq experiments for hundreds to thousands of TFs  
50 in a variety of different cell types and tissues. Hence, there is a demand for computa-  
51 tional methods predicting cell type-specific TF binding with high accuracy. Fortunately,  
52 the existence of genome-wide ChIP data for a subset of TFs and cell types also opens  
53 up the opportunity to generate more accurate models by supervised machine learning  
54 techniques, which may consider further features besides motif matches.

55 High-throughput sequencing may also be used to obtain genome-wide assays of chro-  
56 matin accessibility (e.g., DNase-seq (Hesselberth *et al.*, 2009), ATAC-seq (Buenrostro  
57 *et al.*, 2013)), which has been considered one of the key features distinguishing func-  
58 tional from non-functional TF binding sites (Galas and Schmitz, 1978; Chen *et al.*,  
59 2010). Chromatin accessibility data may yield genome-wide maps of functional binding  
60 sites of a large fraction of TFs but, in contrast to ChIP-seq, does not identify the TF  
61 binding to a specific region. Hence, the association between bound regions (“footprints”)  
62 and TFs is typically derived computationally (Pique-Regi *et al.*, 2011).

63 Following this path, a plenitude of tools (Supplementary Table S1) has been proposed  
64 over the last five years. While the general notion of combining sequence signals with  
65 chromatin accessibility data and, in some cases, other features is common to the majority  
66 of approaches, they differ in several specific aspects. Most approaches (e.g., Pique-Regi  
67 *et al.* (2011); Natarajan *et al.* (2012); Piper *et al.* (2013); Gusmao *et al.* (2014); Chen *et al.*  
68 (2017)) use binding motifs represented as position weight matrix (PWM) models that  
69 have been obtained from databases like TRANSFAC (Matys *et al.*, 2006), Jaspar (Math-  
70 elier *et al.*, 2016), UniProbe (Newburger and Bulyk, 2009) or CisBP (Weirauch *et al.*,  
71 2014), or from motif collections like Factorbook (Wang *et al.*, 2012), the ENCODE-  
72 motif collection (Kheradpour and Kellis, 2014), or Homer (Heinz *et al.*, 2010), while  
73 some perform de-novo motif discovery based on k-mers (Arvey *et al.*, 2012) or as part of  
74 convolutional neural networks (Quang and Xie, 2017; Qin and Feng, 2017). Irrespective  
75 of the source of the motifs considered, three general schemas are have been established for  
76 combining motif predictions with chromatin accessibility data. First, motif matches (i.e.,  
77 predicted binding sites) may be used as prior information and combined with DNase-seq

78 data to distinguish functional from non-functional binding sites (e.g., Pique-Regi *et al.*  
79 (2011); Jankowski *et al.* (2016); Raj *et al.* (2015)), Second, TF footprints may be first  
80 identified from DNase-seq data and then annotated with specific TFs based on motif  
81 matches afterwards (Gusmao *et al.*, 2014). Third, both sources of information are com-  
82 bined in a holistic approach (Quang and Xie, 2017; Qin and Feng, 2017). DNase-seq  
83 (and ATAC-seq) data are employed in different ways by existing approaches including  
84 i) binning of chromatin accessibility statistics in larger genomic regions around putative  
85 binding sites (Luo and Hartemink, 2012), ii) association of chromatin accessibility with  
86 specific genes (Schmidt *et al.*, 2017), or iii) high-resolution maps of DNase cut sites (Sher-  
87 wood *et al.*, 2014; Raj *et al.*, 2015), which may additionally be considered separately for  
88 each DNA strand (Piper *et al.*, 2013). On the methodological level, approaches either fol-  
89 low a supervised approach based on training examples labeled as “bound” or “unbound”,  
90 typically derived from TF ChIP-seq data (e.g., Arvey *et al.* (2012); Luo and Hartemink  
91 (2012); Kähärä and Lähdesmäki (2015); Liu *et al.* (2017)), or an unsupervised approach  
92 clustering regions into “bound” and “unbound” based on their experimental properties  
93 (e.g., DNase-seq data or histone modifications (Pique-Regi *et al.*, 2011; Sherwood *et al.*,  
94 2014; Gusmao *et al.*, 2014)), while others base their predictions on statistical tests (Piper  
95 *et al.*, 2013) or scores related to binding affinity predictions (Schmidt *et al.*, 2017). Su-  
96 pervised approaches use a variety of methods like support vector machines (Arvey *et al.*,  
97 2012; Kumar and Bucher, 2016), (sparse) logistic regression (Natarajan *et al.*, 2012;  
98 Luo and Hartemink, 2012; Kähärä and Lähdesmäki, 2015; Chen *et al.*, 2017), random  
99 forests (Liu *et al.*, 2017), or neural networks adapted by deep learning (Quang and Xie,  
100 2017; Qin and Feng, 2017). Unsupervised approaches use hierarchical mixture mod-  
101 els (Pique-Regi *et al.*, 2011), hierarchical multi-scale models (Raj *et al.*, 2015), hidden  
102 Markov models (Gusmao *et al.*, 2014), or other probabilistic models (Sherwood *et al.*,  
103 2014). In some approaches, sequence-based features besides motif matches (Kumar and  
104 Bucher, 2016; Gusmao *et al.*, 2014; Chen *et al.*, 2017), sequence conservation (Kumar  
105 and Bucher, 2016; Liu *et al.*, 2017; Chen *et al.*, 2017), or additional experimental data  
106 like histone modification (Pique-Regi *et al.*, 2011; Arvey *et al.*, 2012; Gusmao *et al.*,  
107 2014) are included into the model. Finally, a subset of approaches uses the prediction  
108 of TF binding regions as an intermediate step for predicting gene regulation (Natarajan  
109 *et al.*, 2012) or tissue-specific gene expression (Schmidt *et al.*, 2017).

110 Each of these previous approaches has its benefits and downsides, and the results of  
111 benchmark studies in the respective original publications are ambiguous with regard to  
112 their prediction performance. Against this background, the “ENCODE-DREAM in vivo  
113 Transcription Factor Binding Site Prediction Challenge” (ENCODE-DREAM Consor-  
114 tium, 2017) aimed at assessing the performance of tools for predicting cell type-specific  
115 TF binding in human using a minimal set of experimental data in a fair and unbiased  
116 manner. The challenge setting has several important advantages over typical benchmark  
117 studies. First, approaches are typically applied to the challenge data by their authors,  
118 which reduces the risk of, for instance, sub-optimal parameter settings or mode choices.  
119 Second, the ground truth data used for evaluation are known only to the challenge or-  
120 ganizers, which ensures a fair and unbiased comparison. Third, at least in DREAM  
121 challenges, participants are required to document their method such that the submit-

122 ted predictions can be reproduced by the challenge organizers. In addition, preliminary  
123 assessments on dedicated leaderboard data may help to judge ranking relative to com-  
124 petitors and also limited tuning of methods in a realistic setting.

125 Participants in the ENCODE-DREAM challenge were allowed to use binding motifs  
126 from any source, genomic sequence, gene annotations, *in-silico* DNA shape predictions,  
127 and cell type-specific DNase-seq and RNA-seq data. In addition, TF ChIP-seq data and  
128 ChIP-seq-derived labels (“bound”, “unbound”, “ambiguous”) were provided for training  
129 cell types and chromosomes. Predictions had then to be made for combinations of TF  
130 and cell type not present in the training data on held-out chromosomes. Predictions  
131 were evaluated against labels derived from TF ChIP-seq data for that specific TF and  
132 test cell type.

133 Here, we present our approach for predicting cell type-specific TF binding regions  
134 earning a shared first rank among 40 international teams, including some of the devel-  
135 opers of those approaches mentioned above (<https://www.synapse.org/#!Synapse:syn6131484/wiki/405275>, (ENCODE-DREAM Consortium, 2017)). Following the cat-  
136 egorization applied to previous approaches above, the approach presented in this paper  
137 combines several novel ideas. First, we consider motifs from databases, but also motifs  
138 learned by de-novo motif discovery from ChIP-seq and DNase-seq data using sparse local  
139 inhomogeneous mixture (Slim) models (Keilwagen and Grau, 2015), which may capture  
140 short to mid-range intra-motif dependencies. Second, we process DNase-seq data fol-  
141 lowing the binning idea of previous approaches but defining novel statistics computed  
142 from the data in those bins, and derive several sequence-based, annotation-based, and  
143 RNA-seq-based features. Third, we apply a supervised machine learning approach that  
144 employs a discriminative learning principle, which is related to logistic regression but  
145 allows for explicit model assumptions with regard to different features. Fourth, dis-  
146 criminative learning is combined with an iterative training approach for refining sets  
147 of representative negative examples. Finally, we combine the predictions of classifiers  
148 trained in different of these iterations and on different training cell types in an ensemble-  
149 like approach.

151 As this novel approach has already been benchmarked against a large number of  
152 competing approaches as part of the ENCODE-DREAM challenge (ENCODE-DREAM  
153 Consortium, 2017), we focus on the analysis for the contributions of different aspects of  
154 this approach on the final prediction performance in this paper. Specifically, we evaluate  
155 the contribution of related subsets of features, we compare the performance achieved by  
156 training on an initial negative set with that achieved by the iterative training proce-  
157 dure complementing this initial set with further negative examples, and we assess the  
158 performance of individual classifiers compared with their ensemble prediction. Based  
159 on these analyses, we define and benchmark a simplified variant of the proposed ap-  
160 proach. Finally, we provide a large collection of predicted, cell type-specific tracks of  
161 binding regions for 31 TFs in 22 primary cell types and tissues to make predictions by  
162 this approach readily accessible.

## 163 2 Methods

### 164 2.1 Data

165 We use the following types of input data sets as provided by the challenge organizers  
166 (<https://www.synapse.org/#!/Synapse:syn6131484/wiki/402033>):

- 167 • the raw sequence of the human genome (hg19) and gene annotations according  
168 to the gencode v19 annotation (<http://www.gencodegenes.org/releases/19.html>) (Harrow *et al.*, 2012),  
169
- 170 • cell type-specific DNase-seq “fold-enrichment coverage” tracks, which represent  
171 DNase-seq signal relative to a pseudo control smoothed in a 150 bp window,
- 172 • cell type-specific DNase-seq peak files in “conservative” (IDR threshold of 10% in  
173 pseudo replicates) and “relaxed” (no IDR threshold) flavors,
- 174 • cell type-specific TPM values from RNA-seq experiments in two bio-replicates for  
175 all gencode v19 genes as estimated by RSEM (Li and Dewey, 2011),
- 176 • cell type-specific and TF-specific ChIP-seq peak files in “conservative” (IDR thresh-  
177 old of 10% in pseudo replicates) and “relaxed” (no IDR threshold) flavors,
- 178 • cell type-specific and TF-specific label files classifying genome-wide 200 bp regions  
179 shifted by 50 bp into B=“bound”, A=“ambiguous”, and U=“unbound” according  
180 to the respective conservative and relaxed ChIP-seq peak files; an overview of the  
181 combinations of TF and cell type in the training data, the leaderboard data, and  
182 the test data used for evaluation in the final challenge round is given in Supple-  
183 mentary Figure S1.

184 In addition, we download sequence motifs represented as PWMs from the following  
185 collections:

- 186 • TF-specific motifs from the databases HOCOMOCO (Kulakovskiy *et al.*, 2016)  
187 and DBcorrDB (Grau *et al.*, 2015a),
- 188 • motifs related to epigenetic markers from the epigram pipeline (Whitaker *et al.*,  
189 2015).

190 Details about the motifs considered are given in section Features and Supplementary  
191 Text S1.

192 For predicting cell type-specific binding of TFs in additional cell types beyond those  
193 considered in the challenge, we downloaded DNase-seq data (FastQ format) from the  
194 ENCODE project ([encodeproject.org](http://encodeproject.org)). Specifically, we selected all DNase-seq exper-  
195 iments that i) were flagged as “released”, ii) have FastQ files available, iii) are not from  
196 immortalized cell lines, iv) have no entry in one of the “Audit error” categories, and v) are  
197 not in the “insufficient replicate concordance” category of “Audit no compliant”. A list

198 of the corresponding experiments was obtained from the ENCODE project (S3) and ex-  
199 periments were filtered for the existence of at least two replicates, yielding 23 experiments  
200 in total. One of these experiments had to be excluded later, because a different DNase  
201 protocol with much shorter reads had been used. For the remaining 22 experiments  
202 (Supplementary Table S3), all FastQ files were downloaded from ENCODE and pro-  
203 cessed using ATAC-Seq/DNase-Seq Pipeline ([https://github.com/kundajelab/atac\\_](https://github.com/kundajelab/atac_)  
204 `dnase_pipelines`, latest git commit: c1d07d38a02af2f0319a69707eee047ab6112ecc (Tue  
205 Mar 21 20:31:25 2017)). The data sets were analyzed using the following parameters  
206 `-species hg19 -type dnase-seq -subsample 50M -se`. For further analyzes, the re-  
207 laxed (`./out/peak/idr/pseudo_reps/rep1/*.filt.narrowPeak.gz`) and conservative  
208 peaks (`./out/peak/macs2/overlap/*pval0.1*.filt.narrowPeak.gz`) as well as the  
209 DNase coverage (`./out/signal/macs2/rep1/*.fc.signal.bigwig`) were used.

210 In addition, we download ChIP-seq peak files (Supplementary Table S4) matching  
211 these cell types and one of the TFs considered. Based on the “relaxed” (i.e., “optimal  
212 idr thresholded peaks”) and “conservative” (i.e., “conservative idr thresholded peaks”)  
213 peak files, we derive labels for 200 bp windows every 50 bp as proposed for the challenge.  
214 Specifically, we labels each 200 bp region overlapping a conservative peak by at least  
215 100 bp as “bound”. Of the remaining regions, all regions that overlap a relaxed peak  
216 by at least 1 bp are labeled “ambiguous”, while all other regions are labeled “unbound”.  
217 For a subset of TFs, no conservative peaks are available due to the lack of replicates. In  
218 such cases, we also use the relaxed peaks to assign “bound” labels.

## 219 2.2 Binning the genome

220 Given the large number of ChIP-seq data sets for diverse TFs in the training, leader-  
221 board, and test cell types, defining features with base-pair resolution would have been  
222 a major challenge with regard to memory requirements (hard disk as well as main  
223 memory) as well as runtime. As the final prediction is requested for overlapping 200 bp  
224 regions with an offset of 50 bp, we decide to compute features with a matching resolu-  
225 tion of 50 bp. To this end, the genome is divided into non-overlapping bins of 50 bp.  
226 Features are then either computed directly with that resolution (where possible, e.g.,  
227 distance to the closest TSS), or first computed with base-pair resolution and afterwards  
228 summarized as aggregate values (minimum, maximum, median, or similar statistics) for  
229 each 50 bp bin. By this means, e.g., a score profile of a motif model or a DNase coverage  
230 profile is represented by a few aggregate values instead of 50 individual values, which  
231 substantially reduces memory requirements. An odd number of several, adjacent bins  
232 represented by the respective feature values (see below) is then considered as input of  
233 the classifier composed of statistical models for the training process as well as for making  
234 predictions. Conceptually, the classifier uses the information from those bins to compute  
235 a-posteriori probabilities  $P_i$  that center bin  $i$  contains a peak summit. The number of  
236 adjacent bins considered is determined from the median across cell types of the median  
237 peak width of a given TF in the individual training cell types.

## 238 2.3 Features

239 The set of features considered may be roughly classified by the source of information  
240 (raw sequence, motif profiles, DNase-seq data, RNA-seq data). Here, we give a brief  
241 overview of these features, while we provide a complete list of definitions of all features  
242 in Supplementary Text S1.

243 The set of sequence-based features comprises the raw sequence (i.e., in 1 bp resolution)  
244 around the center bin and several measures computed from this sequence, for instance  
245 G/C-content, the frequency of CG di-nucleotides, or the length of homo-polymer tracts.  
246 Based on the gencode v19 genome annotation, we additionally define features based on  
247 overlapping annotation elements like CDS, UTRs, or TSS annotations and based on  
248 the distance to the closest TSS annotation in either strand orientation. All of these  
249 features are neither cell type-specific nor TF-specific. However, they may represent gen-  
250 eral features of genomic regions bound by TFs (like CpG islands, GC-rich promoters,  
251 or preference for non-coding regions), which might be helpful to rule out false posi-  
252 tive predictions based on TF-specific features like motif scores. In addition, the model  
253 parameters referring to those features may be adapted in a TF-specific and cell type-  
254 specific manner, which may yield auxiliary information for cell type-specific prediction  
255 of TF binding as well.

256 The most informative features with regard to the challenge task, however, are likely  
257 motif-based and chromatin accessibility-based features. For obtaining a broad set of  
258 binding motifs for each TF at hand, we combine motifs from databases with motifs ob-  
259 tained by de-novo motif discovery from the challenge data. We retrieve PWM models  
260 of the TF at hand from the HOCOMOCO database (Kulakovskiy *et al.*, 2016) and our  
261 in-house database DBcorrDB (Grau *et al.*, 2015a). We perform de-novo motif discov-  
262 ery by the in-house approach Dimont (Grau *et al.*, 2013) learning PWM and LSlim(3)  
263 models (Keilwagen and Grau, 2015) on the “conservative” and “relaxed” ChIP-seq peak  
264 files, and also based on the peak files obtained from DNase-seq experiments. In addition,  
265 we obtain motifs from the epigram pipeline (Whitaker *et al.*, 2015), which are related to  
266 DNA methylation and histone marks of active promoters and enhancers. For a specific  
267 combination of cell type and TF, we also consider motifs of a set of “peer” motifs, which  
268 are determined from the literature (Factorbook, Wang *et al.* (2012)) and by comparing  
269 the overlaps between the respective peak lists.

270 All of these motifs are then used in a sliding window approach to obtain base-pair  
271 resolution score profiles, which are then summarized by aggregate statistics representing  
272 the binding affinity to the strongest binding site (i.e., the maximum log-probability in a  
273 bin according to the motif model) as well as general affinity to broader regions (i.e, the  
274 logarithm of the average probability in a bin). The set of motifs may comprise models  
275 of general binding affinity of the TF at hand but may also capture cell type-specific  
276 differences in the binding regions, which could be caused by interaction with other TFs  
277 including competition for similar binding sites.

278 DNase-seq-based features are computed from the “fold-enrichment coverage” tracks  
279 and DNase-seq peak files provided with the challenge data. These features quantify  
280 short and long range chromatin accessibility, stability of the DNase signal in the region

281 of interest and across different cell types, and overlaps with DNase-seq peak regions.

282 Finally, RNA-seq data are represented by the TPM value of the gene closest to the  
283 bin of interest as well as measures of stability within biological replicates and across  
284 different cell types.

285 DNase-seq and RNA-seq-based features are cell type-specific but not TF-specific by  
286 design. However, model parameters may adapt to situations where one TF preferentially  
287 binds to open chromatin, whereas another TF may also bind in nucleosomal regions.

288 Feature values are computed using a combination of Perl scripts and Java classes  
289 implemented using the Java library Jstacs (Grau *et al.*, 2012). Genome wide feature  
290 values with bin-level resolution are pre-computed and stored in a sparse, compressed  
291 text format.

## 292 **2.4 Model & basic learning principle**

293 We model the joint distribution of these different features by a simple product of indepen-  
294 dent densities or discrete distributions (Supplementary Text S2). Specifically, we model  
295 numeric features (e.g., DNase-based statistics, motif scores, RNA-seq-based features) by  
296 Gaussian densities, discrete, annotation-based features by independent binomial distri-  
297 butions for each type and strand of annotation, and raw sequence by a homogeneous  
298 Markov model of order 3. All distributions are in the exponential family and parameter-  
299 ized using their natural parameterization (Bishop, 2006; Keilwagen *et al.*, 2010), which  
300 allows for unconstrained numerical optimization.

301 As learning principle, we use a weighted variant (Grau, 2010) of the discriminative  
302 maximum conditional likelihood principle (Roos *et al.* (2005), Supplementary Text S2),  
303 which is closely related to logistic regression but allows for making explicit assumptions  
304 about the distribution of the underlying data.

## 305 **2.5 Prediction schema**

306 In the challenge, final predictions are requested for 200 bp windows shifted by 50 bp  
307 along the genome, while the proposed classifier predicts a-posteriori probabilities that  
308 the current center bin contains a peak summit. To yield the predictions requested, we  
309 use these original prediction values to compute the probability that the 200 bp window  
310 overlaps at least one predicted peak by at least 100 bp (Figure 1). Assume that we  
311 already computed the a-posterior probabilities  $P_i$  that bin  $i$  contains the summit of a  
312 ChIP-seq peak according to the trained model. Further assume that for the current TF,  
313 a peak typically spans two bins before and two bins after the center bin, yielding 5 bins  
314 in total. Putative peaks overlapping the current 200 bp window starting at bin  $i$  are  
315 those centered at bin  $i - 1$  to  $i + 4$ . Hence, the probability  $S_i$  that this window overlaps a  
316 peak may be computed as the complementary probability of the event that this window  
317 overlaps no predicted peaks, which in turn is just the product of the complementary  
318 a-posteriori probabilities  $P_\ell$  of these bins.



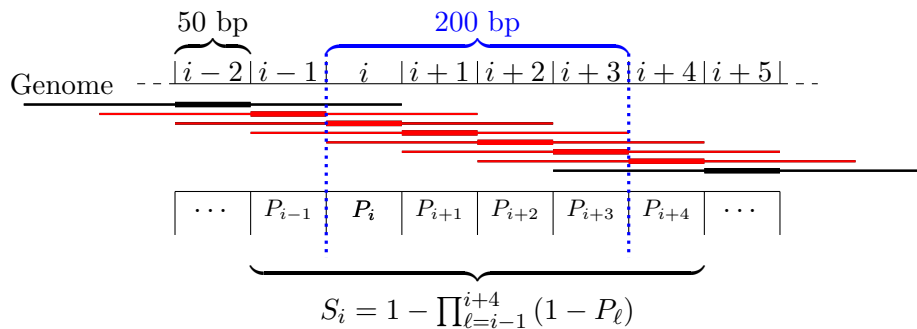


Figure 1: Schema for computing probabilities for 200 bp regions overlapping with predicted peaks spanning five bins in this example. The center bin is indicated by a thick line. Putative peaks are annotated with the probability  $P_i$  of being a true peak. All peaks marked in red overlap the region of interest (dotted blue lines) by at least 100 bp and are considered for the prediction. The prediction  $S_i$  for the 200 bp region is then computed as the probability that this region overlaps with at least one of the peaks.

## 319 2.6 Initial training data

320 For training the model parameters by the discriminative maximum condition likelihood  
 321 principle, we need labeled input data comprising a set of positive (bound) regions and a  
 322 set of negative (unbound) regions. In general, a training region is represented by a vector  
 323 of all feature values described in section Features in an odd number of consecutive bins  
 324 (see section Binning the genome). In case of positive regions, these are centered at the  
 325 bin containing the peak summit. We include all such regions around the peak summits  
 326 of the “conservative peaks” for the current TF and cell type as positive regions.

327 Since we face a highly imbalanced classification problem with rather few ChIP-seq  
 328 peaks compared with the large number of bins not covered by a peak, and since the  
 329 inclusion of all such negative regions into the training set would lead to an inaccept-  
 330 able runtime, we decided to derive representative negative regions by different sampling  
 331 strategies.

332 All sampling steps are performed stratified by chromosome. First, we sample on each  
 333 training chromosome 10 times as many negative regions (spanning an odd number of  
 334 consecutive bins) as we find positive regions on that chromosome, where center positions  
 335 are sampled uniformly over all bins not covered by a “relaxed” peak for the same cell  
 336 type and TF.

337 Second, we over-sample negative regions to yield a representative set of negative re-  
 338 gions with large DNase-seq median values similar to those of positive examples. This is  
 339 especially important as these will be regions that are hard to classify using DNase-seq  
 340 based features but are only lowly represented by the uniform sampling schema. The  
 341 over-sampling is adjusted for by down-weighting the drawn negative examples to the  
 342 corresponding frequency among all negative regions (see Supplementary Text S3).

343 Third, we sample four times as many negative regions as we have positives from regions  
344 that are ChIP-seq positive for one of the other cell types (if more than one training cell  
345 type exists for that TF), but do not overlap a “relaxed peak” in the current cell type.  
346 The latter negative regions are weighted such that the sum of their weights matches  
347 the rate of such regions among all putative negative regions. This sampling schema is  
348 intended to foster learning cell type-specific properties as opposed to general properties  
349 of the binding regions of the current TF.

350 Together, these three sampling schemas yield an initial set of negative regions, which  
351 serve as input of the discriminative maximum conditional likelihood principle in addition  
352 to the positive regions. However, in preliminary tests during the leaderboard round of the  
353 challenge, we observed that even this non-trivial sampling schema is not fully satisfactory.  
354 As testing (a large number of) other sampling schemas seemed futile, we designed an  
355 iterative training schema (Figure 2) that is loosely related to boosting (Freund and  
356 Schapire, 1996) and successively complements the initial set of negative training regions  
357 with further, informative examples.

## 358 2.7 Iterative training

359 The iterative training procedure is illustrated in Figure 2. Initially, we train a classifier on  
360 the negative regions obtained from the sampling schema explained above and all positive  
361 regions using the weighted variant of the maximum conditional likelihood principle. We  
362 then use this classifier to obtain a-posteriori probabilities  $P_i$  for bin  $i$  on the training  
363 chromosomes. To limit the runtime required for this prediction step, we restrict the  
364 prediction to chromosomes chr10 to chr14. These probabilities are then used as input  
365 of the prediction schema (section Prediction schema) to yield predictions for the 200 bp  
366 regions labeled by the challenge organizers based on the ChIP-seq training peaks. Hence,  
367 we may distinguish prediction values of positive regions (label B=“bound”) and negative  
368 regions (label U=“unbound”), while regions labeled as A=“ambiguous” are ignored.  
369 To select additional negative regions that are likely false positive predictions, we first  
370 collect the prediction scores of all positive regions (labeled as B) and determine the  
371 corresponding 1% percentile. We then select from the negative regions (labeled as U)  
372 all those with a predictions score larger than this 1% percentile, which are subsequently  
373 added to the set of negative regions with a weight of 1 per region selected.

374 In the next iteration, we train a second classifier, again using all positive regions but  
375 the initial negative regions complemented with the additional negative regions identified  
376 in the previous step. Prediction is then performed using both classifiers, where the pre-  
377 dictions of the individual two classifiers (or all previously trained classifiers in subsequent  
378 iterations) are averaged per region. Again, regions labeled U with large prediction scores  
379 are identified and added to the set of negative regions, which then serve as input of the  
380 following iteration. After five rounds of training yielding five classifiers, the iterative  
381 training procedure is terminated.

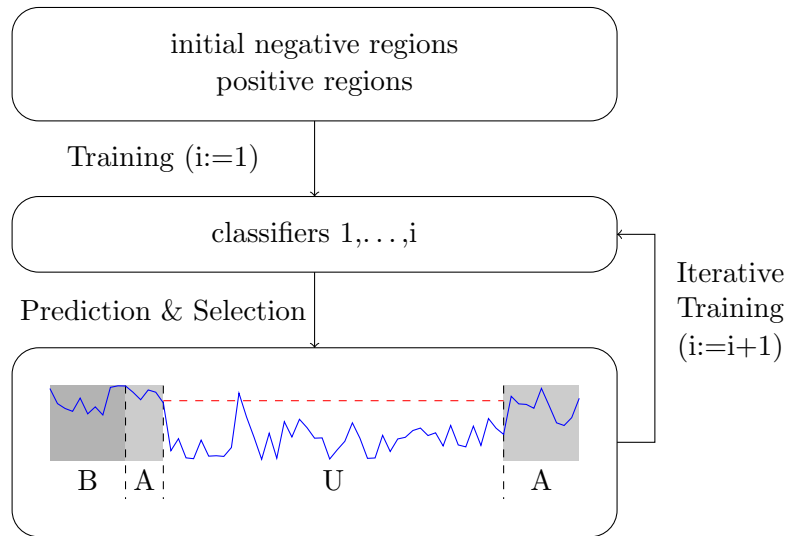


Figure 2: Iterative training procedure. Starting from an initial set of negative regions and the complete set of positive regions, a first classifier is trained, applied to the training data, and putative false positive (i.e., “unbound” regions with large prediction scores) are identified. In each of the subsequent iterations, such regions are added to the set of negative regions, which are in turn used for training refined classifiers. The result of this iterative training procedure is a set of  $K = 5$  classifiers trained in 5 cycles of the iterative training procedure.

## 382 2.8 Final prediction

383 The iterative training procedure is executed for all  $K$  training cell types with ChIP-seq  
384 data for the TF of interest, which yields a total of  $5 \cdot K$  classifiers. For the final prediction,  
385 the prediction schema (section Prediction schema) is applied to all chromosomes and  
386 each classifier. These predictions are finally averaged per 200 bp region to yield the final  
387 prediction result.

## 388 2.9 Deriving peak lists

389 For the additional primary cell types and tissues beyond those considered in the chal-  
390 lenge, we further process final predictions to yield peak lists in narrowPeak format,  
391 which are smaller and easier to handle than the genome-wide probability tracks with  
392 50 bp resolution. To this end, we join contiguous stretches of regions with predicted  
393 binding probability above a pre-defined threshold  $t$  into a common peak region. For  
394 each region, we record the maximum probability  $p$ , and discard bordering regions with a  
395 probability below  $0.8 \cdot p$ . The resulting regions are then annotated according to the nar-  
396 rowPeak format with a “peak summit” at the center of the region yielding  $p$ , a “score”  
397 of  $-100 \cdot \log_{10}(1 - p)$ , and a “signal value” equal to  $p$ . We generate “relaxed” peak  
398 predictions using  $t = 0.6$  and “conservative” peak prediction using  $t = 0.8$ .

## 399 2.10 Availability

400 The approach presented here has been implemented using the Java library Jstacs (Grau  
401 *et al.*, 2012) combined with custom Perl and bash scripts for data extraction, conversion,  
402 and pipelining. ENCODE-DREAM-specific Java classes will be part of the next Jstacs  
403 release. The complete code accompanying the challenge submission is, in accordance  
404 with the challenge guidelines, available from <https://www.synapse.org/#!/Synapse:syn8009967/wiki/412123> including a brief method writeup.  
405

## 406 3 Results

407 During the ENCODE-DREAM challenge, a large number of approaches created by 40  
408 international teams has been benchmarked on 13 cell type-specific ChIP-seq assays for  
409 12 different TFs in human (Supplementary Figure S1). A set of 109 data sets for the  
410 same (and additional) TFs in other cell types was provided for training. In addition,  
411 teams could submit predictions for 27 further combinations of TF and cell type in a  
412 leaderboard round and evaluation results for submitted predictions were made available  
413 to all participants. Training data comprised cell type-specific DNase-seq data, cell type-  
414 specific RNA-seq data, genomic sequence and annotations, and *in-silico* DNA shape  
415 predictions. In addition, cell type-specific and TF-specific ChIP-seq data and derived  
416 labels were provided for training chromosomes, while predictions were evaluated only  
417 on the remaining, held-out chromosomes chr1, chr8, and chr21 that were not provided

418 with any of the ChIP-seq training data. For 200 bp regions shifted by 50 bp, genome-  
419 wide predictions of the probability that a specific region overlaps a ChIP-seq peak were  
420 requested from the participating teams. Predictions were evaluated by i) the area under  
421 the ROC curve (AUC-ROC), ii) the area under the precision-recall curve (AUC-PR), iii)  
422 recall at 10% FDR, and iv) recall at 50% FDR on each of the 13 test data sets. These  
423 were aggregated per data set based on the average, normalized rank earned for each of  
424 these measures in 10 bootstrap samples of the held-out chromosomes, and a final ranking  
425 was obtained as the average of these rank statistics (ENCODE-DREAM Consortium,  
426 2017).

427 As a result of this ranking, the approach presented in this paper (team “J-Team”)  
428 earned a shared first rank together with the approach created by team “Yuanfang  
429 Guan” (<https://www.synapse.org/#!Synapse:syn6131484/wiki/405275>, ENCODE-  
430 DREAM Consortium (2017)).

431 In the following, we investigate the influence of different aspects of the proposed  
432 approach on the final prediction performance. First, we inspect the impact of different  
433 sets of related features (DNase-seq data, motif scores, RNA-seq data, sequence-based and  
434 annotation-based features) on prediction performance. Second, we study the importance  
435 of the iterative training approach as opposed to a training on initial training data.  
436 Third, we compare the performance of the predictions gained by classifiers trained on  
437 training data for individual cell types with the performance of the aggregated prediction  
438 obtained by averaging over these predictions. Finally, we apply the proposed method for  
439 predicting cell type-specific TF binding for 31 TFs in 22 additional primary cell types  
440 yielding a total of 682 prediction tracks.

### 441 **3.1 Impact of feature sets on prediction performance**

442 We use the prediction performance obtained by the proposed approach using all sets  
443 of features (section Features), the iterative training procedure (section Iterative train-  
444 ing), and the aggregation over all training cell types (section Prediction schema) as a  
445 baseline for all further comparisons (Figure 3). Throughout this manuscript, we con-  
446 sider AUC-PR as the primary performance measure, since AUC-PR is more informative  
447 about classification performance for heavily imbalanced classification problems (Keilwa-  
448 gen *et al.*, 2014; Saito and Rehmsmeier, 2015), and recall at the different FDR levels is  
449 rather unstable since it corresponds to single points on the precision-recall curve. AUC-  
450 PR values are computed using the R-package PRROC (Grau *et al.*, 2015b), which has  
451 also been used in the ENCODE-DREAM challenge.

452 We find that prediction performance as measured by AUC-PR varies greatly among  
453 the different transcription factors (Figure 3) with a median AUC-PR value of 0.4098. The  
454 best prediction performance is achieved for CTCF, which has a long and information-rich  
455 binding motif, in two different cell types (IPSC and PC-3). Above-average performance  
456 is also obtained for FOXA1 and HNF4A in liver cells. For most other TFs, we find  
457 AUC-PR values around 0.4, whereas we observe a rather low prediction accuracy for  
458 NANOG and REST.

459 To analyze the contribution of selected features on the final prediction performance,

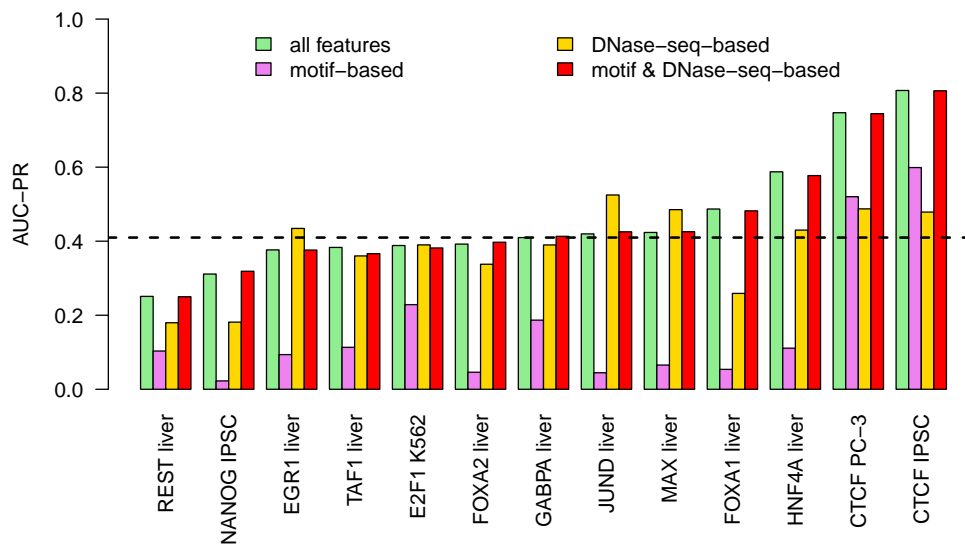


Figure 3: Across cell type performance. For each of the 13 combinations of TF and cell type within the test data, we compute the prediction performance (AUC-PR) on the held-out chromosomes of classifiers i) using all features considered, ii) using only motif-based features, iii) using only DNase-seq-based features, and iv) using only motif-based and DNase-seq-based features. Median performance of classifiers using all features is indicated by a dashed line.

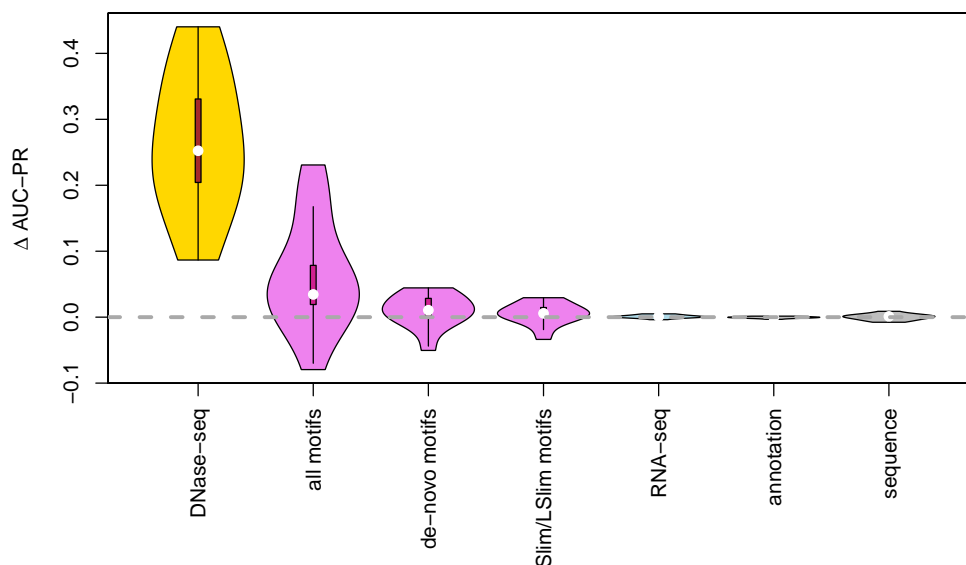


Figure 4: Importance of feature sets. We test the importance of related sets of features by excluding one set of features from the training data, measuring the performance (AUC-PR) of the resulting classifier, and subtracting this AUC-PR value from the corresponding value achieved by the classifier using all features. Hence, if  $\Delta$  AUC-PR is above zero, the left-out set of features improved the final prediction performance, whereas  $\Delta$  AUC-PR values below zero indicate a negative effect on prediction performance. We collect the  $\Delta$  AUC-PR values for all 13 test data sets and visualize these as violin plots.

460 we systematically exclude sets of related features from the input data in training and  
461 prediction. Specifically, we exclude i) all DNase-seq-based features, ii) all motif-derived  
462 features, iii) all motif-derived features of motifs obtained by de-novo motif discovery on  
463 the challenge ChIP-seq peak files, iv) all motif-derived features based on LSlim models,  
464 v) all RNA-seq-based features, vi) all annotation-based features, and vii) all sequence-  
465 based features. As a baseline, we measure AUC-PR for the classifier using all feature  
466 sets. In addition, we measure AUC-PR when excluding each individual feature set,  
467 where the difference of these two AUC-PR values quantifies the improvement gained  
468 by including the feature set. We collect these differences for all 13 test data sets and  
469 visualize them as violin plots in Figure 4.

470 We observe the greatest impact for the set of features derived from DNase-seq data.  
471 The improvement in AUC-PR gained by including DNase-seq data varies between 0.087  
472 for E2F1 and 0.440 for HNF4A with a median of 0.252.

473 Features based on motif scores (including de-novo discovered motifs and those from  
474 databases) also contribute substantially to the final prediction performance. Here, we  
475 observe large improvements for some TFs, namely 0.231 for CTCF in iPSC cells, 0.175

476 for CTCF in PC-3 cells, and 0.167 for FOXA1. By contrast, we observe a decrease in  
477 prediction performance in case of JUND ( $-0.080$ ) when including motif-based features.  
478 For the remaining TFs, we find improvements of AUC-PR between 0.008 and 0.079.  
479 We further consider two subsets of motifs, namely all motifs obtained by de-novo motif  
480 discovery on the challenge data and all Slim/LSlim models capturing intra motif depen-  
481 dencies. For motifs from de-novo motif discovery, we find an improvement for 9 of the  
482 13 data sets and for Slim/LSlim model we find an improvement for 10 of the 13 data  
483 sets. However, the absolute improvements (median of 0.011 and 0.006, respectively)  
484 are rather small, possibly because i) motifs obtained by de-novo motif discovery might  
485 be redundant to those found in databases and ii) intra motif dependencies and hetero-  
486 geneities captured by Slim/LSlim models (Keilwagen and Grau, 2015) might be partly  
487 covered by variations in the motifs from different sources.

488 Notably, RNA-seq-based features (median 0.001), annotation-based features (0.000),  
489 and sequence-based features (0.001) have almost no influence on prediction performance.

490 Having established that DNase-seq-based and motif-based features have a large impact  
491 on prediction performance, we also tested the prediction performance of the proposed  
492 approach using *only* features based on DNase-seq data and TF motifs, respectively. We  
493 find (Figure 3) that classifiers using exclusively motif-based features already yield a  
494 reasonable prediction performance for some TFs (CTCF and, to some extent, E2F1 and  
495 GABPA), whereas we observe AUC-PR values below 0.12 for the remaining of TFs. This  
496 may be explained by the large number of false positive predictions typically generated  
497 by approaches using exclusively motif information, which may only be avoided in case  
498 of long, specific motifs as it is the case for CTCF.

499 By contrast, classifiers using only DNase-seq-based features yield a remarkable perfor-  
500 mance for many of the TFs studied (Figure 3), which is lower than for the motif-based  
501 classifier only for the two CTCF datasets. For some datasets (especially JUND but  
502 also EGR1, MAX), we even observe that a classifier based on DNase-seq data alone  
503 outperforms the classifier utilizing all features. For EGR1 and MAX, we observe a  
504 drop in prediction performance when excluding only motif-based features and only a  
505 slight increase in performance when excluding one of the other non-DNase feature sets  
506 (Supplementary Table S2, Figure 4). Hence, the inclusion of non-DNase feature sets  
507 individually may not explain the apparent difference between the classifier using only  
508 DNase-seq-based features and the classifier based on all features, which suggests mutual  
509 interactions between the different feature sets.

510 In case of JUND, however, the increase in performance when neglecting all non-DNase  
511 features can likely be attributed to a strong adaptation of classifier parameters to either  
512 cell type-specific binding motifs or cell type-specific co-binding with other TFs, because  
513 JUND is the only dataset with an improved performance when excluding motif-based  
514 features as discussed above. For all three TFs, we do find an improvement of prediction  
515 performance if classifier parameters are trained on the training chromosomes of the test  
516 cell type (Supplementary Figure S2).

517 Since DNase-seq-based and motif-based features appear to be the primary feature sets  
518 affecting prediction performance, we finally study prediction performance of a classifier  
519 using only these two feature sets. We observe that prediction performance using only



520 DNase-seq-based and motif-based features is largely identical to that of the classifier  
521 using all features (Figure 3), where we observe the largest loss in AUC-PR for TAF1  
522 (0.017) and the largest gain in AUC-PR for NANOG (0.007). We notice a similar  
523 behaviour for the within-cell type case (Supplementary Figure S2). As the left-out  
524 feature sets include all RNA-seq-based features, this also has the consequence that one  
525 cell type-specific assay (namely DNase-seq) is sufficient for predicting TF binding, which  
526 broadens scope of cell types with readily available experimental data that the proposed  
527 approach may be applied to.

### 528 **3.2 Iterative training improves prediction performance**

529 As a second key aspect of the proposed approach, we investigate the impact of the  
530 iterative training procedure on the final prediction performance. To this end, we compare  
531 for each TF the AUC-PR values obtained by averaging over the predictions all five  
532 classifiers resulting from the iterative training procedure for all training cell types with  
533 the AUC-PR values obtained by only averaging over the initially trained classifiers for all  
534 training cell types, i.e., classifiers trained only on the initial training data (section Initial  
535 training data).

536 For 11 of the 13 test data sets, we observe an improvement of prediction performance  
537 by the iterative training procedure (Figure 5). The largest improvements are achieved  
538 for E2F1 (0.114), FOXA2 (0.085), NANOG (0.08), FOXA1 (0.063), and MAX (0.061).  
539 Among these are TFs for which we observed a good performance using only DNase-  
540 seq-based features (E2F1, MAX) and TFs for which the combination with motif-based  
541 features was beneficial (FOXA1, FOXA2, NANOG), which indicates that the additional  
542 negative regions added in iterations 2 to 5 do not induce a bias towards either of these  
543 two feature types. For four of these five TFs, only one (FOXA2, NANOG, FOXA1) or  
544 two (E2F1) training cell types were provided, and the variation between the different  
545 classifiers from iterative training may help to avoid overfitting. By contrast, we find a  
546 decrease in performance for JUND (0.041) and also TAF1 (0.01), which might be caused  
547 by a stronger emphasis on cell type-specific binding regions in subsequent iterations of the  
548 iterative training procedure. This hypothesis is also supported by the observation that  
549 the iterative training procedure always leads to an increase in prediction performance  
550 if classifier parameters are trained on the training chromosomes of the test cell type  
551 (Supplementary Figure S3).

### 552 **3.3 Averaging predictions improves over random selection of cell types**

553 For 9 of the 12 TFs considered, data for more than one training cell type is provided  
554 with the challenge data. Hence, one central question might be the choice of the cell  
555 type used for training and, subsequently, for making predictions for the test cell type.  
556 However, the only cell type-specific experimental data available for making that choice  
557 are DNase-seq and RNA-seq data, whereas similarity of cell types might depend on the  
558 TF considered. Indeed, similarity measures derived from DNase-seq data (e.g., Jaccard  
559 coefficients of overlapping DNase-seq peaks, correlation of profiles) or from RNA-seq

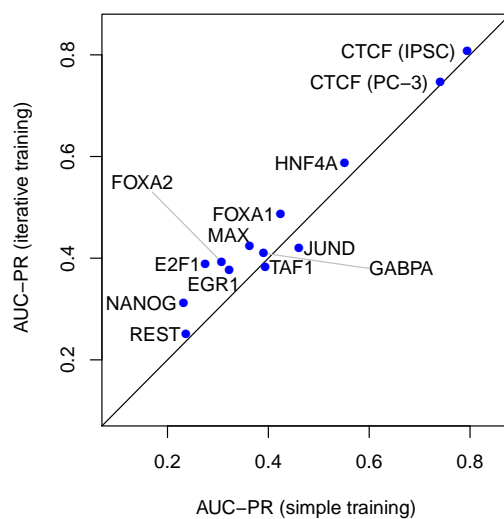


Figure 5: Relevance of the iterative training procedure. For each of the 13 test data sets, we compare the performance (AUC-PR) achieved by the (set of) classifier(s) trained on the initial negative regions (abscissa) with the performance achieved by averaging over all classifiers from the iterative training procedure (ordinate).

560 data (e.g., correlation of TPM values) showed to be non-informative with regard to the  
561 similarity of TF binding regions in preliminary studies on the training cell types.

562 Hence, we consider the choice of the training cell type a latent variable, and average  
563 over the predictions generated by the respective classifiers (see section 2.5). As labels of  
564 the test cell types have been made available after the challenge, we may now evaluate the  
565 impact of this choice on prediction performance and also test the prediction performance  
566 of classifiers trained on individual cell types (Figure 6).

567 For all test data sets with multiple training cell types available, we find that the  
568 averaged prediction yields AUC-PR values above the median of the AUC-PR values  
569 achieved for individual training cell types. This improvement is especially pronounced  
570 for REST, GABPA, and MAX. Hence, we may argue that averaging over the cell type-  
571 specific classifiers generally yields more accurate predictions than would be achieved by  
572 an uninformed choice of one specific training cell types.

573 However, we also notice for almost all test data sets with multiple training cell types  
574 (the only exception being CTCF for the PC-3 cell type) that the best prediction perfor-  
575 mance achieved using one of the individual training cell types would have gained, in some  
576 cases considerable, improvements over the proposed averaging procedure. Notably, the  
577 variance of AUC-PR between the different training cell types is especially pronounced for  
578 JUND, which supports the previous hypothesis that some features, for instance binding  
579 motifs or co-binding of TFs, are highly cell type-specific for JUND. In general, deriving  
580 informative measures of TF-specific cell type similarity based on cell type-specific assays

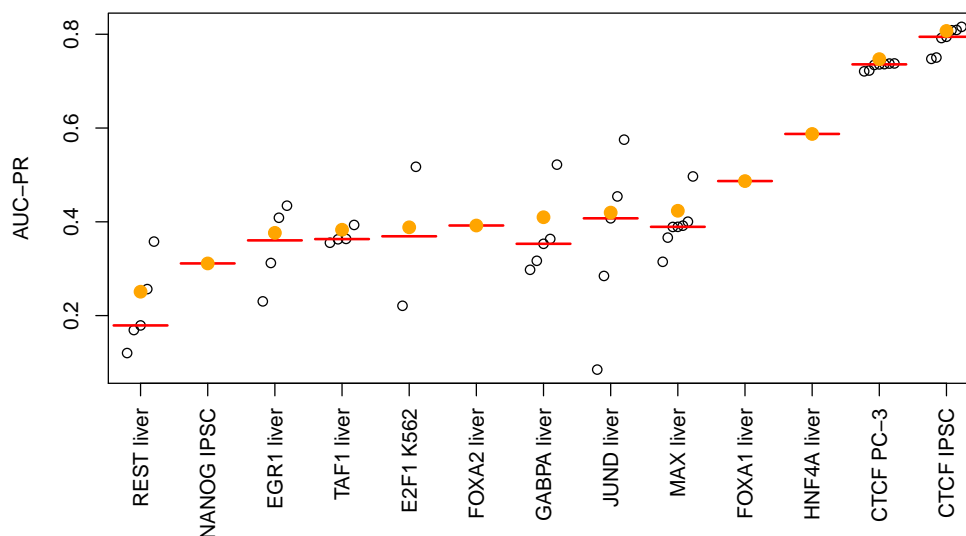


Figure 6: Performance of ensemble classifiers. For each of the 13 test data sets, we compare the performance (AUC-PR) of the individual classifiers trained on single cell types (open circles) to that of the ensemble classifier averaging over all classifiers trained on all training cell types (filled, orange circles). As a reference, we also plot the median of the individual classifiers as a red bar.

581 and, for instance, preliminary binding site predictions, would likely lead to a further  
582 boost of the performance of computational approaches for predicting cell type-specific  
583 TF binding.

### 584 3.4 Creating a collection of cell type-specific TF binding tracks

585 Having established that a single type of experimental assay, namely DNase-seq, is suffi-  
586 cient for predicting cell type-specific TF binding with state-of-the-art accuracy, we may  
587 now use the classifiers obtained on the training cell types and TFs for predictions on  
588 further cell types. To this end, we download DNase-seq data for a collection of pri-  
589 mary cell types and tissues (see section Data), process these in the same manner as the  
590 original challenge data and, subsequently, extract DNase-seq-dependent features (sec-  
591 tion Features). We then applied the trained classifiers for all 31 TFs considered in the  
592 challenge to these 22 DNase-seq feature sets to yield a total of 682 prediction tracks with  
593 a resolution of 200 bp windows shifted by 50 bp.

594 For the selected cell types (Supplementary Table S3), only few cell type and TF-  
595 specific ChIP-seq data are available (Supplementary Table S4). On the one hand, this  
596 means that the predicted TF binding tracks provide valuable, novel information for the  
597 collection of 31 TFs studied. On the other hand, this provides the opportunity to perform

598 benchmarking and sanity checks with regard to the predictions for the subset of these  
599 TFs and cell types with corresponding ChIP-seq data available. For benchmarking,  
600 we additionally obtain the “relaxed” and (where available) “conservative” peak files  
601 from ENCODE and derive the associated labels (“bound”, “unbound”, “ambiguous”)  
602 according to the procedure proposed for the ENCODE-DREAM challenge.

603 For CTCF with ChIP-seq peaks available for multiple cell types, we generally find a  
604 prediction performance that is comparable to the performance observed on the challenge  
605 data (cf. Supplementary Table S2). For these cell types, AUC-PR values (Supplemen-  
606 tary Table S5) range between 0.7720 and 0.8197 if conservative and relaxed peaks are  
607 available and the donors match between the DNase-seq and ChIP-seq experiments, while  
608 performance is slightly lower for non-matching donors (0.7322) and in case of missing  
609 conservative peaks (0.7270). For JUN, MAX, and MYC, only relaxed peaks are available  
610 from ENCODE due to missing replicates. Here, we find AUC-PR values of 0.6310 for  
611 JUN, which is substantially larger than for the challenge data, 0.4004 for MAX, which  
612 is slightly lower than for the challenge data, and 0.1989 for MYC, which has not been  
613 among the test TFs in the challenge but obtained substantially better performance in  
614 the leaderboard round.

615 The 682 genome-wide prediction tracks are still rather large (approx. 880 MB per  
616 track) and, hence, demand for substantial storage space that might not be available to  
617 the typical user, while the majority of regions are likely not bound by the TF of interest.  
618 Hence, we further condense these predictions into predicted peak lists in narrowPeak  
619 format by joining contiguous stretches with high binding probability and applying a  
620 threshold of 0.6 (relaxed) and 0.8 (conservative) on the maximum probability observed  
621 in a predicted “peak”. We provide these peak files for download at [https://www.  
622 synapse.org/#!Synapse:syn11526239](https://www.synapse.org/#!Synapse:syn11526239) (doi:10.7303/syn11526239).

623 To get an impression of the quality of the predicted peaks, we further compute  
624 Jaccard coefficients based on peak overlaps (computed using the GenomicRanges R-  
625 package (Lawrence *et al.*, 2013)) between the predicted peak files and those from the  
626 corresponding, available ChIP-seq peaks (Supplementary Tables S6 and S7), and find  
627 those to be widely concordant to the previous assessment based on the derived labels.

628 For CTCF, we may also employ Jaccard coefficients to study cell type specificity (Sup-  
629 plementary Table S6). We find that many of the cell type-specific predictions for CTCF  
630 are more similar to the ChIP-seq peaks determined for “endothelial cells of umbilical  
631 vein” than to those of their cell type of origin according to the DNase-seq data. One  
632 reason might be that only for this experiment (ENCSR000DLW), peaks have not been  
633 called using the uniform ENCODE pipeline including SPP (Kharchenko *et al.*, 2008),  
634 but by another, “unknown” software. However, if we in turn ask for each experimentally  
635 determined peak list, which of the predicted peak lists is the most similar one, this pic-  
636 ture becomes more encouraging. For 7 of the 8 cell types with matching donor between  
637 ChIP-seq and DNase-seq data, the most similar prediction is obtained for the true cell  
638 type, while in one case (“fibroblast of lung”), the most similar cell type is “foreskin  
639 fibroblast”.

640 Based on the predicted peak lists, we may also compare the predicted binding charac-  
641 teristics of the different TFs across cell types. First, we inspect the number of predicted

642 peaks per TF and cell type (Supplementary Figure S4). We find a distinct group of  
643 highly abundant TFs (CTCF, GATA3, SPI1, CEBPB, FOXA1, FOXA2, MAX), which  
644 typically also show large numbers of peaks in the training data. Among these, we  
645 find patterns of cell type specificity from the ubiquitously abundant CTCF to larg-  
646 erly varying abundance for GATA3. The remainder of TFs obtains substantially lower  
647 numbers of predicted peaks with similar patterns, e.g, for ATF7/ARID3A/NANOG or  
648 EP300/TEAD4/JUND, where the latter group has been reported to co-bind in distal  
649 enhancers (Xie *et al.*, 2013). Next, we study the stability of peak predictions, i.e., the  
650 Jaccard coefficients of peaks predicted for each of the TFs in different cell types (Supple-  
651 mentary Figure S5). Again, we find substantial variation among the TFs with GABPA,  
652 CTCF, and REST with median Jaccard coefficients above 0.7. Notably, CTCF has been  
653 one of the TFs with the largest number of predicted peaks (median 37 455), whereas  
654 we observed an order of magnitude less predicted peaks for REST (median 3 364) and  
655 GABPA (median 5 430). At the other end of the scale, we find indirectly binding TFs  
656 like EP300, or TFs that are highly specific to cell types under-represented in our data  
657 like NANOG (stem cells) and HNF4A (liver, kidney, intestines). Finally, we investigate  
658 co-binding of TFs by computing the average Jaccard coefficient across cell types for each  
659 pair of TFs (Supplementary Figure S6). Here, we observe distinct groups of co-occurring  
660 TFs like CTCF/ZNF143 or FOXA1/FOXA2, which are known to interact *in-vivo* (Bai-  
661 ley *et al.*, 2015; Ye *et al.*, 2016; Motalebipour *et al.*, 2009). In addition, we find a larger  
662 cluster of TFs with substantial overlaps between their predicted peaks comprising YY1,  
663 MAX, CREB1, MYC, E2F6, E2F1, and TAF1. As TAF1 (TATA-Box Binding Protein  
664 Associated Factor 1) is associated with transcriptional initiation at the TATA box, one  
665 explanation might be that binding sites of these TFs are enriched at core promoters.  
666 Indeed, binding to proximal promoters has been reported for MYC/MAX (Guo *et al.*,  
667 2014), CREB1 (Zhang *et al.*, 2005), YY1 (Li *et al.*, 2008), and E2F factors (Rabinovich  
668 *et al.*, 2008).

## 669 4 Discussion

670 Predicting *in-vivo* binding sites of a TF of interest *in-silico* is still one of the central  
671 challenges in regulatory genomics. A variety of tools and approaches for this purpose  
672 have been created over the last years and, among these, the approach presented here is  
673 not exceptional in many of its aspects. Specifically, it works on hand-crafted features  
674 derived from genomic and experimental data, it considers TF binding motifs and chro-  
675 matin accessibility as its major sources of information, and it uses supervised learning  
676 related to logistic regression. Here, we focus on the impact of further, novel aspects of  
677 the proposed approach on prediction performance.

678 With regard to the features considered, we find that motif-based and DNase-seq-  
679 based features are pivotal for yielding a reasonable prediction performance for most  
680 TFs, while other sequence-based, annotation-based, or RNA-seq-based features have  
681 only marginal influence on the prediction result. In case of RNA-seq-based features,  
682 however, more sophisticated features than those employed in our approach might have

683 a positive influence on prediction accuracy. In addition, DNA shape might also be  
684 informative about true TF binding sites, although *in-silico* shape predictions provided  
685 in ENCODE-DREAM are determined based on k-mers, and their influence might also be  
686 captures by higher-order Markov models or Slim/LSlim models (Keilwagen and Grau,  
687 2015) employed in the approach presented here.

688 Previous studies have shown that additional features like sequence conservation (Ku-  
689 mar and Bucher, 2016; Liu *et al.*, 2017), histone marks (Pique-Regi *et al.*, 2011; Arvey  
690 *et al.*, 2012; Gusmao *et al.*, 2014), or ChIP-seq data of co-factors (Kumar and Bucher,  
691 2016) might also help to predict *in-vivo* TF binding. However, these were not allowed  
692 to be used in the ENCODE-DREAM challenge and further experimental assays were  
693 unavailable for the training cell types. Hence, we decided to also exclude such features  
694 from the studies presented in this paper.

695 Two further novel aspects of the presented approach, namely the iterative training  
696 procedure and aggregation of predictions across training cell types, also contribute sub-  
697 stantially to the final prediction performance. Both ideas might also be of relevance in  
698 related fields. Specifically, the iterative training procedure provides a general schema  
699 applicable to imbalanced classification problems, especially when these require sampling  
700 of negative examples. In an abstract sense, the aggregation across training cell types  
701 corresponds to favoring model averaging over model selection if good selection criteria  
702 are hard to find or might yield highly varying results.

703 Despite its state-of-the-art performance proven in the ENCODE-DREAM challenges,  
704 the approach presented here has important limitations. First, the large number of mo-  
705 tifs (including those from *de-novo* motif discovery) and DNase-seq-based features lead  
706 to high demands with regard to disk space but also runtime, which are likely beyond  
707 reach for wet-lab biologists. Disk requirements could be reduced by computing features  
708 from (smaller) raw files on demand. However, this would in turn increase running time  
709 considerably.

710 Second, the approach proposed here, like any of the other supervised approaches (Nataraj-  
711 an *et al.*, 2012; Arvey *et al.*, 2012; Luo and Hartemink, 2012; Kähärä and Lähdesmäki,  
712 2015; Kumar and Bucher, 2016; Quang and Xie, 2017; Liu *et al.*, 2017; Qin and Feng,  
713 2017; Chen *et al.*, 2017), requires labeled training data for at least one cell type and the  
714 TF of interest to make predictions for this TF in another cell type.

715 While the latter limitation is partly overcome by unsupervised approaches (Pique-  
716 Regi *et al.*, 2011; Sherwood *et al.*, 2014; Gusmao *et al.*, 2014; Raj *et al.*, 2015; Jankowski  
717 *et al.*, 2016), this typically comes at the cost of reduced prediction accuracy (Kähärä  
718 and Lähdesmäki, 2015; Liu *et al.*, 2017). We address the former limitation by providing  
719 a large collection of 682 predicted peak files for 31 TFs using 22 DNase-seq data sets for  
720 primary cell types and tissues. Benchmarks based on the limited number of available  
721 ChIP-seq data indicate that prediction performance on these cell types is comparable  
722 to that achieved in the ENCODE-DREAM challenge, where absolute values of AUC-  
723 PR measuring prediction accuracy vary greatly between different TFs. For the wide  
724 majority of these combinations of TF and cell type, no experimental data about cell  
725 type-specific TF binding is available so far, which renders these predictions a valuable  
726 resource for questions related to regulatory genomics in these primary cell types and

727 tissues. Preliminary studies raise our confidence that the predicted peak files may indeed  
728 help to solve biological questions related to these cell types and TFs.

## 729 Acknowledgements

730 We would like to express our gratitude to the ENCODE-DREAM organizers, who com-  
731 posed an excellent challenge with clear rules and meaningful performance measures.  
732 We would also like to thank Ivan Kulakovskiy, Andrey Lando, and Vsevolod Makeev  
733 (team autosome.ru), Wolfgang Kopp (team BlueWhale), Daniel Quang, and Simon van  
734 Heeringen for openly sharing their ideas and thoughts during the challenge.

## 735 References

- 736 Arvey, A., Agius, P., Noble, W. S., and Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific  
737 transcription factor binding. *Genome Research*, **22**(9), 1723–1734.
- 738 Bailey, S. D., Zhang, X., Desai, K., Aid, M., Corradin, O., Cowper-Sallari, R., Akhtar-Zaidi, B., Scacheri, P. C., Haibe-  
739 Kains, B., and Lupien, M. (2015). ZNF143 provides sequence specificity to secure chromatin interactions at gene  
740 promoters. **2**, 6186 EP –.
- 741 Berg, O. G. and von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins: Statistical-mechanical  
742 theory and application to operators and promoters. *Journal of Molecular Biology*, **193**(4), 723 – 743.
- 743 Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New  
744 York, 1st edition.
- 745 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin  
746 for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat*  
747 *Meth*, **10**(12), 1213–1218.
- 748 Bulyk, M. L. (2003). Computational prediction of transcription-factor binding site locations. *Genome Biology*, **5**(1),  
749 201.
- 750 Chen, X., Hoffman, M. M., Bilmes, J. A., Hesselberth, J. R., and Noble, W. S. (2010). A dynamic Bayesian network for  
751 identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics*, **26**(12), i334–i342.
- 752 Chen, X., Yu, B., Carriero, N., Silva, C., and Bonneau, R. (2017). Mocap: large-scale inference of transcription factor  
753 binding sites from chromatin accessibility. *Nucleic Acids Research*, **45**(8), 4315–4329.
- 754 ENCODE-DREAM Consortium (2017). Systematic evaluation of multi-modal approaches to predict in vivo transcription  
755 factor binding across cell types. *bioRxiv*.
- 756 Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the 13th*  
757 *International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann.
- 758 Galas, D. J. and Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding  
759 specificity. *Nucleic Acids Research*, **5**(9), 3157–3170.
- 760 Grau, J. (2010). *Discriminative Bayesian principles for predicting sequence signals of gene regulation*. Ph.D. thesis,  
761 Martin Luther University Halle–Wittenberg.
- 762 Grau, J., Keilwagen, J., Gohr, A., Haldemann, B., Posch, S., and Grosse, I. (2012). Jstacs: A Java framework for  
763 statistical analysis and classification of biological sequences. *Journal of Machine Learning Research*, **13**(Jun), 1967–  
764 1971.
- 765 Grau, J., Posch, S., Grosse, I., and Keilwagen, J. (2013). A general approach for discriminative de novo motif discovery  
766 from high-throughput data. *Nucleic Acids Research*, **41**(21), e197.

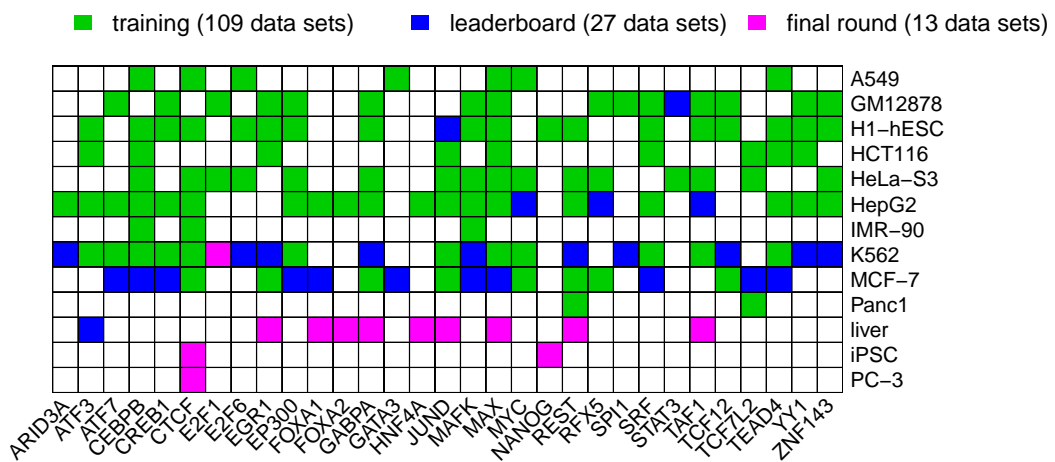
- 767 Grau, J., Grosse, I., Posch, S., and Keilwagen, J. (2015a). Motif clustering with implications for transcription factor  
768 interactions. In *German Conference on Bioinformatics*, volume 3 of *PeerJ Preprints*, page e1601.
- 769 Grau, J., Grosse, I., and Keilwagen, J. (2015b). PRROC: computing and visualizing precision-recall and receiver operating  
770 characteristic curves in R. *Bioinformatics*, **31**(15), 2595–2597.
- 771 Guo, J., Li, T., Schipper, J., Nilson, K. A., Fordjour, F. K., Cooper, J. J., Gordân, R., and Price, D. H. (2014). Sequence  
772 specificity incompletely defines the genome-wide occupancy of Myc. *Genome Biology*, **15**(10), 482.
- 773 Gusmao, E. G., Dieterich, C., Zenke, M., and Costa, I. G. (2014). Detection of active transcription factor binding sites  
774 with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, **30**(22), 3143–3151.
- 775 Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa,  
776 A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes,  
777 G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N.,  
778 Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D.,  
779 Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T. J. (2012). GENCODE: The  
780 reference human genome annotation for the ENCODE project. *Genome Research*, **22**(9), 1760–1774.
- 781 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass,  
782 C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required  
783 for macrophage and B cell identities. *Molecular Cell*, **38**(4), 576–589.
- 784 Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn,  
785 M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions  
786 in vivo by digital genomic footprinting. *Nat Meth*, **6**(4), 283–289.
- 787 Jankowski, A., Tiuryn, J., and Prabhakar, S. (2016). Romulus: robust multi-state identification of transcription factor  
788 binding sites from DNase-seq data. *Bioinformatics*, **32**(16), 2419–2426.
- 789 Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA  
790 interactions. *Science*, **316**(5830), 1497–1502.
- 791 Kähärä, J. and Lähdesmäki, H. (2015). BinDNase: a discriminatory approach for transcription factor binding prediction  
792 using DNase I hypersensitivity data. *Bioinformatics*, **31**(17), 2852–2859.
- 793 Keilwagen, J. and Grau, J. (2015). Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids  
794 Research*.
- 795 Keilwagen, J., Grau, J., Posch, S., and Grosse, I. (2010). Apples and oranges: avoiding different priors in Bayesian DNA  
796 sequence analysis. *BMC Bioinformatics*, **11**(1), 149.
- 797 Keilwagen, J., Grosse, I., and Grau, J. (2014). Area under precision-recall curves for weighted and unweighted data.  
798 *PLoS ONE*, **9**(3), e92209.
- 799 Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-  
800 binding proteins. *Nat Biotech*, **26**(12), 1351–1359.
- 801 Kheradpour, P. and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF  
802 binding experiments. *Nucleic Acids Research*, **42**(5), 2976–2987.
- 803 Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Soboleva, A. V., Kasianov, A. S., Ashoor, H., Ba-alawi, W., Bajic,  
804 V. B., Medvedeva, Y. A., Kolpakov, F. A., and Makeev, V. J. (2016). HOCOMOCO: expansion and enhancement of  
805 the collection of transcription factor binding sites models. *Nucleic Acids Research*, **44**(D1), D116–D125.
- 806 Kumar, S. and Bucher, P. (2016). Predicting transcription factor site occupancy using DNA sequence intrinsic and  
807 cell-type specific chromatin features. *BMC Bioinformatics*, **17**(1), S4.
- 808 Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., and Carey, V. (2013).  
809 Software for computing and annotating genomic ranges. *PLoS Computational Biology*, **9**.
- 810 Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference  
811 genome. *BMC Bioinformatics*, **12**(1), 323.
- 812 Li, H., Liu, H., Wang, Z., Liu, X., Guo, L., Huang, L., Gao, L., McNutt, M. A., and Li, G. (2008). The role of  
813 transcription factors Sp1 and YY1 in proximal promoter region in initiation of transcription of the mu opioid receptor  
814 gene in human lymphocytes. *Journal of Cellular Biochemistry*, **104**(1), 237–250.



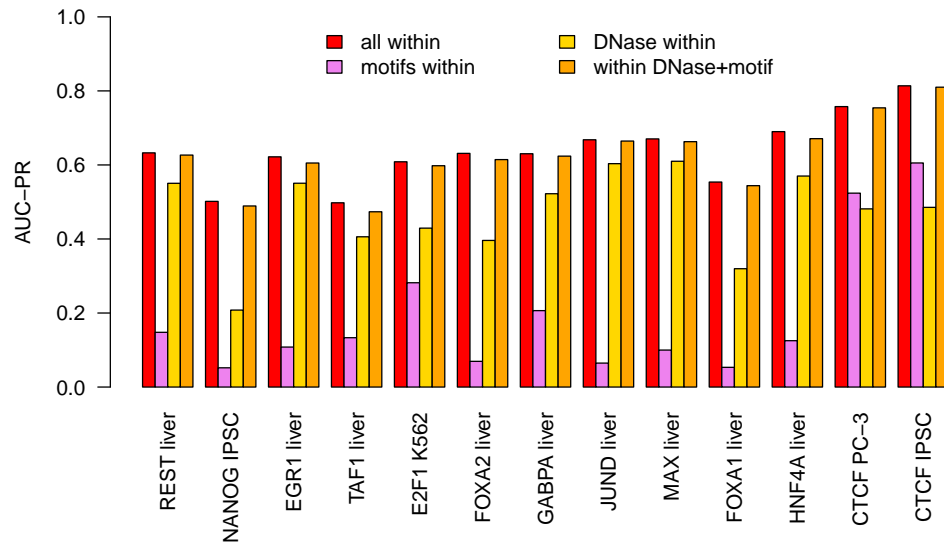
- 815 Liu, S., Zibetti, C., Wan, J., Wang, G., Blackshaw, S., and Qian, J. (2017). Assessing the model transferability for  
816 prediction of transcription factor binding sites based on chromatin accessibility. *BMC Bioinformatics*, **18**(1), 355.
- 817 Luo, K. and Hartemink, A. J. (2012). Using DNase digestion data to accurately identify transcription factor binding  
818 sites. In *Pacific Symposium on Biocomputing*, pages 80–91. World Scientific.
- 819 Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R.,  
820 Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2016). Jasp2016: a major expansion  
821 and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **44**(D1),  
822 D110–D115.
- 823 Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull,  
824 M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006).  
825 TRANSFAC(R) and its module TRANSCOMP(R): transcriptional gene regulation in eukaryotes. *Nucleic Acids  
826 Research*, **34**(suppl.1), D108–110.
- 827 Motallebipour, M., Ameer, A., Reddy Bysani, M. S., Patra, K., Wallerman, O., Mangion, J., Barker, M. A., McKernan,  
828 K. J., Komorowski, J., and Wadelius, C. (2009). Differential binding and co-binding pattern of FOXA1 and FOXA3  
829 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biology*, **10**(11), R129.
- 830 Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E., and Ohler, U. (2012). Predicting cell-type-specific  
831 gene expression from regions of open chromatin. *Genome Research*, **22**(9), 1711–1722.
- 832 Newburger, D. E. and Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on  
833 protein-DNA interactions. *Nucleic Acids Research*, **37**(suppl 1), D77–D82.
- 834 Piper, J., Elze, M. C., Cauchy, P., Cockerill, P. N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the  
835 accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Research*, **41**(21), e201.
- 836 Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of  
837 transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, **21**(3), 447–455.
- 838 Qin, Q. and Feng, J. (2017). Imputation for transcription factor binding predictions based on deep learning. *PLOS  
839 Computational Biology*, **13**(2), 1–20.
- 840 Quang, D. and Xie, X. (2017). FactorNet: a deep learning framework for predicting cell type specific transcription factor  
841 binding from nucleotide-resolution sequential data. *bioRxiv*.
- 842 Rabinovich, A., Jin, V. X., Rabinovich, R., Xu, X., and Farnham, P. J. (2008). E2f in vivo binding specificity: Comparison  
843 of consensus versus nonconsensus binding sites. *Genome Research*, **18**(11), 1763–1777.
- 844 Raj, A., Shim, H., Gilad, Y., Pritchard, J. K., and Stephens, M. (2015). msCentipede: Modeling heterogeneity across  
845 genomic sites and replicates improves accuracy in the inference of transcription factor binding. *PLOS ONE*, **10**(9),  
846 1–15.
- 847 Roos, T., Wettig, H., Grünwald, P., Myllymäki, P., and Tirri, H. (2005). On discriminative Bayesian network classifiers  
848 and logistic regression. *Machine Learning*, **59**(3), 267–296.
- 849 Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating  
850 binary classifiers on imbalanced datasets. *PLOS ONE*, **10**(3), 1–21.
- 851 Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J. K., Ebert, P., Nordström, K.,  
852 Barann, M., Sinha, A., Fröhler, S., Xiong, J., Dehghani Amirabad, A., Behjati Ardakani, F., Hutter, B., Zipprich,  
853 G., Felder, B., Eils, J., Brors, B., Chen, W., Hengstler, J. G., Hamann, A., Lengauer, T., Rosenstiel, P., Walter, J.,  
854 and Schulz, M. H. (2017). Combining transcription factor binding affinities with open-chromatin data for accurate  
855 gene expression prediction. *Nucleic Acids Research*, **45**(1), 54–66.
- 856 Sherwood, R. I., Hashimoto, T., O’Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., Karun, V., Jaakkola, T., and  
857 Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase  
858 profile magnitude and shape. *Nat Biotech*, **32**(2), 171–178.
- 859 Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, **12**, 505–519.
- 860 Stormo, G. D. and Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions.  
861 *Trends in Biochemical Sciences*, **23**(3), 109 – 113.

- 862 Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng,  
863 Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., and Weng, Z. (2012). Sequence features  
864 and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*,  
865 **22**(9), 1798–1812.
- 866 Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert,  
867 S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E.,  
868 Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J., Bouget, F.-Y., Ratsch,  
869 G., Larrondo, L. F., Ecker, J. R., and Hughes, T. R. (2014). Determination and inference of eukaryotic transcription  
870 factor sequence specificity. *Cell*, **158**(6), 1431 – 1443.
- 871 Whitaker, J. W., Chen, Z., and Wang, W. (2015). Predicting the human epigenome from DNA motifs. *Nat Meth*, **12**(3),  
872 265–272.
- 873 Wu, J., Smith, L. T., Plass, C., and Huang, T. H.-M. (2006). ChIP-chip Comes of Age for Genome-wide Functional  
874 Analysis. *Cancer Research*, **66**(14), 6899–6902.
- 875 Xie, D., Boyle, A. P., Wu, L., Zhai, J., Kawli, T., and Snyder, M. (2013). Dynamic trans-acting factor colocalization in  
876 human cells. *Cell*, **155**(3), 713 – 724.
- 877 Ye, B.-Y., Shen, W.-L., Wang, D., Li, P., Zhang, Z., Shi, M.-L., Zhang, Y., Zhang, F.-X., and Zhao, Z.-H. (2016). ZNF143  
878 is involved in CTCF-mediated chromatin interactions by cooperation with cohesin and other partners. *Molecular*  
879 *Biology*, **50**(3), 431–437.
- 880 Zhang, X., Odom, D. T., Koo, S.-H., Conkright, M. D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E.,  
881 Jacobsen, E., Kadam, S., Ecker, J. R., Emerson, B., Hogenesch, J. B., Unterman, T., Young, R. A., and Montminy,  
882 M. (2005). Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target  
883 gene activation in human tissues. *Proceedings of the National Academy of Sciences of the United States of America*,  
884 **102**(12), 4459–4464.

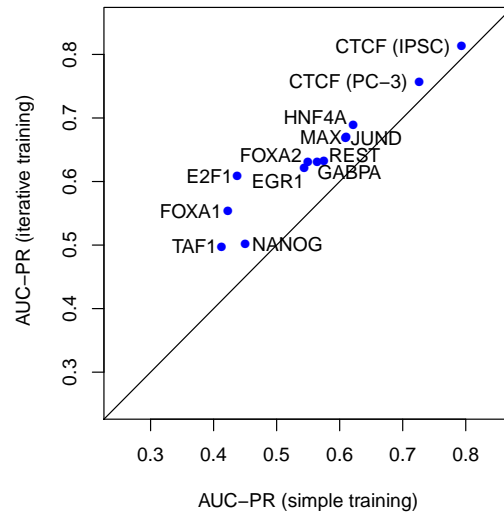
885 **Supplementary Tables and Figures**



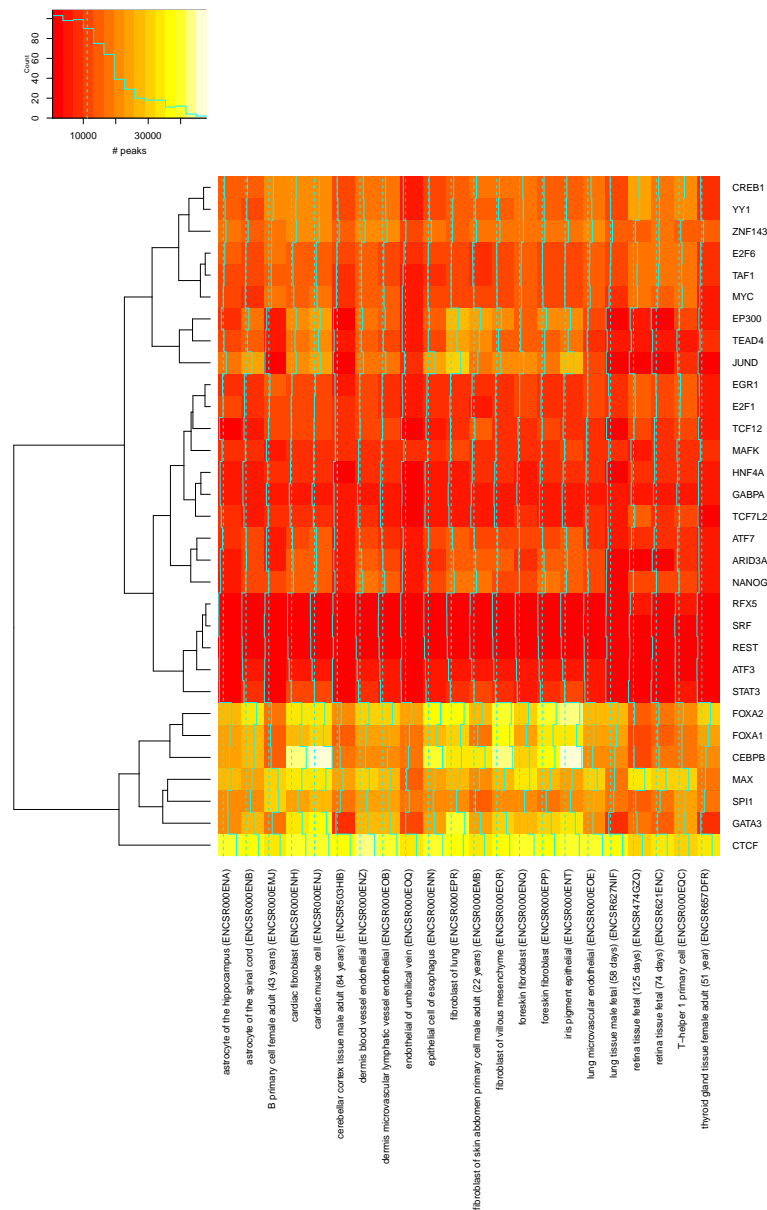
Supplementary Figure S1: Overview of the combinations of cell type and TF in the ENCODE-DREAM training, leaderboard, and final round sets.



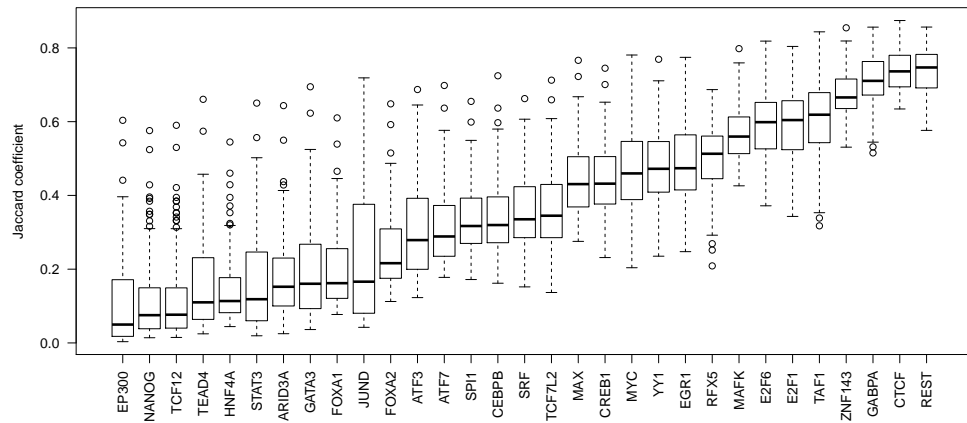
Supplementary Figure S2: Within cell type performance. For each of the 13 combinations of TF and cell type within the test data, we compute the prediction performance (AUC-PR) on the held-out chromosomes of classifiers i) using all features considered, ii) using only motif-based features, iii) using only DNase-seq-based features, and iv) using only motif-based and DNase-seq-based features. The training data comprises the training chromosomes of the same (test) cell type, while predictions are made for the held-out test chromosomes of that cell type.



Supplementary Figure S3: Relevance of the iterative training procedure for within-cell type predictions. For each of the 13 test data sets, we compare the performance (AUC-PR) achieved by the (set of) classifier(s) trained on the initial negative regions (abscissa) with the performance achieved by averaging over all classifiers from the iterative training procedure (ordinate). The training data comprises the training chromosomes of the same (test) cell type, while predictions are made for the held-out test chromosomes of that cell type.



Supplementary Figure S4: Number of predicted peaks in “conservative” peak files for the studied TFs (rows) in the collection of primary cell types and tissues (columns). In each column of the heatmap, cyan trace lines in addition to colors indicate the corresponding values in each cell. In the color scale, the solid cyan line represents the histogram of values observed in the heatmap. Dashed lines indicate median values across all displayed numbers. Rows are clustered by the R `hclust` function using complete linkage.



Supplementary Figure S5: Jaccard coefficients of the different TFs computed on the overlap of the peak files between all pairs of the 22 individual cell types.

Supplementary Table S1: Previous approaches for predicting *in-vivo* transcription factor binding sites and their properties, listed in chronological order.

Approach	Motifs	Accessibility	additional Features	Learning	Model	specifics
CENTIPEDe (Pique-Regi <i>et al.</i> , 2011)	PWMs (TRANS-FAC, Jaspar), de-novo k-mers	DNase-seq	histone modifications	unsupervised	hierarchical mixture model	first predict motifs then matched to DNase-seq profiles
(Natarajan <i>et al.</i> , 2012)	PWMs (TRANS-FAC, Jaspar, UniProbe)	DNase-seq		supervised	Sparse logistic regression	predict gene regulation
(Arvey <i>et al.</i> , 2012)	k-mer based	DNase-seq	histone modifications	supervised	SVMs	cell type-specific binding motifs
Millipede (Luo and Hartemink, 2012)	PWMs (TRANS-FAC, Jaspar)	DNase-seq		supervised	logistic regression	same motifs as CENTIPEDe, binned DNase-seq cuts
Wellington (Piper <i>et al.</i> , 2013)	PWMs (Homer)	DNase-seq		-	based on statistical tests	strand specific cut profiles
PIQ (Sherwood <i>et al.</i> , 2014)	PWMs (TRANS-FAC, Jaspar, UniProbe)	DNase-seq		unsupervised	probabilistic model	first predict motifs then matched to DNase-seq profiles, high resolution
(Gusmao <i>et al.</i> , 2014)	PWMs (TRANS-FAC, Jaspar, UniProbe)	DNase-seq	histone modifications	unsupervised	Hidden Markov model	active footprints annotated with motifs
msCentipede (Raj <i>et al.</i> , 2015)	PWMs (from SE-LEX)	DNase-seq, ATAC-seq		unsupervised	hierarchical multi-scale model	explicitly models heterogeneities in cut profiles
BinDNase (Kähärä and Lähdesmäki, 2015)	PWMs (Factor-book)	DNase-seq		supervised	logistic regression	high resolution, DNase-seq signal TF-specific
(Kumar and Bucher, 2016)	PWMs (Jaspar)	DNase-seq	<i>in-silico</i> nucleosome occupancy, structural features, conservation, CHIP-seq of co-factors (histone modifications, conservation)	supervised	SVMs	also regression
Romulus (Jankowski <i>et al.</i> , 2016)	motif matches (Homer)	DNase-seq		unsupervised (EM)	probabilistic model	motif matches used as prior information
FactorNet (Quang and Xie, 2017)	convolutional neural network	DNase-seq	mappability, annotations, CpG islands, expression	supervised (Deep learning)	convolutional-recurrent neural network	motif discovery part of deep learning
(Liu <i>et al.</i> , 2017)	PWMs (TRANS-FAC)	DNase-seq (ATAC-seq)	conservation, distance to TSS	supervised	Random forests	model based on motif and DNase may be transferred across cell types and TFs
TFImpute (Qin and Feng, 2017)	convolutional neural network	(DNase-seq: negative training regions)		supervised (Deep learning), multi-task learning	deep neural network	complete matrix of cell type-TF combinations
TEPIC (Schmidt <i>et al.</i> , 2017)	PWMs (Jaspar, UniProbe)	DNase-seq	(histone modifications)	-	TRAP scores with exponential distance prior	predict gene expression using elastic net
Mocap (Chen <i>et al.</i> , 2017)	PWMs (EN-CODE, CisBP)	DNase-seq	GC/CpG-content, mappability, distance to TSS, conservation	supervised	sparse logistic regression	three-stage model, ensemble classifier



Supplementary Table S2: Performance (AUC-PR) on the test cell types using different sets of features. Columns “all features”, “motif-based”, “DNase-seq-based”, and “motif & DNase-seq-based” correspond to classifiers using only those feature sets, while columns with prefix “w/o” indicate that the given feature set has been excluded when training the classifiers (for details see main text, Figures 3 and 4).

TF	cell type	all features	motif-based	DNase-seq-based	motif & DNase-seq-based	w/o DNase	w/o motifs	w/o de-novo motifs	w/o Slim/Lsim motifs	w/o RNA-seq	w/o annotation	w/o sequence
CTCF	IPSC	0.807	0.5989	0.479	0.806	0.6028	0.576	0.763	0.778	0.807	0.807	0.807
CTCF	PC-3	0.747	0.5202	0.487	0.745	0.5307	0.572	0.707	0.721	0.747	0.747	0.746
E2F1	K562	0.388	0.2287	0.390	0.382	0.3017	0.326	0.366	0.382	0.384	0.390	0.385
EGR1	liver	0.377	0.0937	0.435	0.376	0.1242	0.354	0.366	0.375	0.375	0.377	0.378
FOXA1	liver	0.487	0.0538	0.259	0.482	0.0713	0.321	0.458	0.478	0.487	0.488	0.482
FOXA2	liver	0.392	0.0460	0.338	0.397	0.0642	0.358	0.443	0.426	0.396	0.392	0.391
GABPA	liver	0.410	0.1868	0.390	0.413	0.2289	0.391	0.412	0.409	0.409	0.411	0.414
HNF4A	liver	0.587	0.1110	0.430	0.577	0.1471	0.509	0.573	0.573	0.586	0.586	0.579
JUND	liver	0.420	0.0446	0.525	0.425	0.0588	0.499	0.438	0.435	0.418	0.419	0.427
MAX	liver	0.424	0.0654	0.485	0.426	0.0928	0.411	0.424	0.422	0.424	0.425	0.426
NANOG	IPSC	0.311	0.0226	0.181	0.319	0.0304	0.291	0.306	0.306	0.313	0.312	0.317
REST	liver	0.251	0.1033	0.180	0.250	0.1315	0.178	0.220	0.230	0.248	0.254	0.250
TAF1	liver	0.383	0.1133	0.360	0.366	0.1720	0.375	0.382	0.384	0.378	0.382	0.381

Supplementary Table S3: Experiment IDs, tissue/cell type information, and biosample “Term ID” of the ENCODE DNase-seq data used in this study. The list of experiments was obtained from [\(https://www.encodeproject.org/report.tsv?type=Experiment&assay\\_title=DNase-seq&status=released&assembly=hg19&files.file\\_type=fastq&audit.NOT\\_COMPLIANT.category\)](https://www.encodeproject.org/report.tsv?type=Experiment&assay_title=DNase-seq&status=released&assembly=hg19&files.file_type=fastq&audit.NOT_COMPLIANT.category). (accessed March 2, 2017)

Experiment ID	Donor ID	Tissue/Cell Type	Term ID
ENCSR000ENA	ENCDO223AAA	astrocyte of the hippocampus	CL:0002604
ENCSR000ENB	ENCDO224AAA	astrocyte of the spinal cord	CL:0002606
ENCSR000ENH	ENCDO095AAA	cardiac fibroblast	CL:0002548
ENCSR000ENJ	ENCDO330AAA	cardiac muscle cell	CL:0000746
ENCSR000ENN	ENCDO104AAA	epithelial cell of esophagus	CL:0002252
ENCSR000ENQ	ENCDO232AAA	foreskin fibroblast	CL:1001608
ENCSR000ENT	ENCDO100AAA	iris pigment epithelial	CL:0002565
ENCSR000EOE	ENCDO238AAA	lung microvascular endothelial	CL:2000016
ENCSR000ENZ	ENCDO241AAA	dermis blood vessel endothelial	CL:2000010
ENCSR000EOB	ENCDO243AAA	dermis microvascular lymphatic vessel endothelial	CL:2000041
ENCSR000EOQ	ENCDO000AAS	endothelial of umbilical vein	CL:0002618
ENCSR000EOR	ENCDO253AAA	fibroblast of villous mesenchyme	CL:0002558
ENCSR000EPP	ENCDO191CQJ	foreskin fibroblast	CL:1001608
ENCSR000EPR	ENCDO269AAA	fibroblast of lung	CL:0002553
ENCSR000EQC	ENCDO334AAA	T-helper 1 primary cell	CL:0000545
ENCSR000EMB	ENCDO442SWC	fibroblast of skin abdomen male adult (22 years)	CL:2000013
ENCSR000EMJ	ENCDO114AAA	B primary cell female adult (43 years)	CL:0000236
ENCSR621ENC	ENCDO539WIJ	retina tissue fetal (74 days)	UBERON:0000966
ENCSR474GZQ	ENCDO225GSN	retina tissue fetal (125 days)	UBERON:0000966
ENCSR503HIB	ENCDO240JUB	cerebellar cortex tissue male adult (84 years)	UBERON:0002129
ENCSR627NIF	ENCDO652XOU	lung tissue male fetal (58 days)	UBERON:0002048
ENCSR657DFR	ENCDO271OUW	thyroid gland tissue female adult (51 year)	UBERON:0002046

Supplementary Table S4: ChIP-seq data sets available for the primary cell types and tissues. The last seven ChIP-seq data sets provide only “relaxed” peak lists.

TF	Experiment ID	File ID	Donor ID	Tissue/Cell Type	Type
CTCF	ENCSR000DSU	ENCF12HCK	ENCDO224AAA	astrocyte of the spinal cord	relaxed
CTCF	ENCSR000DSU	ENCF787GLH	ENCDO224AAA	astrocyte of the spinal cord	conservative
CTCF	ENCSR000DTI	ENCF266GGD	ENCDO330AAA	cardiac muscle cell	relaxed
CTCF	ENCSR000DTI	ENCF386NQE	ENCDO330AAA	cardiac muscle cell	conservative
CTCF	ENCSR000DTR	ENCF528VFN	ENCDO104AAA	epithelial cell of esophagus	relaxed
CTCF	ENCSR000DTR	ENCF373BXG	ENCDO104AAA	epithelial cell of esophagus	conservative
CTCF	ENCSR000DPM	ENCF681OWQ	ENCDO001AAA	fibroblast of lung	relaxed
CTCF	ENCSR000DPM	ENCF138PXI	ENCDO001AAA	fibroblast of lung	conservative
CTCF	ENCSR000DVQ	ENCF738CXX	ENCDO253AAA	fibroblast of villous mesenchyme	relaxed
CTCF	ENCSR000DVQ	ENCF199ZDU	ENCDO253AAA	fibroblast of villous mesenchyme	conservative
CTCF	ENCSR000DWQ	ENCF337WIE	ENCDO191CQJ	foreskin fibroblast	relaxed
CTCF	ENCSR000DWQ	ENCF275AVH	ENCDO191CQJ	foreskin fibroblast	conservative
CTCF	ENCSR000DLW	ENCF002DBA	ENCDO000AAS	endothelial cell of umbilical vein	relaxed
CTCF	ENCSR000DWY	ENCF002DDO	ENCDO269AAA	fibroblast of lung	relaxed
CTCF	ENCSR000DUH	ENCF002DCY	ENCDO232AAA	foreskin fibroblast	relaxed
CTCF	ENCSR000DQI	ENCF649IRT	ENCDO000AAG	foreskin fibroblast	relaxed
JUN	ENCSR000EFA	ENCF002CVC	ENCDO000AAS	endothelial cell of umbilical vein	relaxed
MAX	ENCSR000EEZ	ENCF002CVE	ENCDO000AAS	endothelial cell of umbilical vein	relaxed
MYC	ENCSR000DLU	ENCF002DAZ	ENCDO000AAS	endothelial cell of umbilical vein	relaxed

Supplementary Table S5: Prediction performance on primary cell types and tissues using labels derived from ChIP-seq data. Here, we include all performance measures considered in the ENCODE-DREAM challenge.  
\*: labels determined from only relaxed peaks.

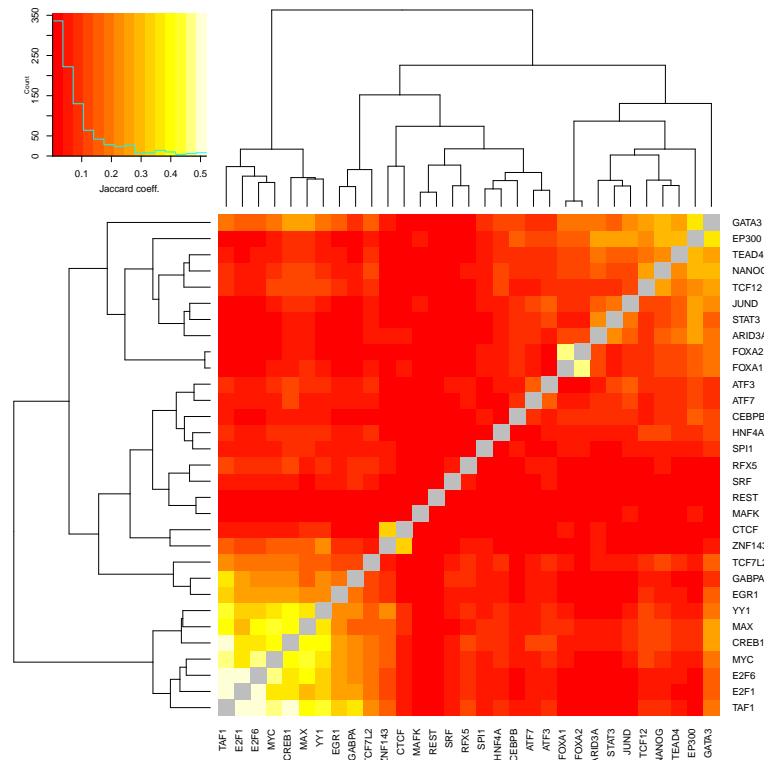
TF	DNase ID	Experiment ID	Matching donor	AUC-ROC	AUC-PR	recall @ 10% FDR	recall @ 50% FDR
CTCF	ENCSR000ENB	ENCSR000DSU	yes	0.9953	0.7895	0.5603	0.8240
CTCF	ENCSR000ENJ	ENCSR000DTI	yes	0.9950	0.8197	0.6316	0.8486
CTCF	ENCSR000ENN	ENCSR000DTR	yes	0.9932	0.7788	0.5621	0.8098
CTCF	ENCSR000EPP	ENCSR000DWQ	yes	0.9939	0.7720	0.5517	0.7975
CTCF	ENCSR000EOR	ENCSR000DVQ	yes	0.9939	0.8048	0.6094	0.8319
CTCF	ENCSR000EPR	ENCSR000DPM	no	0.9913	0.7322	0.4834	0.7600
CTCF*	ENCSR000EOQ	ENCSR000DLW	yes	0.9962	0.7270	0.4030	0.7868
JUN*	ENCSR000EOQ	ENCSR000EFA	yes	0.9965	0.631	0.1644	0.6996
MAX*	ENCSR000EOQ	ENCSR000EEZ	yes	0.9967	0.4004	0.0255	0.3327
MYC*	ENCSR000EOQ	ENCSR000DLU	yes	0.9977	0.1989	0.000	0.0336

Supplementary Table S6: Jaccard coefficients between predicted (columns) and experimentally determined (rows) peak files for CTCF. Entries of matching tissues/cell types are marked in bold. In each row, we mark the largest value in green for matching cell types and in red for differing cell types. We mark matching donor with “(y)”. Jaccard coefficients are computed using the `intersect` and `union` of the `GenomicRanges` R package. For each peak list, entries are sorted by score and limited to the minimum number of peaks across all peak lists.

astrocyte of the spinal cord (ENCSTR000DSU)	0.5852 (y)	0.5732	0.5465	0.5555	0.5675	0.5659	0.5725	0.5537
cardiac muscle cell (ENCSTR000ENJ)	0.5503	<b>0.5754</b> (y)	0.5309	0.5287	0.5408	0.5540	0.5517	0.5395
endothelial cell of umbilical vein (ENCSTR000DLW)	0.6995	0.7049	<b>0.7090</b> (y)	0.6849	0.6730	0.6924	0.6888	0.6664
endothelial cell of esophagus (ENCSTR000DTR)	0.5228	0.5230	0.4998	<b>0.5576</b> (y)	0.5123	0.5147	0.5214	0.5014
epithelial cell of esophagus (ENCSTR000EPM)	0.5361	0.5372	0.5119	0.5094	<b>0.5257</b> (y)	0.5338	<b>0.5383</b>	0.5290
fibroblast of lung (ENCSTR000DWM)	0.6451	0.6478	0.6197	0.6204	<b>0.6391</b> (y)	0.6497	<b>0.6530</b>	0.6501
fibroblast of villous mesenchyme (ENCSTR000DVQ)	0.5818	0.5977	0.5629	0.5663	0.5737	<b>0.6084</b> (y)	0.5938	0.5758
foreskin fibroblast (ENCSTR000DQI)	0.5752	0.5787	0.5558	0.5549	0.5723	0.5837	<b>0.5920</b>	<b>0.5850</b>
foreskin fibroblast (ENCSTR000DUH)	0.6714	0.6755	0.6447	0.6509	0.6637	0.6787	<b>0.6934</b> (y)	<b>0.6770</b>
foreskin fibroblast (ENCSTR000DWQ)	0.5400	0.5402	0.5141	0.5086	0.5301	0.5418	<b>0.5469</b>	<b>0.5680</b> (y)
astrocyte of the spinal cord (ENCSTR000ENB)								
cardiac muscle cell (ENCSTR000ENJ)								
endothelial cell of umbilical vein (ENCSTR000EQ)								
epithelial cell of esophagus (ENCSTR000ENN)								
fibroblast of lung (ENCSTR000EPR)								
fibroblast of villous mesenchyme (ENCSTR000EOR)								
foreskin fibroblast (ENCSTR000ENQ)								
foreskin fibroblast (ENCSTR000EPP)								

Supplementary Table S7: Jaccard coefficient between experimentally determined and predicted peak files. Jaccard coefficients are computed using the **intersect** and **union** of the GenomicRanges R package. For each TF, entries of the experimentally determined and predicted peak lists are sorted by score and limited to the minimum number of peaks in either of the two peak lists.

TF	ChIP-seq	Predicted	Matching donor	Jaccard coefficient
JUN	endothelial cell of umbilical vein (ENC5R000EFA)	endothelial cell of umbilical vein (ENC5R000EOQ)	yes	0.4500
MAX	endothelial cell of umbilical vein (ENC5R000EEZ)	endothelial cell of umbilical vein (ENC5R000EOQ)	yes	0.3634
MYC	endothelial cell of umbilical vein (ENC5R000DLU)	endothelial cell of umbilical vein (ENC5R000EOQ)	yes	0.2221



Supplementary Figure S6: Average Jaccard coefficients computed on the overlap of the peak files of pairs of TFs for matched cell types. In the color scale, the solid cyan line represents the histogram of values observed in the heatmap. Dashed lines indicate the value at the center bin of the color scale. Rows and columns are clustered by the R `hclust` function using complete linkage.

## 886 **Supplementary Methods**

### 887 **Supplementary Text S1 – Features**

888 The features described in the following are all determined on the level of genome bins.  
889 We refer to the bin for which the a-posteriori probability of being peak center should  
890 be computed (i.e., the bin containing the peak summit in case of positive examples) as  
891 *center bin*. Further, adjacent bins considered are defined relative to that center bin (see  
892 also section Prediction schema).

#### 893 **S1.1 Sequence-based features**

894 As a first sequence-based feature, we consider the raw DNA sequence according to the  
895 *hg19* human genome sequence in the center bin and the directly preceding and the  
896 directly following bin. In total, this corresponds to 150 bp of sequence, centered at the  
897 center bin.

898 We further consider the mean G/C-content, and the relative frequency of CG di-  
899 nucleotides in the raw sequence spanning those three bins centered at the center bin.  
900 G/C-content might be an informative property of promoters bound by a certain TF,  
901 and an enrichment of CG di-nucleotides might be informative about the presence of  
902 CpG islands.

903 We also compute the Kullback-Leibler divergence between the relative frequencies of  
904 all tri-nucleotides in each of these three bins compared with their relative frequencies  
905 in the complete genome. As a feature, we then consider the maximum of those three  
906 Kullback-Leibler divergence values obtained for the three bins. Here, the reasoning is  
907 that a deviation from the genomic distribution of tri-nucleotides might be a sign of the  
908 general information content of a sequence, which might help to distinguish coding and  
909 non-coding DNA regions as well as identifying regions that encode regulatory informa-  
910 tion.

911 Finally, we consider the length of the longest poly-A or poly-T tract, the length of the  
912 longest poly-C or poly-G tract, the length of the longest poly-A/T tract, and the length  
913 of the longest poly-G/C tract in these three bins.

914 All of those sequence-based features are neither TF-specific nor cell type-specific, but  
915 model parameters learned on their feature values might well be different for different  
916 training TFs or cell types.

#### 917 **S1.2 Annotation-based features**

918 Based on the Gencode v19 genome annotation of the *hg19* genome, we derive a set  
919 of annotation-based features. First, we consider the distance of the current center bin  
920 to the closest TSS annotation (regardless of its strand orientation), which might be  
921 informative about core promoter regions. Second, we collect the binary information if  
922 the current center bin overlaps with annotations of i) a CDS, ii) a UTR, iii) an exon,  
923 iv) a transcript, or v) a TSS annotation, separately for each of the two possible strand  
924 orientations. Like some of the previous features, this helps to identify coding, non-coding



925 but transcribed, core promoter, and intergenic regions. Again, these features are not TF  
926 or cell type-specific, but model parameters may be adapted specifically for a TF or cell  
927 type.

### 928 **S1.3 Motif-based features**

929 As it might be expected that binding motifs are pivotal for predicting TF-specific bind-  
930 ing regions, we create a large collection of motifs for each of the TFs considered. For  
931 each of the TFs, we collect all position weight matrix models from the HOCOMOCO  
932 database (Kulakovskiy *et al.*, 2016) as well as our in-house database DBcorrDB (Grau  
933 *et al.*, 2015a), and Slim/LSlim models of the respective TFs from a previous publica-  
934 tion (Keilwagen and Grau, 2015). In addition, we learn a large set of motifs from the  
935 data provided in the challenge using our motif discovery tools Dimont (Grau *et al.*, 2013)  
936 using PWM as well as LSlim(3) models (Keilwagen and Grau, 2015). Specifically, we  
937 perform motif discovery for

- 938 • PWM models from the “conservative” peak files for each training cell type,
- 939 • PWM models from the “relaxed” peak files complemented by negative regions se-  
940 lected to be DNase positive (i.e., open chromatin) but ChIP-seq negative according  
941 to the ChIP-seq and DNase-seq peak files provided with the challenge data,
- 942 • LSlim(3) models from the “conservative” peak files for each training cell type,
- 943 • LSlim(3) models from the “relaxed” peak files for each training cell type,
- 944 • LSlim(3) models from the “relaxed” peak files complemented by negative regions  
945 selected to be DNase positive (i.e., open chromatin) but ChIP-seq negative accord-  
946 ing to the ChIP-seq and DNase-seq peak files provided with the challenge data.

947 LSlim(3) may capture intra-motif dependencies between binding site position with a  
948 distance of at most three nucleotides.

949 Motifs discovered using models of different complexity on these different sets of training  
950 data (“conservative” and “relaxed” peaks, and “relaxed” peaks complemented by DNase  
951 positive regions) should i) capture the breadth of the binding landscape of a TF as  
952 represented by the different levels of stringency (“conservative” vs. “relaxed”), and ii)  
953 represent potential intra-motif dependencies as well as traditional, “additive” binding  
954 affinities. In addition, we learn motifs from the DNase-seq peak files as well, considering

- 955 • LSlim(3) models from the “conservative” and “relaxed” DNase-seq peak files,
- 956 • LSlim(3) models from the regions in the intersection of all “relaxed” DNase-seq  
957 peak files.

958 Learning motifs from the DNase-seq data alone might have the potential to capture  
959 additional binding motifs of TFs that are important for cell type-specific predictions but  
960 are not represented in the ChIP-seq data provided with the challenge data.

961 Regardless of the TF considered, we further include PWM and Slim/LSlim motifs  
962 discovered previously (Keilwagen and Grau, 2015; Grau *et al.*, 2015a) for CTCF, SP1,  
963 JUND, and MAX, as those i) mark boundaries between regulatory regions, ii) frequently  
964 interact with other transcriptions factor, or iii) bind to a large fraction of active promot-  
965 ers. Further TFs that might interact with the currently considered TF as determined i)  
966 from the literature, specifically from Factorbook (Wang *et al.*, 2012), ii) determined from  
967 the overlap between the ChIP-seq peaks provided with the challenge data. The latter is  
968 accomplished by computing for each TF and cell type i) the TF with the largest overlap  
969 (F1 measure computed on the peaks) and ii) the TF with the lowest overlap between the  
970 peak files. The former might be indicative of co-binding, while the latter might indicate  
971 mutually exclusive binding, both of which might help to predict TF-specific binding  
972 regions.

973 Finally, we consider motifs determined by the epigram pipeline (Whitaker *et al.*, 2015),  
974 which mark epigenetic modifications. Specifically, we select the top 10 motifs reported  
975 for “single mark” analyses for methylation, and H3K4me3 and H3K27ac histone mod-  
976 ifications (downloaded from <http://wanglab.ucsd.edu/star/epigram/mods/index.html>).  
977 [html](http://wanglab.ucsd.edu/star/epigram/mods/index.html)).

978 We use all motif models described above to scan the hg19 genome for potential binding  
979 regions. To this end, we apply a sliding window approach across the genome, and  
980 aggregate the motif scores obtained according to the genomic bins. For the TF-specific  
981 motifs obtained by de-novo motif discovery from ChIP-seq data, we consider as features

- 982 • the maximum log-probability of all sliding windows starting in the center bin,
- 983 • the logarithm of the sum of binding probabilities in all sliding windows starting in  
984 the center bin or its two adjacent bins, and
- 985 • the logarithm of the sum of binding probabilities in all sliding windows starting in  
986 any of the bins considered.

987 The first feature should capture the binding affinity at the strongest binding site around  
988 the peak summit, while the latter two features represent the general binding affinity of  
989 a region with different levels of resolution.

990 For all of the remaining motifs, we consider the maximum of the bin-wise logarithm  
991 of the sum of binding probabilities over all bins considered (see section Binning the  
992 genome), as this reduces memory requirements as well as model complexity and this  
993 level of detail might be sufficient to capture TF interactions.

#### 994 **S1.4 DNase-based features**

995 For the DNase-seq data, the challenge provided tracks with a “fold-enrichment coverage”  
996 track, peak files, and the original BAM files from mapping the DNase-seq reads, of which  
997 we consider only the former two. From the fold-enrichment coverage track, we compute  
998 the following statistics:

- 999 • the minimum value across the center bin and its two adjacent bins,

- 1000 • the minimum of the maximum value within each bin considered,
- 1001 • the minimum of the 25% percentile within each bin considered, and
- 1002 • the median values of all the bins considered.

1003 After extracting those feature values for all genomic bins, we quantile normalize each  
1004 of the features independently across the challenge cell types. Before normalization, we  
1005 randomize the order of values to avoid systematic effects due to genomic order, which  
1006 might especially occur for the large number of very low values. For the additional,  
1007 primary cell types, we do not perform an independent quantile normalization but instead  
1008 map the DNase-seq features (according to their numerical order) to the corresponding,  
1009 quantile normalized values of the challenge cell types.

1010 In addition to these short-range DNase features, we also determine a set of long-range  
1011 features, which are computed from i) 10 bins ii) 20 bins, and iii) 40 bins preceding and  
1012 succeeding the current center bin. These features are

- 1013 • the minimum value across all bins,
- 1014 • the maximum value across all bins,
- 1015 • the minimum value across the bins preceding the center bin,
- 1016 • the minimum value across the bins succeeding the center bin,
- 1017 • the maximum value across the bins preceding the center bin, and
- 1018 • the maximum value across the bins succeeding the center bin.

1019 Together, these features capture chromatin accessibility on a short and long range level  
1020 with reasonable resolution, which should be highly informative with regard to the general  
1021 TF-binding potential. Model parameters should then be able to adapt for TF-specific  
1022 preferences of chromatin accessibility.

1023 For the current center bin, we additionally determine features of stability across the  
1024 different cell types, namely

- 1025 • the ratio of the minimum value in the current cell type divided by the average of  
1026 the minimum values across all cell types,
- 1027 • the ratio of the maximum value in the current cell type divided by the average of  
1028 the maximum values across all cell types,
- 1029 • the coefficient of variation (standard deviation divided by mean) of the minimum  
1030 values across all cell types, and
- 1031 • the coefficient of variation of the maximum values across all cell types,

1032 where the latter two features are identical for all cell types by design.

1033 We also determine several features that represent the monotonicity/stability of these  
1034 DNase-seq signals. Specifically, these features are

- 1035 ● the number of steps (increasing or decreasing) in the track profile in a 450 bp  
1036 interval centered at the center bin,
- 1037 ● the longest strictly monotonically increasing stretch in the four bins preceding the  
1038 center bin,
- 1039 ● the longest strictly monotonically decreasing stretch in the four bins preceding the  
1040 center bin,
- 1041 ● the longest strictly monotonically increasing stretch in the four bins succeeding  
1042 the center bin, and
- 1043 ● the longest strictly monotonically decreasing stretch in the four bins succeeding  
1044 the center bin.

1045 The first of these features has been inspired by the “orange” feature coined by team  
1046 autosome.ru in the challenge.

1047 Finally, we define further features based on the “conservative” and “relaxed” DNase-  
1048 seq peak files as provided with the challenge data. These are

- 1049 ● the distance of the center bin to the summit of the closest conservative peak,
- 1050 ● the distance of the center bin to the summit of the closest relaxed peak,
- 1051 ● the peak statistic of a conservative peak overlapping the center bin (or zero if no  
1052 such overlapping peak exists) multiplied by the length of the overlap,
- 1053 ● the peak statistic of a relaxed peak overlapping the center bin (or zero if no such  
1054 overlapping peak exists) multiplied by the length of the overlap,
- 1055 ● the maximum of the q-values of an overlapping conservative peak (or zero if no  
1056 such overlapping peak exists) multiplied by the length of the overlap across the  
1057 five central bins,
- 1058 ● the maximum of the q-values of an overlapping relaxed peak (or zero if no such  
1059 overlapping peak exists) multiplied by the length of the overlap across the five  
1060 central bins.

### 1061 **S1.5 RNA-seq-based features**

1062 The RNA-seq data provided with the challenge data included the TPM values of genes  
1063 according to the gencode v19 genome annotation. TPM values are also quantile normal-  
1064 ized across the cell types. As features, we consider

- 1065 ● the maximum TPM value (averaged over the two bio-replicates per cell type) of  
1066 genes in at most 2.5 kb distance
- 1067 ● the coefficient of variation of the bio-replicated of the corresponding gene,

- 1068 • the relative difference (difference of values in bio-replicated divided by their mean  
1069 value) of the corresponding gene.

1070 In analogy to the DNase-based features, we computed from the first feature as measures  
1071 of stability across the different cell types

- 1072 • the ratio of the maximum TPM value in the current cell type divided by the average  
1073 of the maximum values across all cell types, and

- 1074 • the coefficient of variation of the maximum TPM values across all cell types.

## 1075 Supplementary Text S2 – Model & learning principle

1076 For numerical features  $x$ , we use independent Gaussian densities parameterized as

$$\mathcal{N}(x|\lambda, \mu) := \sqrt{\frac{e^\lambda}{2\pi}} \cdot e^{-\frac{e^\lambda}{2}(x-\mu)^2},$$

1077 which allows for unconstrained numerical optimization of both,  $\lambda$  and  $\mu$ .

1078 For features  $y$  with  $K$  possible discrete values  $v_1, \dots, v_K$ , we use (unnormalized) multi-  
1079 nomial distributions with parameters  $\beta = (\beta_1, \dots, \beta_K)$  defined as

$$\mathcal{B}(y|\beta) := \prod_{k=1}^K \left( \frac{\exp(\beta_k)}{\sum_{\ell} \exp(\beta_{\ell})} \right)^{\delta(y=v_k)}.$$

1080 The multinomial coefficient is neglected in this case, since it only depends on the in-  
1081 put data but not on the model parameters. In case of binary features, i.e.,  $K=2$ , this  
1082 corresponds to an (unnormalized) binomial distribution.

1083 For modeling the raw sequence  $\mathbf{s} = s_1 s_2 \dots s_L$ ,  $s_{\ell} \in \Sigma = \{A, C, G, T\}$ , we use a  
1084 homogeneous Markov model of order 3 parameterized as

$$\mathcal{M}(\mathbf{s}|\beta_s) := \frac{\exp(\beta_{1,s_1})}{\sum_{a \in \Sigma} \exp(\beta_{1,a})} \cdot \frac{\exp(\beta_{2,s_2|s_1})}{\sum_{a \in \Sigma} \exp(\beta_{2,a|s_1})} \cdot \frac{\exp(\beta_{3,s_3|s_1 s_2})}{\sum_{a \in \Sigma} \exp(\beta_{3,a|s_1 s_2})} \cdot \prod_{\ell=4}^L \frac{\exp(\beta_{h,s_{\ell}|s_{\ell-3} s_{\ell-2} s_{\ell-1}})}{\sum_{a \in \Sigma} \exp(\beta_{h,a|s_{\ell-3} s_{\ell-2} s_{\ell-1}})},$$

1085 where  $\beta_{h,a|b}$ ,  $a \in \Sigma$ ,  $b \in \Sigma^3$  are the homogeneous parameters and

1086  $\beta_s = (\beta_{1,A}, \dots, \beta_{1,T}, \beta_{2,A|A}, \dots, \beta_{2,T|T}, \beta_{3,A|AA}, \dots, \beta_{3,T|TT}, \beta_{h,A|AAA}, \dots, \beta_{h,T|TTT})$  de-  
1087 notes the vector of all model parameters.

1088 Let  $\mathbf{x} = (x_1, \dots, x_N)$  denote the vector of all numerical features,  $\mathbf{y} = (y_1, \dots, y_M)$  de-  
1089 note the vector of all discrete features, and  $\mathbf{s}$  denote the raw sequence of one region repre-  
1090 sented by its feature values  $\mathbf{z} = (\mathbf{x}, \mathbf{y}, \mathbf{s})$ . Let  $\theta = (\lambda_1, \dots, \lambda_N, \mu_1, \dots, \mu_N, \beta_1, \dots, \beta_M, \beta_s)$   
1091 denote the set of all model parameters. We compute the likelihood of  $\mathbf{z}$  as an independent  
1092 product of the terms for the individual features, i.e.,

$$P(\mathbf{z}|\theta) := \left( \prod_{\ell=1}^N \mathcal{N}(x_{\ell}|\lambda_{\ell}, \mu_{\ell}) \right) \cdot \left( \prod_{\ell=1}^M \mathcal{B}(y_{\ell}|\beta_{\ell}) \right) \cdot \mathcal{M}(\mathbf{s}|\beta_s).$$

1093 For modeling the distribution in the positive (foreground) and negative (background)  
1094 class, we use likelihoods  $P(\mathbf{z}|\boldsymbol{\theta}_{fg})$  and  $P(\mathbf{z}|\boldsymbol{\theta}_{bg})$  with independent sets of parameters  
1095  $\boldsymbol{\theta}_{fg}$  and  $\boldsymbol{\theta}_{bg}$ , respectively. In addition, we define the a-priori class probabilities as  
1096  $P(fg|\gamma_1, \gamma_2) := \frac{\exp(\gamma_1)}{\exp(\gamma_1) + \exp(\gamma_2)}$  and  $P(bg|\gamma_1, \gamma_2) = \frac{\exp(\gamma_2)}{\exp(\gamma_1) + \exp(\gamma_2)}$ .  
1097 Based on these definitions, we may compute the a-posteriori class probability of the  
1098 positive class as

$$P(fg|\mathbf{z}, \boldsymbol{\theta}_{fg}, \boldsymbol{\theta}_{bg}, \gamma) = \frac{P(fg|\gamma_1, \gamma_2) \cdot P(\mathbf{z}|\boldsymbol{\theta}_{fg})}{P(fg|\gamma_1, \gamma_2) \cdot P(\mathbf{z}|\boldsymbol{\theta}_{fg}) + P(bg|\gamma_1, \gamma_2) \cdot P(\mathbf{z}|\boldsymbol{\theta}_{bg})},$$

1099 and the a-posteriori class probability of the negative class in complete analogy.

1100 Using the discriminative maximum conditional likelihood principle (Roos *et al.*, 2005),  
1101 the parameters are optimized such that the a-posteriori probabilities of the correct class  
1102 labels given data and parameters are maximized. Here, we use a variant (Grau, 2010)  
1103 of the maximum conditional likelihood principle that incorporates weights. Let  $\mathbf{F} =$   
1104  $(\mathbf{z}_1, \dots, \mathbf{z}_I)$  denote the set of positive examples and let  $\mathbf{B} = (\mathbf{z}_{I+1}, \dots, \mathbf{z}_J)$  denote the  
1105 set of negative examples, where  $\mathbf{z}_i$  is assigned weight  $w_i$ . The parameters are then  
1106 optimized with regard to

$$(\boldsymbol{\theta}_{fg}^*, \boldsymbol{\theta}_{bg}^*, \gamma^*) = \underset{(\boldsymbol{\theta}_{fg}, \boldsymbol{\theta}_{bg}, \gamma)}{\operatorname{argmax}} \left[ \sum_{i=1}^I w_i \cdot \log P(fg|\mathbf{z}_i, \boldsymbol{\theta}_{fg}, \boldsymbol{\theta}_{bg}, \gamma) + \sum_{i=I+1}^J w_i \cdot \log P(bg|\mathbf{z}_i, \boldsymbol{\theta}_{fg}, \boldsymbol{\theta}_{bg}, \gamma) \right].$$

### 1107 **Supplementary Text S3 – Sampling of DNase-matched negative regions**

1108 We sample negative regions with chromatin accessibility values matched to the positive  
1109 regions (following an idea related to importance sampling) as explained in the following.  
1110 We consider the center bins of all positive regions, collect the corresponding DNase-  
1111 seq median feature values (see Supplementary Text S1) of those bins, and determine a  
1112 histogram of the collected values. The histogram is composed of 20 equally sizes bins  
1113 between the observed maximum and minimum values of the DNase-seq median values.  
1114 This histograms represents an approximation of the distribution of DNase-seq median  
1115 values in the positive regions. As we expect DNase-seq values to be highly informative  
1116 about TF binding, we aim at sampling a representative set of negative regions that  
1117 exhibit similar DNase-seq values but might be distinguished from positive regions by  
1118 other features.

1119 To this end, we assign each of the negative regions to the same histogram bins based  
1120 on their respective DNase-seq median values at their center bins. This also yields an  
1121 analogous histogram of the DNase-seq median values for the negative regions, which will  
1122 usually be different from the histogram for the positive regions.

1123 Within each histogram bin, we then draw a subset of the negative regions assigned to  
1124 that bin by i) drawing a subset of these regions four times as large as the corresponding

1125 positive set, and ii) weighting the drawn negative regions such that the sum of weights  
1126 matches the relative abundance of that histogram bin in the histogram on all negative  
1127 region.

1128 Conceptually, this procedure yields an over-sampling of negative regions with large  
1129 DNase-seq median features, which is adjusted for by down-weighting such examples to  
1130 the corresponding frequency on the chromosome level. This is especially important as  
1131 these will be regions that are hard to classify using DNase-seq based features but are  
1132 only lowly represented by the uniform sampling schema.