

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

Jens Keilwagen¹, Stefan Posch², and Jan Grau²

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

²Institute 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 features derived from chromatin accessibility, binding motifs, gene expression, genomic sequence and annotation to train classifiers using a supervised, discriminative learning principle. Two further key aspects of this approach are 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 may help to yield biological conclusions. Finally, we provide a user-friendly version of our approach as open source software at <http://jstacs.de/index.php/Catchitt>.
Contact: 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 gene regulation depends on transcription factors (TFs), which
38 specifically bind directly to sites in promoters or enhancers of regulated genes or bind
39 indirectly via other, sequence specific TFs. Modeling binding specificities, typically rep-
40 resented as sequence motifs, has been an important topic of bioinformatics since its
41 infancy (Staden, 1984; Berg and von Hippel, 1987). However, it soon became evident
42 that *in-silico* binding site predictions based on sequence motifs alone are insufficient to
43 explain *in-vivo* binding of TFs and that determinants beyond sequence specificity likely
44 play an important 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-specific and cell type-specific binding, the experimental
49 effort and costs currently preclude ChIP-seq experiments for hundreds to thousands of
50 TFs 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 beyond 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; detailed discussion
64 in Supplementary Text Supplementary Text S1) has been proposed over the last years
65 (e.g., Pique-Regi *et al.* (2011); Natarajan *et al.* (2012); Arvey *et al.* (2012); Luo and
66 Hartemink (2012); Piper *et al.* (2013); Sherwood *et al.* (2014); Gusmao *et al.* (2014);
67 Raj *et al.* (2015); Kähärä and Lähdesmäki (2015); Kumar and Bucher (2016); Jankowski
68 *et al.* (2016); Quang and Xie (2017); Liu *et al.* (2017); Qin and Feng (2017); Schmidt
69 *et al.* (2017); Chen *et al.* (2017)). While the general notion of combining sequence signals
70 with chromatin accessibility data and, in some cases, other features is common to the
71 majority of approaches, they differ in several aspects. Specifically, approaches differ
72 in the source of motif information, which may stem from motif databases or from *de-*
73 *novo* motif discovery. Matches to these motifs are either used as prior information and
74 filtered by their respective DNase-seq signals in a subsequent step, or DNase footprints
75 are first detected and annotated with TFs based on motif matches in those footprints, or,
76 finally, motif and DNase-seq information are processed jointly. Supervised approaches
77 rely on labeled training data, whereas unsupervised approaches may be applied without

78 any a-priorily known binding sites of the TF at hand. Finally, motif and chromatin
79 accessibility data may be complemented with further experimental or computational
80 assays like histone modifications or sequence conservation.

81 Each of these approaches has its benefits and downsides, and the results of bench-
82 mark studies in the respective original publications are ambiguous with regard to their
83 prediction performance. Against this background, the “ENCODE-DREAM in vivo
84 Transcription Factor Binding Site Prediction Challenge” ([https://www.synapse.org/
85 #!Synapse:syn6131484](https://www.synapse.org/#!Synapse:syn6131484)) aimed at assessing the performance of tools for predicting cell
86 type-specific TF binding in human using a minimal set of experimental data in a fair
87 and unbiased manner. The challenge setting has advantages over typical benchmark
88 studies, because approaches are typically applied to the challenge data by their authors,
89 ground truth is known only by the challenge organizers, and participants are typically
90 required to provide code and documentation for their method such that predictions can
91 be reproduced.

92 Participants in the ENCODE-DREAM challenge were allowed to use binding motifs
93 from any source, genomic sequence, gene annotations, *in-silico* DNA shape predictions,
94 and cell type-specific DNase-seq and RNA-seq data. In addition, TF ChIP-seq data and
95 ChIP-seq-derived labels (“bound”, “unbound”, “ambiguous”) were provided for training
96 cell types and training chromosomes. Predictions had then to be made for combinations
97 of TF and cell type not present in the training data on held-out chromosomes. Predic-
98 tions were evaluated against labels derived from TF ChIP-seq data for that specific TF
99 and test cell type.

100 Here, we present our approach for predicting cell type-specific TF binding regions
101 earning a shared first rank among 40 international teams, including developers of sev-
102 eral established methods ([https://www.synapse.org/#!Synapse:syn6131484/wiki/
103 405275](https://www.synapse.org/#!Synapse:syn6131484/wiki/405275)). The approach presented in this paper combines several novel ideas. First,
104 we consider motifs from databases, but also motifs learned by de-novo motif discovery
105 from ChIP-seq and DNase-seq data using sparse local inhomogeneous mixture (Slim)
106 models (Keilwagen and Grau, 2015), which may capture short to mid-range intra-motif
107 dependencies. Second, we process DNase-seq data following the binning idea of previ-
108 ous approaches but defining novel statistics computed from the data in those bins, and
109 derive several sequence-based, annotation-based, and RNA-seq-based features. Third,
110 we apply a supervised machine learning approach that employs a discriminative learning
111 principle, which is related to logistic regression but allows for explicit model assumptions
112 with regard to different features. Fourth, discriminative learning is combined with an
113 iterative training approach for refining sets of representative negative examples. Finally,
114 we combine the predictions of classifiers trained in different of these iterations and on
115 different training cell types in an ensemble-like approach.

116 As this novel approach has already been benchmarked against a large number of com-
117 peting approaches as part of the ENCODE-DREAM challenge ([https://www.synapse.
118 org/#!Synapse:syn6131484/wiki/405275](https://www.synapse.org/#!Synapse:syn6131484/wiki/405275)), we focus on the analysis for the contribu-
119 tions of different aspects of this approach on the final prediction performance in this
120 paper. Specifically, we evaluate the contribution of different features, we compare the
121 performance achieved by standard training with that achieved by the iterative training

122 procedure, and we assess the performance of individual classifiers compared with their
123 ensemble prediction. Based on these analyses, we define and benchmark a simplified
124 variant of the proposed approach. Finally, we provide a large collection of predicted, cell
125 type-specific tracks of binding regions for 31 TFs in 22 primary cell types and tissues to
126 make predictions by this approach readily accessible.

127 2 Methods

128 2.1 Data

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

- 131 • the raw sequence of the human genome (hg19) and gene annotations according
132 to the gencode v19 annotation (<http://www.gencodegenes.org/releases/19.html>) (Harrow *et al.*, 2012),
- 133
- 134 • cell type-specific DNase-seq “fold-enrichment coverage” tracks, which represent
135 DNase-seq signal relative to a pseudo control, smoothed in a 150 bp window,
- 136 • cell type-specific DNase-seq peak files in “conservative” (IDR threshold of 10% in
137 pseudo replicates) and “relaxed” (no IDR threshold) flavors,
- 138 • cell type-specific TPM values from RNA-seq experiments in two bio-replicates for
139 all gencode v19 genes as estimated by RSEM (Li and Dewey, 2011),
- 140 • cell type-specific and TF-specific ChIP-seq peak files in “conservative” (IDR thresh-
141 old of 10% in pseudo replicates) and “relaxed” (no IDR threshold) flavors,
- 142 • cell type-specific and TF-specific label files classifying genome-wide 200 bp regions
143 every 50 bp into B=“bound”, A=“ambiguous”, and U=“unbound” according to
144 the respective conservative and relaxed ChIP-seq peak files; an overview of the
145 combinations of TF and cell type in the training data, the leaderboard data, and
146 the test data used for evaluation in the final challenge round is given in Supple-
147 mentary Figure S1.

148 In addition, we download sequence motifs represented as PWMs from the following
149 collections:

- 150 • TF-specific motifs from the databases HOCOMOCO (Kulakovskiy *et al.*, 2016)
151 and DBcorrDB (Grau *et al.*, 2015a),
- 152 • motifs related to epigenetic markers from the epigram pipeline (Whitaker *et al.*,
153 2015).

154 Details about the motifs considered are given in section “Features” and Supplementary
155 Text S2.

156 For predicting cell type-specific binding of TFs in additional cell types beyond those
157 considered in the challenge, we download DNase-seq data (FastQ format) from the EN-
158 CODE project (encodeproject.org). Specifically, we select all DNase-seq experiments
159 that i) are flagged as “released”, ii) have FastQ files available, iii) are not from immortal-
160 ized cell lines, iv) have no entry in one of the “Audit error” categories, and v) are not in
161 the “insufficient replicate concordance” category of “Audit not compliant”. A list of the
162 corresponding experiments is obtained from the ENCODE project and experiments are
163 filtered for the existence of at least two replicates, yielding 23 experiments in total. One
164 of these experiments had to be excluded later, because a different DNase protocol with
165 much shorter reads had been used. For the remaining 22 experiments (Supplementary
166 Table S3), all FastQ files are downloaded from ENCODE and processed using ATAC-
167 Seq/DNase-Seq Pipeline (https://github.com/kundajelab/atac_dnase_pipelines,
168 latest git commit: c1d07d38a02af2f0319a69707eee047ab6112ecc (Tue Mar 21 20:31:25
169 2017)). The data sets are analyzed using the following parameters
170 `-species hg19 -type dnase-seq -subsample 50M -se`. For further analyzes, the re-
171 laxed (`./out/peak/idr/pseudo_reps/rep1/*.filt.narrowPeak.gz`) and conservative
172 peaks (`./out/peak/macs2/overlap/*pval0.1*.filt.narrowPeak.gz`) as well as the
173 DNase coverage (`./out/signal/macs2/rep1/*.fc.signal.bigwig`) are used.

174 In addition, we download ChIP-seq peak files (Supplementary Table S4) matching
175 these cell types and one of the TFs considered. Based on the “relaxed” (i.e., “optimal
176 idr thresholded peaks”) and “conservative” (i.e., “conservative idr thresholded peaks”)
177 peak files, we derive labels for 200 bp windows every 50 bp as proposed for the challenge.
178 Specifically, we label each 200 bp region overlapping a conservative peak by at least
179 100 bp as “bound”. Of the remaining regions, all regions that overlap a relaxed peak
180 by at least 1 bp are labeled “ambiguous”, while all other regions are labeled “unbound”.
181 For a subset of TFs, no conservative peaks are available due to the lack of replicates. In
182 such cases, we also use the relaxed peaks to assign “bound” labels.

183 **2.2 Binning the genome**

184 As the final prediction is requested for overlapping 200 bp regions with an offset of
185 50 bp, we decide to compute features with a matching resolution of 50 bp. To this
186 end, the genome is divided into non-overlapping bins of 50 bp. Features are then either
187 computed directly with that resolution (where possible, e.g., distance to the closest TSS),
188 or first computed with base-pair resolution and afterwards summarized as aggregate
189 values (minimum, maximum, median, or similar statistics) for each 50 bp bin. An odd
190 number of several, adjacent bins, i.e., the respective feature values (see below), is then
191 considered as input of the classifier composed of statistical models for the training process
192 as well as for making predictions. Conceptually, the classifier uses the information from
193 those bins to compute a-posteriori probabilities P_i that center bin i (i.e., the central bin
194 of those adjacent bins considered, cf. Figure 1) contains a peak summit. The number of
195 adjacent bins considered is determined from the median across cell types of the median
196 peak widths of a given TF in the individual training cell types.

197 2.3 Features

198 The set of features considered may be roughly classified by the source of information:
199 DNase-seq data, motif profiles, raw sequence, RNA-seq data. Here, we give a brief
200 overview of these features, while we provide a complete list of definitions of all features
201 in Supplementary Text S2.

202 The most informative features with regard to the challenge task are likely motif-based
203 and chromatin accessibility-based features. For obtaining a broad set of binding motifs
204 for each TF at hand, we combine motifs from databases with motifs obtained by de-novo
205 motif discovery from the challenge data. We retrieve PWM models of the TF at hand
206 from the databases HOCOMOCO (Kulakovskiy *et al.*, 2016) and DBcorrDB (Grau *et al.*,
207 2015a). We perform de-novo motif discovery with Dimont (Grau *et al.*, 2013) learning
208 PWM and LSlim(3) models (Keilwagen and Grau, 2015) on the “conservative” and
209 “relaxed” ChIP-seq peak files, and also based on the peak files obtained from DNase-seq
210 experiments. In addition, we obtain motifs from the epigram pipeline (Whitaker *et al.*,
211 2015), which are related to DNA methylation and histone marks of active promoters and
212 enhancers. For a specific combination of cell type and TF, we also consider motifs of a
213 set of “peer” motifs, which are determined from the literature (Factorbook, Wang *et al.*
214 (2012)) and by comparing the overlaps between the respective peak lists.

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

223 DNase-seq-based features are computed from the “fold-enrichment coverage” tracks
224 and DNase-seq peak files provided with the challenge data. These features quantify
225 short and long range chromatin accessibility, stability of the DNase signal in the region
226 of interest and across different cell types, and overlaps with DNase-seq peak regions.

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

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

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

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

250 2.4 Model & basic learning principle

251 We model the joint distribution of these different features by a simple product of in-
252 dependent densities or discrete distributions (Supplementary Text S3). Specifically, we
253 model numeric features (e.g., DNase-based statistics, motif scores, RNA-seq-based fea-
254 tures) by Gaussian densities, discrete, annotation-based features by independent bino-
255 mial distributions, and raw sequence by a homogeneous Markov model of order 3. All
256 distributions are in the exponential family and parameterized using their natural pa-
257 rameterization (Bishop, 2006; Keilwagen *et al.*, 2010), which allows for unconstrained
258 numerical optimization.

259 As learning principle, we use a weighted variant (Grau, 2010) of the discriminative
260 maximum conditional likelihood principle (Roos *et al.* (2005), Supplementary Text S3),
261 which is closely related to logistic regression but allows for making explicit assumptions
262 about the distribution of the underlying data.

263 2.5 Prediction schema

264 In the challenge, final predictions have been requested for 200 bp windows shifted by
265 50 bp along the genome, while the proposed classifier predicts a-posteriori probabilities
266 that the current center bin contains a peak summit. To yield the predictions requested,
267 we use these original prediction values (cf. section 2.2) to compute the probability that
268 the 200 bp window overlaps at least one predicted peak by at least 100 bp (Figure 1).
269 Assume that we already computed the a-posteriori probabilities P_i that center bin i
270 contains the summit of a ChIP-seq peak according to the trained model. Further assume
271 that for the current TF, a peak typically spans 5 bins in total, which corresponds to the
272 center bin, and two bins before and two bins after the center bin in our model (cf. regions
273 marked by lines in Figure 1). Putative peaks overlapping the current 200 bp window
274 starting at bin i are those with center bins at $i - 1$ to $i + 4$. Hence, the probability S_i
275 that this window overlaps a peak may be computed as the complementary probability
276 of the event that this window overlaps no predicted peaks, which in turn is just the
277 product of the complementary a-posteriori probabilities P_ℓ of these bins.

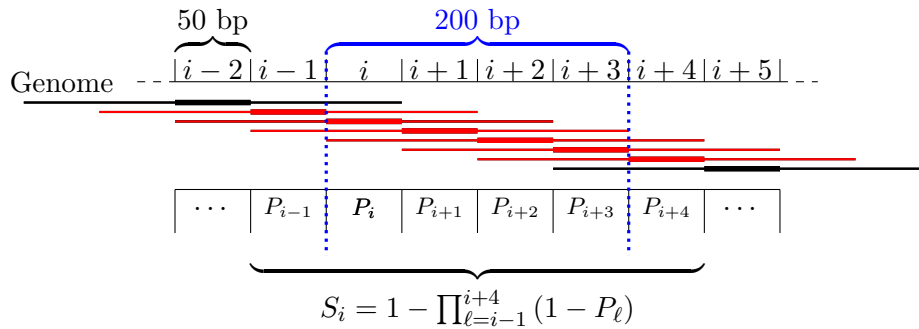


Figure 1: Schema for computing probabilities for 200 bp regions overlapping with predicted peaks spanning five bins in this example. Center bins are indicated by thick lines. Putative peaks are annotated with the probability P_i of being a true peak. All peaks marked in red overlap the region of interest (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.

278 2.6 Initial training data

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

286 Since we face a highly imbalanced classification problem with rather few ChIP-seq
 287 peaks compared with the large number of bins not covered by a peak, and since the
 288 inclusion of all such negative regions into the training set would lead to an unacceptable
 289 runtime, we decided to derive representative negative regions by three different sampling
 290 strategies. All sampling steps are performed stratified by chromosome.

291 First, we sample on each training chromosome 10 times as many negative regions
 292 (spanning an odd number of consecutive bins) as we find positive regions on that chro-
 293 mosome. Center bins are sampled uniformly over all bins not covered by a “relaxed”
 294 peak for the same cell type and TF.

295 Second, we over-sample negative regions with large DNase-seq median values similar
 296 to those of positive examples to yield a representative set of negative regions. This is
 297 especially important as these will be regions that are hard to classify using DNase-seq
 298 based features but are only lowly represented by the uniform sampling schema. The
 299 over-sampling is adjusted for by down-weighting the drawn negative examples to the
 300 corresponding frequency among all negative regions (see Supplementary Text S4).

301 Third, we sample negative regions from regions that are ChIP-seq positive for one of

302 the other cell types (if more than one training cell type exists for that TF), but do not
303 overlap a “relaxed peak” in the current cell type. These negative regions are weighted
304 such that the sum of their weights matches the rate of such regions among all putative
305 negative regions. This sampling schema is intended to foster learning cell type-specific
306 properties as opposed to general properties of the binding regions of the current TF. In
307 this case, we sample four times as many negative regions as we have positives.

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

316 2.7 Iterative training

317 The iterative training procedure is illustrated in Figure 2. Initially, we train a classifier
318 on the negative regions obtained from the sampling schemas explained above and all
319 positive regions. We then use this classifier to obtain a-posteriori probabilities P_i
320 for each bin i on training chromosomes. To limit the runtime required for this prediction
321 step, we restrict the prediction to chromosomes chr10 to chr14. These probabilities
322 are then used as input of the prediction schema (section Prediction schema) to yield
323 predictions for the 200 bp regions labeled based on the ChIP-seq training peaks. Given
324 these labels, we may distinguish prediction values of positive regions (label B=“bound”)
325 and negative regions (label U=“unbound”), while regions labeled as A=“ambiguous” are
326 ignored. To select additional negative regions that are likely false positive predictions,
327 we first collect the prediction scores of all positive regions (labeled as B) and determine
328 the corresponding 1% percentile. We then select from the negative regions (labeled as U)
329 all those with a predictions score larger than this 1% percentile, which are subsequently
330 added to the set of negative regions with a weight of 1 per region selected.

331 In the next iteration, we train a second classifier, again using all positive regions but
332 with negative regions complemented by these additional negative regions. Prediction
333 is then performed using both classifiers, where the predictions of the individual two
334 classifiers (or all previously trained classifiers in subsequent iterations) are averaged per
335 region. Again, regions labeled U with large prediction scores are identified and added to
336 the set of negative regions, which then serve as input of the following iteration. After five
337 rounds of training yielding five classifiers, the iterative training procedure is terminated.

338 2.8 Final prediction

339 The iterative training procedure is executed for all K training cell types with ChIP-seq
340 data for the TF of interest, which yields a total of $5 \cdot K$ classifiers. For the final prediction,
341 the prediction schema (section Prediction schema) is applied to all chromosomes and

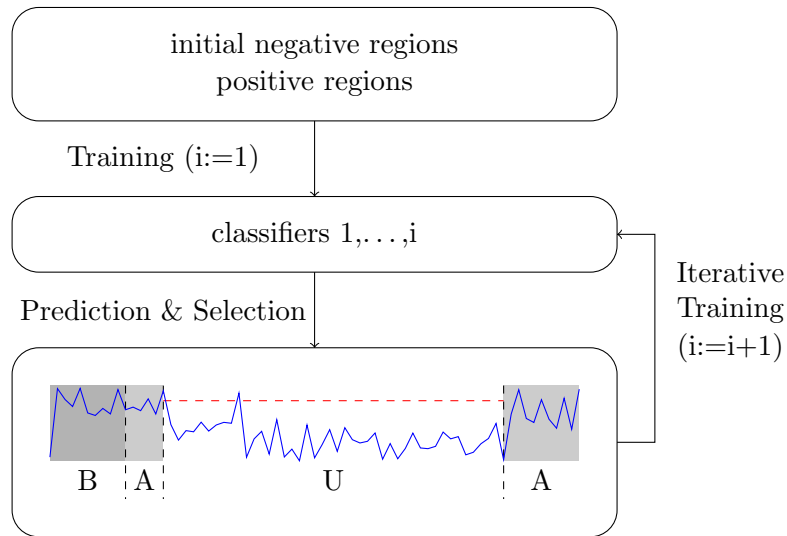


Figure 2: Iterative training procedure. Starting from an initial set of negative regions and the complete set of positive regions, a first classifier is trained, applied to the training data, and putative false positive (i.e., “unbound” regions with large prediction scores) are identified. In each of 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 5 classifiers trained in 5 cycles of the iterative training procedure.

342 each classifier. These predictions are finally averaged per 200 bp region to yield the final
343 prediction result.

344 **2.9 Catchitt: a streamlined open-source implementation**

345 Since the original challenge submission, we have re-implemented the basic approach with
346 the aim of making it more accessible for both, users and developers. Specifically, our
347 objectives were to implement a tool that i) is consolidated into a single runnable JAR
348 file to limit system requirements to a current Java installation only, ii) has an extensible
349 code base eliminating much of the experimental code of the challenge implementation,
350 iii) is applicable to data from individual cell types to reduce data-interdependencies, and
351 iv) may be executed on a standard compute server in acceptable runtime.

352 To achieve these aims, some parts of the methods have been simplified and stream-
353 lined. First, we consider only the most important chromatin accessibility and motif-based
354 features, which reduces runtime and memory consumption. Second, we implement an
355 accelerated motif scanning module that computes whole-genome score profiles even for
356 the complex LSlim models within a few hours. Third, we skip steps that jointly evalu-
357 ate data and/or feature files from multiple cell types. Specifically, we skip quantile
358 normalization of chromatin accessibility features (although normalization could be per-
359 formed externally, still), and we omit the sampling step depending on ChIP-seq data
360 for other cell types for determining initial negative regions. We call this implementa-
361 tion “Catchitt” comprising five modules for i) computing chromatin accessibility features
362 from DNase-seq or ATAC-seq data, ii) computing motif-based features, iii) deriving la-
363 bels from ChIP-seq peak lists, iv) performing iterative training given feature files and
364 labels, and v) predicting binding probabilities for genomic regions.

365 **2.10 Deriving peak lists**

366 For the additional primary cell types and tissues beyond those considered in the chal-
367 lenge, we further process final predictions to yield peak lists in narrowPeak format,
368 which are smaller and easier to handle than the genome-wide probability tracks with
369 50 bp resolution. To this end, we join contiguous stretches of regions with predicted
370 binding probability above a pre-defined threshold t into a common peak region. For
371 each region, we record the maximum probability p , and discard bordering regions with a
372 probability below $0.8 \cdot p$. The resulting regions are then annotated according to the nar-
373 rowPeak format with a “peak summit” at the center of the region yielding p , a “score”
374 of $-100 \cdot \log_{10}(1 - p)$, and a “signal value” equal to p . We generate “relaxed” peak
375 predictions using $t = 0.6$ and “conservative” peak prediction using $t = 0.8$.

376 **2.11 Availability**

377 The original challenge implementation has been developed using the open source Java
378 library Jstacs (Grau *et al.*, 2012) combined with custom Perl and bash scripts for data
379 extraction, conversion, and pipelining. The complete code accompanying the challenge

380 submission is, in accordance with the challenge guidelines, available from [https://www.
381 synapse.org/#!Synapse:syn8009967/wiki/412123](https://www.synapse.org/#!Synapse:syn8009967/wiki/412123) including a brief method writeup.
382 The Catchitt implementation is also based on the Jstacs library and is available as a
383 runnable JAR file at <http://jstacs.de/index.php/Catchitt>, where we also publish
384 the corresponding source code under GPL 3. Catchitt will be integrated into the Jstacs
385 library with its next release.

386 **3 Results**

387 During the ENCODE-DREAM challenge, a large number of approaches created by 40
388 international teams has been benchmarked on 13 cell type-specific ChIP-seq assays for
389 12 different TFs in human (Supplementary Figure S1). A set of 109 data sets for the
390 same (and additional) TFs in other cell types was provided for training. Training data
391 comprised cell type-specific DNase-seq data, cell type-specific RNA-seq data, genomic
392 sequence and annotations, and *in-silico* DNA shape predictions. In addition, cell type-
393 specific and TF-specific ChIP-seq data and derived labels were provided for training
394 chromosomes, while predictions were evaluated only on the remaining, held-out chromo-
395 somes chr1, chr8, and chr21 that were not provided with any of the ChIP-seq training
396 data. For 200 bp regions shifted by 50 bp, genome-wide predictions of the probability that
397 a specific region overlaps a ChIP-seq peak were requested from the participating teams.
398 Predictions were evaluated by i) the area under the ROC curve (AUC-ROC), ii) the area
399 under the precision-recall curve (AUC-PR), iii) recall at 10% FDR, and iv) recall at 50%
400 FDR on each of the 13 test data sets. These were aggregated per data set based on the
401 average, normalized rank earned for each of these measures in 10 bootstrap samples of
402 the held-out chromosomes, and a final ranking was obtained as the average of these rank
403 statistics (cf. <https://www.synapse.org/#!Synapse:syn6131484/wiki/405275>).

404 As a result of this ranking, the approach presented in this paper (team “J-Team”)
405 earned a shared first rank together with the approach created by team “Yuanfang Guan”.

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

416 **3.1 Impact of feature sets on prediction performance**

417 We use the prediction performance obtained by the proposed approach using all sets of
418 features (section Features), the iterative training procedure (section Iterative training),

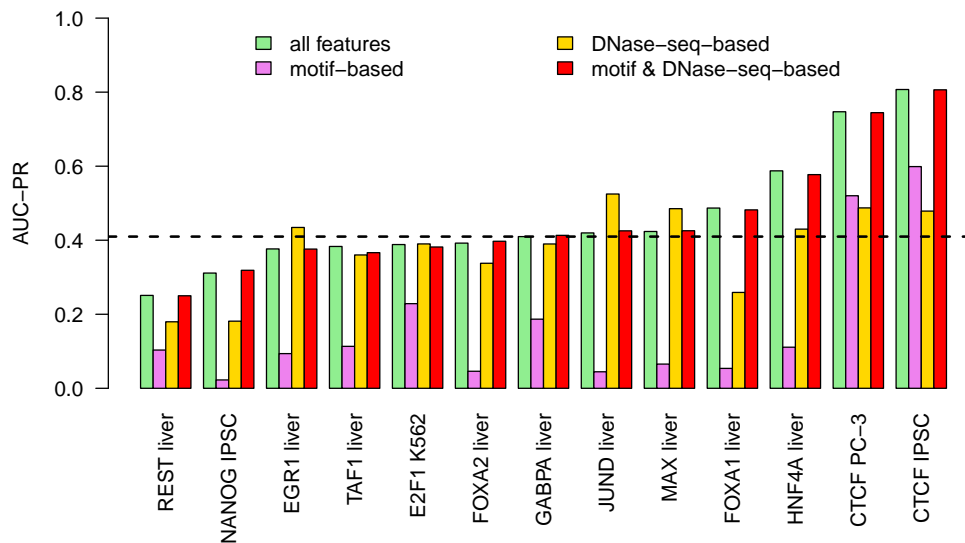


Figure 3: Across cell type performance. For each of the 13 combinations of TF and cell type within the test data, we compute the prediction performance (AUC-PR) on the held-out chromosomes of classifiers i) using all features considered, ii) using only motif-based features, iii) using only DNase-seq-based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

419 and the aggregation over all training cell types (section Prediction schema) as a base-
420 line for all further comparisons (Figure 3; “all features”). Throughout this manuscript,
421 we consider AUC-PR as the primary performance measure, since AUC-PR is more in-
422 formative about classification performance for heavily imbalanced classification prob-
423 lems (Keilwagen *et al.*, 2014; Saito and Rehmsmeier, 2015), and recall at the different
424 FDR levels is rather unstable since it corresponds to single points on the precision-recall
425 curve. AUC-PR values are computed using the R-package PRROC (Grau *et al.*, 2015b),
426 which has also been used in the ENCODE-DREAM challenge.

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

434 To analyze the contribution of selected features on the final prediction performance,
435 we systematically exclude sets of related features from the input data in training and
436 prediction. As a baseline, we measure AUC-PR for the classifier using all feature sets.

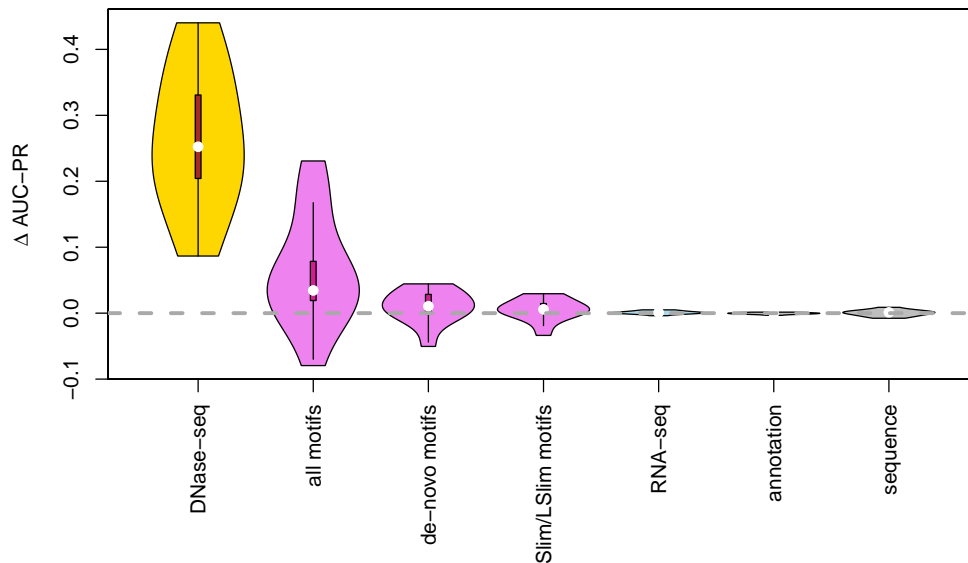


Figure 4: Importance of feature sets. We test the importance of related sets of features by excluding one set of features from the training data, measuring the performance (AUC-PR) of the resulting classifier, and subtracting this AUC-PR value from the corresponding value achieved by the classifier using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

437 In addition, we measure AUC-PR when excluding each individual feature set, where the
438 difference of these two AUC-PR values quantifies the improvement gained by including
439 the feature set (Figure 4).

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

443 Features based on motif scores (including de-novo discovered motifs and those from
444 databases) also contribute substantially to the final prediction performance. Here, we
445 observe large improvements for some TFs, namely 0.231 for CTCF in IPSC cells, 0.175
446 for CTCF in PC-3 cells, and 0.167 for FOXA1. By contrast, we observe a decrease in
447 prediction performance in case of JUND (-0.080) when including motif-based features.
448 For the remaining TFs, we find improvements of AUC-PR between 0.008 and 0.079.
449 We further consider two subsets of motifs, namely all motifs obtained by de-novo motif
450 discovery on the challenge data and all Slim/LSlim models capturing intra motif depen-
451 dencies. For motifs from de-novo motif discovery, we find an improvement for 9 of the
452 13 data sets and for Slim/LSlim model we find an improvement for 10 of the 13 data

453 sets. However, the absolute improvements (median of 0.011 and 0.006, respectively)
454 are rather small, possibly because i) motifs obtained by de-novo motif discovery might
455 be redundant to those found in databases and ii) intra motif dependencies and hetero-
456 geneities captured by Slim/LSlim models (Keilwagen and Grau, 2015) might be partly
457 covered by variations in the motifs from different sources.

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

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

469 Classifiers using only DNase-seq-based features yield a remarkable performance for
470 many of the TFs studied (Figure 3), which is lower than for the motif-based classifier
471 only for the two CTCF datasets. For some datasets (especially JUND but also EGR1,
472 MAX), we even observe that a classifier based on DNase-seq data alone outperforms the
473 classifier utilizing all features.

474 In case of JUND, the increase in performance when neglecting all non-DNase features
475 can likely be attributed to a strong adaptation of classifier parameters to either cell type-
476 specific binding motifs or cell type-specific co-binding with other TFs, because JUND is
477 the only dataset with an improved performance when excluding motif-based features as
478 discussed above. For all three TFs, we do find an improvement of prediction performance
479 if classifier parameters are trained on the training chromosomes of the test cell type (
480 “within cell type” case; Supplementary Figure S2).

481 Since DNase-seq-based and motif-based features appear to be the primary feature sets
482 affecting prediction performance, we finally study prediction performance of a classifier
483 using only these two feature sets. We observe that prediction performance using only
484 DNase-seq-based and motif-based features is largely identical to that of the classifier us-
485 ing all features (Figure 3), where we observe the largest loss in AUC-PR for TAF1 (0.017)
486 and the largest gain in AUC-PR for NANOG (0.007). We notice a similar behaviour for
487 the within cell type case (Supplementary Figure S2). As the left-out feature sets include
488 all RNA-seq-based features, this also has the consequence that one cell type-specific
489 assay (namely DNase-seq) is sufficient for predicting TF binding, which broadens the
490 scope of cell types with readily available experimental data that the proposed approach
491 may be applied to.

492 **3.2 Iterative training improves prediction performance**

493 As a second key aspect of the proposed approach, we investigate the impact of the
494 iterative training procedure on the final prediction performance. To this end, we compare

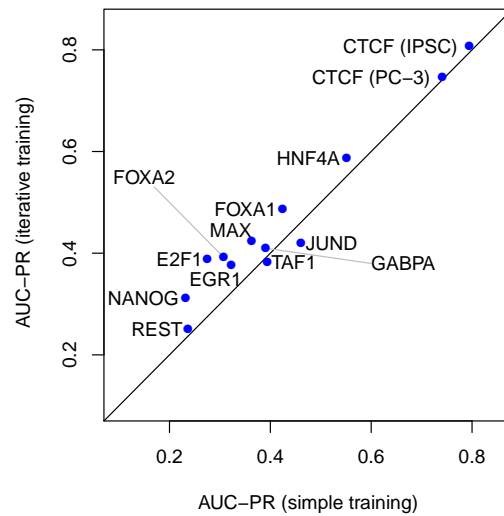


Figure 5: Relevance of the iterative training procedure. For each of the 13 test data sets, we compare the performance (AUC-PR) achieved by the (set of) classifier(s) trained on the initial negative regions (abscissa) with the performance achieved by averaging over all classifiers from the iterative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

495 for each TF the AUC-PR values obtained by averaging over the predictions all five
496 classifiers resulting from the iterative training procedure for all training cell types with
497 the AUC-PR values obtained by only averaging over the initially trained classifiers for all
498 training cell types, i.e., classifiers trained only on the initial training data (section “Initial
499 training data”).

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

516 **3.3 Averaging predictions improves over random selection of cell types**

517 For 9 of the 12 TFs considered, data for more than one training cell type is provided
518 with the challenge data. Hence, one central question might be the choice of the cell
519 type used for training and, subsequently, for making predictions for the test cell type.
520 However, the only cell type-specific experimental data available for making that choice
521 are DNase-seq and RNA-seq data, whereas similarity of cell types might depend on the
522 TF considered. Indeed, similarity measures derived from DNase-seq data (e.g., Jaccard
523 coefficients of overlapping DNase-seq peaks, correlation of profiles) or from RNA-seq
524 data (e.g., correlation of TPM values) showed to be non-informative with regard to the
525 similarity of TF binding regions in preliminary studies on the training cell types.

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

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

537 However, we also notice for almost all test data sets with multiple training cell types
538 (the only exception being CTCF for the PC-3 cell type) that the best prediction perfor-
539 mance achieved for one of the individual training cell types would have gained, in some
540 cases considerable, improvements over the proposed averaging procedure. Notably, the
541 variance of AUC-PR between the different training cell types is especially pronounced for
542 JUND, which supports the previous hypothesis that some features, for instance binding
543 motifs or co-binding of TFs, are highly cell type-specific for JUND. In general, deriving
544 informative measures of TF-specific cell type similarity based on cell type-specific assays
545 and, for instance, preliminary binding site predictions, would likely lead to a further
546 boost of the performance of computational approaches for predicting cell type-specific
547 TF binding.

548 **3.4 Creating a collection of cell type-specific TF binding tracks**

549 Having established that a single type of experimental assay, namely DNase-seq, is suffi-
550 cient for predicting cell type-specific TF binding with state-of-the-art accuracy, we may
551 now use the classifiers obtained on the training cell types and TFs for predictions on fur-
552 ther cell types. To this end, we download DNase-seq data for a collection of primary cell
553 types and tissues (see section Data), process these in the same manner as the original
554 challenge data and, subsequently, extract DNase-seq-dependent features (section Fea-
555 tures). We then applied the trained classifiers for all 31 TFs considered in the challenge
556 to these 22 DNase-seq feature sets to yield a total of 682 prediction tracks.

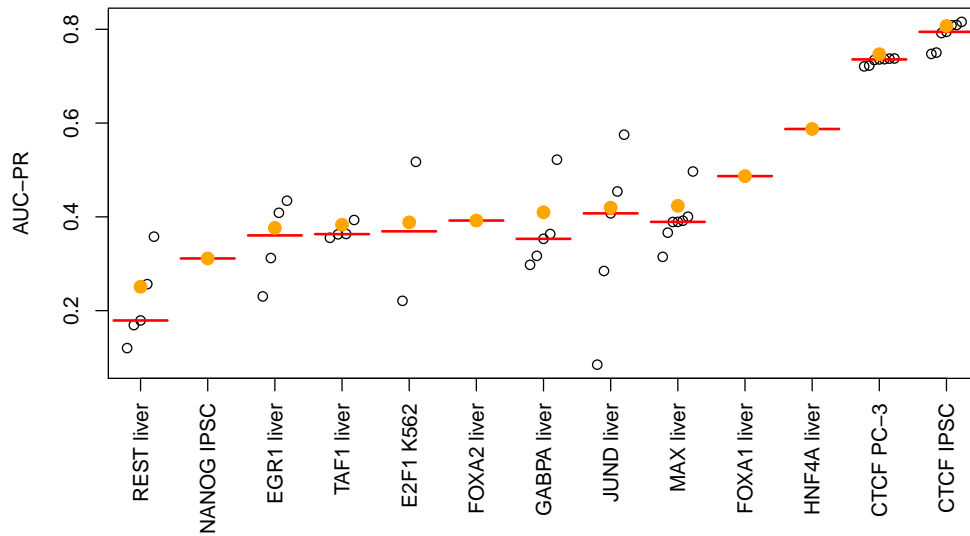


Figure 6: Performance of ensemble classifiers. For each of the 13 test data sets, we compare the performance (AUC-PR) of the individual classifiers trained on single cell types (open circles) to that of the ensemble classifier averaging over all classifiers trained on all training cell types (filled,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

557 For the selected cell types (Supplementary Table S3), only few cell type and TF-
558 specific ChIP-seq data are available (Supplementary Table S4). On the one hand, this
559 means that the predicted TF binding tracks provide valuable, novel information for the
560 collection of 31 TFs studied. On the other hand, this provides the opportunity to perform
561 benchmarking and sanity checks with regard to the predictions for the subset of these
562 TFs and cell types with corresponding ChIP-seq data available. For benchmarking,
563 we additionally obtain the “relaxed” and (where available) “conservative” peak files
564 from ENCODE and derive the associated labels (“bound”, “unbound”, “ambiguous”)
565 according to the procedure proposed for the ENCODE-DREAM challenge.

566 For CTCF with ChIP-seq peaks available for multiple cell types, we generally find a
567 prediction performance that is comparable to the performance observed on the challenge
568 data (cf. Supplementary Table S2). For these cell types, AUC-PR values (Supplemen-
569 tary Table S5) range between 0.7720 and 0.8197 if conservative and relaxed peaks are
570 available and if the donors match between the DNase-seq and ChIP-seq experiments,
571 while performance is slightly lower for non-matching donors (0.7322) and in case of
572 missing conservative peaks (0.7270). For JUN, MAX, and MYC, only relaxed peaks
573 are available from ENCODE due to missing replicates. Here, we find AUC-PR values
574 of 0.6310 for JUN, which is substantially larger than for the challenge data, 0.4004 for
575 MAX, which is slightly lower than for the challenge data, and 0.1989 for MYC, which
576 has not been among the test TFs in the challenge but obtained substantially better
577 performance in the leaderboard round.

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

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

591 Based on the predicted peak lists, we may also compare the predicted binding charac-
592 teristics of the different TFs across cell types. First, we inspect the number of predicted
593 peaks per TF and cell type (Supplementary Figure S4). We find a distinct group of
594 highly abundant TFs (CTCF, GATA3, SPI1, CEBPB, FOXA1, FOXA2, MAX), which
595 typically also show large numbers of peaks in the training data. Among these, we
596 find patterns of cell type specificity from the ubiquitously abundant CTCF to larg-
597 erly varying abundance for GATA3. The remainder of TFs obtains substantially lower
598 numbers of predicted peaks with similar patterns, e.g, for ATF7/ARID3A/NANOG or
599 EP300/TEAD4/JUND, where the latter group has been reported to co-bind in distal
600 enhancers (Xie *et al.*, 2013). Next, we study the stability of peak predictions, i.e., the

601 Jaccard coefficients of peaks predicted for each of the TFs in different cell types (Supple-
602 mentary Figure S5). Again, we find substantial variation among the TFs with GABPA,
603 CTCF, and REST having median Jaccard coefficients above 0.7. Notably, CTCF has
604 been one of the TFs with the largest number of predicted peaks (median 37 455), whereas
605 we observed an order of magnitude less predicted peaks for REST (median 3 364) and
606 GABPA (median 5 430). At the other end of the scale, we find indirectly binding TFs
607 like EP300, or TFs that are highly specific to cell types under-represented in our data
608 like NANOG (stem cells) and HNF4A (liver, kidney, intestines). Finally, we investigate
609 co-binding of TFs by computing the average Jaccard coefficient across cell types for each
610 pair of TFs (Supplementary Figure S6). Here, we observe distinct groups of co-occurring
611 TFs like CTCF/ZNF143 or FOXA1/FOXA2, which are known to interact *in-vivo* (Bai-
612 ley *et al.*, 2015; Ye *et al.*, 2016; Motalebipour *et al.*, 2009). In addition, we find a larger
613 cluster of TFs with substantial overlaps between their predicted peaks comprising YY1,
614 MAX, CREB1, MYC, E2F6, E2F1, and TAF1. As TAF1 (TATA-Box Binding Protein
615 Associated Factor 1) is associated with transcriptional initiation at the TATA box, one
616 explanation might be that binding sites of these TFs are enriched at core promoters.
617 Indeed, binding to proximal promoters has been reported for MYC/MAX (Guo *et al.*,
618 2014), CREB1 (Zhang *et al.*, 2005), YY1 (Li *et al.*, 2008), and E2F factors (Rabinovich
619 *et al.*, 2008).

620 **3.5 Streamlined Catchitt implementation yields competitive performance**

621 We finally compare Catchitt, the simplified implementation of the iterative training
622 approach combining chromatin accessibility and motif scores, to the challenge imple-
623 mentation using DNase-seq-based and motif-based features for the within cell type case.
624 To this end, we select five combinations of cell type and transcription factor spanning
625 the range of performance values observed in the challenge. Specifically, we consider
626 NANOG and TAF1, which obtained the lowest AUC-PR values (cf. Figure S2) for the
627 challenge implementation, CTCF in iPSC cells, which obtained the largest AUC-PR
628 value, and FOXA1 and HNF4A, which obtained medium AUC-PR values but profited
629 substantially from iterative training (cf. Figure S3). We summarize the results of this
630 comparison in Supplementary Table S8. Despite approximately ten-fold reduction in the
631 number of motifs considered and further simplifications (section 2.9), Catchitt still yields
632 competitive AUC-PR values. Ranking the Catchitt results within the original challenge
633 results, we find that performance achieved by Catchitt scores only two ranks lower than
634 the challenge implementation using DNase-seq-based and motif-based features. As be-
635 fore, we find a substantial improvement of prediction performance due to the iterative
636 training procedure.

637 **4 Discussion**

638 Predicting *in-vivo* binding sites of a TF of interest *in-silico* is still one of the central
639 challenges in regulatory genomics. A variety of tools and approaches for this purpose
640 have been created over the last years and, among these, the approach presented here is

641 not exceptional in many of its aspects. Specifically, it works on hand-crafted features
642 derived from genomic and experimental data, it considers TF binding motifs and chro-
643 matin accessibility as its major sources of information, and it uses supervised learning
644 related to logistic regression. Yet, this approach gained the best performance in the
645 ENCODE-DREAM challenge. Here, we focus on the impact of further, novel aspects of
646 the proposed approach on prediction performance.

647 With regard to the features considered, we find that motif-based and DNase-seq-
648 based features are pivotal for yielding a reasonable prediction performance for most
649 TFs, while other sequence-based, annotation-based, or RNA-seq-based features have
650 only marginal influence on the prediction result. In case of RNA-seq-based features,
651 however, more sophisticated features than those employed in our approach might have
652 a positive influence on prediction accuracy. In addition, DNA shape might also be
653 informative about true TF binding sites, although *in-silico* shape predictions provided
654 in ENCODE-DREAM are determined based on k-mers, and their influence might also be
655 captured by higher-order Markov models or Slim/LSlim models (Keilwagen and Grau,
656 2015) employed in the approach presented here.

657 Previous studies have shown that additional features like sequence conservation (Ku-
658 mar and Bucher, 2016; Liu *et al.*, 2017), histone marks (Pique-Regi *et al.*, 2011; Arvey
659 *et al.*, 2012; Gusmao *et al.*, 2014), or ChIP-seq data of co-factors (Kumar and Bucher,
660 2016) might also help to predict *in-vivo* TF binding. However, these were not allowed
661 to be used in the ENCODE-DREAM challenge and further experimental assays were
662 unavailable for the training cell types. Hence, we decided to also exclude such features
663 from the studies presented in this paper.

664 Two aspects of the presented approach, namely the iterative training procedure and
665 aggregation of predictions across training cell types, contribute substantially to the fi-
666 nal prediction performance. Both ideas might also be of relevance in related fields.
667 Specifically, the iterative training procedure provides a general schema applicable to
668 imbalanced classification problems, especially when these require sampling of negative
669 examples. In an abstract sense, the aggregation across training cell types corresponds
670 to favoring model averaging over model selection if good selection criteria are hard to
671 find or might yield highly varying results.

672 Despite its state-of-the-art performance proven in the ENCODE-DREAM challenges,
673 the approach presented here has important limitations. First, the large number of mo-
674 tifs (including those from *de-novo* motif discovery) and DNase-seq-based features lead to
675 high demands with regard to disk space but also runtime, which are likely beyond reach
676 for wet-lab biologists. Disk requirements could be reduced by computing features from
677 (smaller) raw files on demand. However, this would in turn increase running time con-
678 siderably. Hence, we chose to implement a simplified version of the approach presented
679 here in an open source software available at <http://jstacs.de/index.php/Catchitt>,
680 which only uses a combination of chromatin accessibility features and motif-based fea-
681 tures. In preliminary benchmarks (Supplementary Table S8), this implementation still
682 achieved competitive performance.

683 Second, the approach proposed here, like any of the other supervised approaches (Nataraj-
684 an *et al.*, 2012; Arvey *et al.*, 2012; Luo and Hartemink, 2012; Kähärä and Lähdesmäki,

685 2015; Kumar and Bucher, 2016; Quang and Xie, 2017; Liu *et al.*, 2017; Qin and Feng,
686 2017; Chen *et al.*, 2017), requires labeled training data for at least one cell type and the
687 TF of interest to make predictions for this TF in another cell type. While the latter
688 limitation is partly overcome by unsupervised approaches (Pique-Regi *et al.*, 2011; Sher-
689 wood *et al.*, 2014; Gusmao *et al.*, 2014; Raj *et al.*, 2015; Jankowski *et al.*, 2016), this
690 typically comes at the cost of reduced prediction accuracy (Kähärä and Lähdesmäki,
691 2015; Liu *et al.*, 2017).

692 We also provide a large collection of 682 predicted peak files for 31 TFs using 22
693 DNase-seq data sets for primary cell types and tissues. Benchmarks based on the limited
694 number of available ChIP-seq data indicate that prediction performance on these cell
695 types is comparable to that achieved in the ENCODE-DREAM challenge, where absolute
696 values of AUC-PR measuring prediction accuracy vary greatly between different TFs.
697 For the wide majority of these combinations of TF and cell type, no experimental data
698 about cell type-specific TF binding is available so far, which renders these predictions
699 a valuable resource for questions related to regulatory genomics in these primary cell
700 types and tissues. Preliminary studies raise our confidence that the predicted peak files
701 may indeed help to solve biological questions related to these cell types and TFs.

702 Acknowledgements

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

708 References

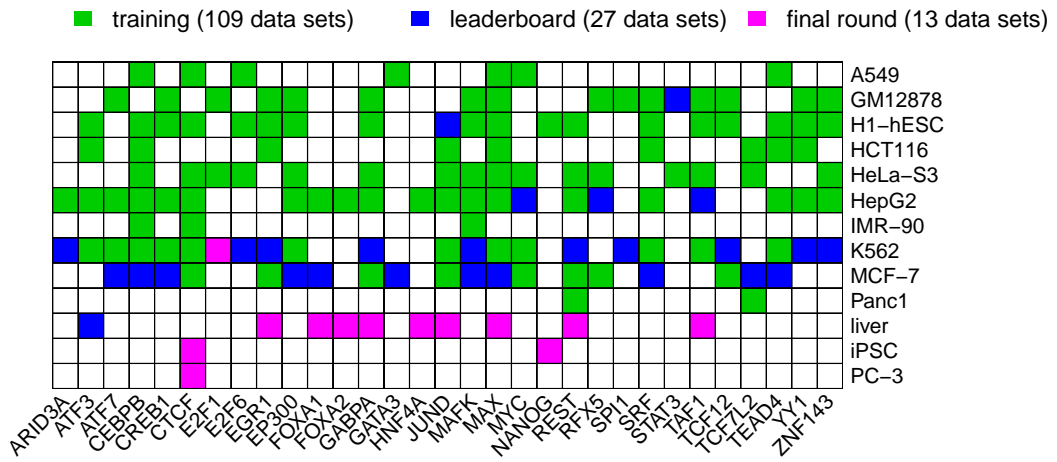
- 709 Arvey, A., Agius, P., Noble, W. S., and Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific
710 transcription factor binding. *Genome Research*, **22**(9), 1723–1734.
- 711 Bailey, S. D., Zhang, X., Desai, K., Aid, M., Corradin, O., Cowper-Sallari, R., Akhtar-Zaidi, B., Scacheri, P. C., Haibe-
712 Kains, B., and Lupien, M. (2015). ZNF143 provides sequence specificity to secure chromatin interactions at gene
713 promoters. **2**, 6186 EP –.
- 714 Berg, O. G. and von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins: Statistical-mechanical
715 theory and application to operators and promoters. *Journal of Molecular Biology*, **193**(4), 723 – 743.
- 716 Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New
717 York, 1st edition.
- 718 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin
719 for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat*
720 *Meth*, **10**(12), 1213–1218.
- 721 Bulyk, M. L. (2003). Computational prediction of transcription-factor binding site locations. *Genome Biology*, **5**(1),
722 201.
- 723 Chen, X., Hoffman, M. M., Bilmes, J. A., Hesselberth, J. R., and Noble, W. S. (2010). A dynamic Bayesian network for
724 identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics*, **26**(12), i334–i342.

- 725 Chen, X., Yu, B., Carriero, N., Silva, C., and Bonneau, R. (2017). Mocap: large-scale inference of transcription factor
726 binding sites from chromatin accessibility. *Nucleic Acids Research*, **45**(8), 4315–4329.
- 727 Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the 13th*
728 *International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann.
- 729 Galas, D. J. and Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding
730 specificity. *Nucleic Acids Research*, **5**(9), 3157–3170.
- 731 Grau, J. (2010). *Discriminative Bayesian principles for predicting sequence signals of gene regulation*. Ph.D. thesis,
732 Martin Luther University Halle–Wittenberg.
- 733 Grau, J., Keilwagen, J., Gohr, A., Haldemann, B., Posch, S., and Grosse, I. (2012). Jstacs: A Java framework for
734 statistical analysis and classification of biological sequences. *Journal of Machine Learning Research*, **13**(Jun), 1967–
735 1971.
- 736 Grau, J., Posch, S., Grosse, I., and Keilwagen, J. (2013). A general approach for discriminative de novo motif discovery
737 from high-throughput data. *Nucleic Acids Research*, **41**(21), e197.
- 738 Grau, J., Grosse, I., Posch, S., and Keilwagen, J. (2015a). Motif clustering with implications for transcription factor
739 interactions. In *German Conference on Bioinformatics*, volume 3 of *PeerJ Preprints*, page e1601.
- 740 Grau, J., Grosse, I., and Keilwagen, J. (2015b). PRROC: computing and visualizing precision-recall and receiver operating
741 characteristic curves in R. *Bioinformatics*, **31**(15), 2595–2597.
- 742 Guo, J., Li, T., Schipper, J., Nilson, K. A., Fordjour, F. K., Cooper, J. J., Gordân, R., and Price, D. H. (2014). Sequence
743 specificity incompletely defines the genome-wide occupancy of Myc. *Genome Biology*, **15**(10), 482.
- 744 Gusmao, E. G., Dieterich, C., Zenke, M., and Costa, I. G. (2014). Detection of active transcription factor binding sites
745 with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, **30**(22), 3143–3151.
- 746 Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa,
747 A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes,
748 G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N.,
749 Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D.,
750 Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T. J. (2012). GENCODE: The
751 reference human genome annotation for the ENCODE project. *Genome Research*, **22**(9), 1760–1774.
- 752 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass,
753 C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required
754 for macrophage and B cell identities. *Molecular Cell*, **38**(4), 576–589.
- 755 Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn,
756 M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions
757 in vivo by digital genomic footprinting. *Nat Meth*, **6**(4), 283–289.
- 758 Jankowski, A., Tiuryn, J., and Prabhakar, S. (2016). Romulus: robust multi-state identification of transcription factor
759 binding sites from DNase-seq data. *Bioinformatics*, **32**(16), 2419–2426.
- 760 Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA
761 interactions. *Science*, **316**(5830), 1497–1502.
- 762 Kähärä, J. and Lähdesmäki, H. (2015). BinDNase: a discriminatory approach for transcription factor binding prediction
763 using DNase I hypersensitivity data. *Bioinformatics*, **31**(17), 2852–2859.
- 764 Keilwagen, J. and Grau, J. (2015). Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids*
765 *Research*.
- 766 Keilwagen, J., Grau, J., Posch, S., and Grosse, I. (2010). Apples and oranges: avoiding different priors in Bayesian DNA
767 sequence analysis. *BMC Bioinformatics*, **11**(1), 149.
- 768 Keilwagen, J., Grosse, I., and Grau, J. (2014). Area under precision-recall curves for weighted and unweighted data.
769 *PLoS ONE*, **9**(3), e92209.
- 770 Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-
771 binding proteins. *Nat Biotech*, **26**(12), 1351–1359.

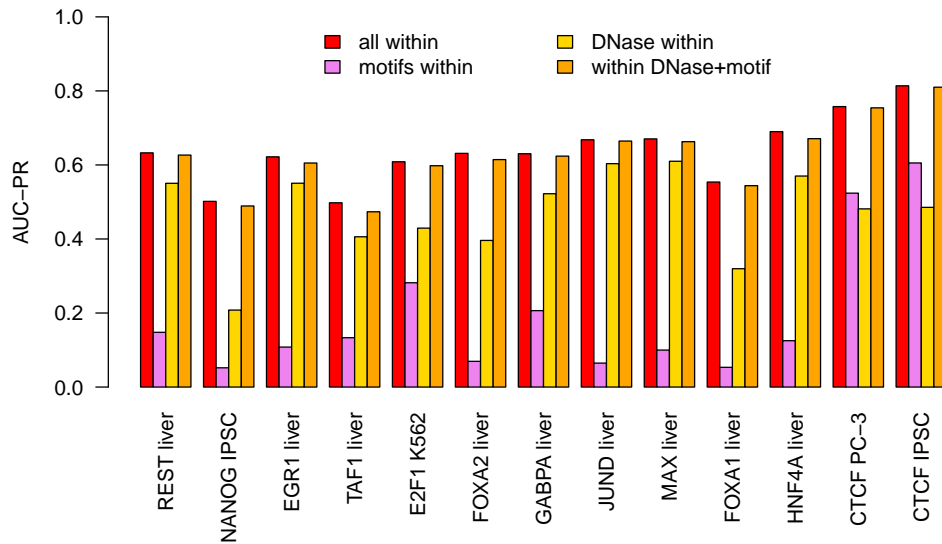
- 772 Kheradpour, P. and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF
773 binding experiments. *Nucleic Acids Research*, **42**(5), 2976–2987.
- 774 Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Soboleva, A. V., Kasianov, A. S., Ashoor, H., Ba-alawi, W., Bajic,
775 V. B., Medvedeva, Y. A., Kolpakov, F. A., and Makeev, V. J. (2016). HOCOMOCO: expansion and enhancement of
776 the collection of transcription factor binding sites models. *Nucleic Acids Research*, **44**(D1), D116–D125.
- 777 Kumar, S. and Bucher, P. (2016). Predicting transcription factor site occupancy using DNA sequence intrinsic and
778 cell-type specific chromatin features. *BMC Bioinformatics*, **17**(1), S4.
- 779 Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., and Carey, V. (2013).
780 Software for computing and annotating genomic ranges. *PLoS Computational Biology*, **9**.
- 781 Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference
782 genome. *BMC Bioinformatics*, **12**(1), 323.
- 783 Li, H., Liu, H., Wang, Z., Liu, X., Guo, L., Huang, L., Gao, L., McNutt, M. A., and Li, G. (2008). The role of
784 transcription factors Sp1 and YY1 in proximal promoter region in initiation of transcription of the mu opioid receptor
785 gene in human lymphocytes. *Journal of Cellular Biochemistry*, **104**(1), 237–250.
- 786 Liu, S., Zibetti, C., Wan, J., Wang, G., Blackshaw, S., and Qian, J. (2017). Assessing the model transferability for
787 prediction of transcription factor binding sites based on chromatin accessibility. *BMC Bioinformatics*, **18**(1), 355.
- 788 Luo, K. and Hartemink, A. J. (2012). Using DNase digestion data to accurately identify transcription factor binding
789 sites. In *Pacific Symposium on Biocomputing*, pages 80–91. World Scientific.
- 790 Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R.,
791 Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2016). Jaspur 2016: a major expansion
792 and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **44**(D1),
793 D110–D115.
- 794 Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull,
795 M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006).
796 TRANSFAC(R) and its module TRANSCOMP(R): transcriptional gene regulation in eukaryotes. *Nucleic Acids*
797 *Research*, **34**(suppl_1), D108–110.
- 798 Motallebipour, M., Ameer, A., Reddy Bysani, M. S., Patra, K., Wallerman, O., Mangion, J., Barker, M. A., McKernan,
799 K. J., Komorowski, J., and Wadelius, C. (2009). Differential binding and co-binding pattern of FOXA1 and FOXA3
800 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biology*, **10**(11), R129.
- 801 Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E., and Ohler, U. (2012). Predicting cell-type-specific
802 gene expression from regions of open chromatin. *Genome Research*, **22**(9), 1711–1722.
- 803 Newburger, D. E. and Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on
804 protein-DNA interactions. *Nucleic Acids Research*, **37**(suppl 1), D77–D82.
- 805 Piper, J., Elze, M. C., Cauchy, P., Cockerill, P. N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the
806 accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Research*, **41**(21), e201.
- 807 Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of
808 transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, **21**(3), 447–455.
- 809 Qin, Q. and Feng, J. (2017). Imputation for transcription factor binding predictions based on deep learning. *PLOS*
810 *Computational Biology*, **13**(2), 1–20.
- 811 Quang, D. and Xie, X. (2017). FactorNet: a deep learning framework for predicting cell type specific transcription factor
812 binding from nucleotide-resolution sequential data. *bioRxiv*.
- 813 Rabinovich, A., Jin, V. X., Rabinovich, R., Xu, X., and Farnham, P. J. (2008). E2f in vivo binding specificity: Comparison
814 of consensus versus nonconsensus binding sites. *Genome Research*, **18**(11), 1763–1777.
- 815 Raj, A., Shim, H., Gilad, Y., Pritchard, J. K., and Stephens, M. (2015). msCentipede: Modeling heterogeneity across
816 genomic sites and replicates improves accuracy in the inference of transcription factor binding. *PLOS ONE*, **10**(9),
817 1–15.
- 818 Roos, T., Wettig, H., Grünwald, P., Myllymäki, P., and Tirri, H. (2005). On discriminative Bayesian network classifiers
819 and logistic regression. *Machine Learning*, **59**(3), 267–296.

- 820 Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating
821 binary classifiers on imbalanced datasets. *PLOS ONE*, **10**(3), 1–21.
- 822 Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J. K., Ebert, P., Nordström, K.,
823 Barann, M., Sinha, A., Fröhler, S., Xiong, J., Dehghani Amirabad, A., Behjati Ardakani, F., Hutter, B., Zipprich,
824 G., Felder, B., Eils, J., Brors, B., Chen, W., Hengstler, J. G., Hamann, A., Lengauer, T., Rosenstiel, P., Walter, J.,
825 and Schulz, M. H. (2017). Combining transcription factor binding affinities with open-chromatin data for accurate
826 gene expression prediction. *Nucleic Acids Research*, **45**(1), 54–66.
- 827 Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., Karun, V., Jaakkola, T., and
828 Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase
829 profile magnitude and shape. *Nat Biotech*, **32**(2), 171–178.
- 830 Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, **12**, 505–519.
- 831 Stormo, G. D. and Fields, D. S. (1998). Specificity, free energy and information content in protein–DNA interactions.
832 *Trends in Biochemical Sciences*, **23**(3), 109 – 113.
- 833 Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng,
834 Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., and Weng, Z. (2012). Sequence features
835 and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*,
836 **22**(9), 1798–1812.
- 837 Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert,
838 S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E.,
839 Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J., Bouget, F.-Y., Ratsch,
840 G., Larrondo, L. F., Ecker, J. R., and Hughes, T. R. (2014). Determination and inference of eukaryotic transcription
841 factor sequence specificity. *Cell*, **158**(6), 1431 – 1443.
- 842 Whitaker, J. W., Chen, Z., and Wang, W. (2015). Predicting the human epigenome from DNA motifs. *Nat Meth*, **12**(3),
843 265–272.
- 844 Wu, J., Smith, L. T., Plass, C., and Huang, T. H.-M. (2006). ChIP-chip Comes of Age for Genome-wide Functional
845 Analysis. *Cancer Research*, **66**(14), 6899–6902.
- 846 Xie, D., Boyle, A. P., Wu, L., Zhai, J., Kawli, T., and Snyder, M. (2013). Dynamic trans-acting factor colocalization in
847 human cells. *Cell*, **155**(3), 713 – 724.
- 848 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
849 is involved in CTCF-mediated chromatin interactions by cooperation with cohesin and other partners. *Molecular*
850 *Biology*, **50**(3), 431–437.
- 851 Zhang, X., Odom, D. T., Koo, S.-H., Conkright, M. D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E.,
852 Jacobsen, E., Kadam, S., Ecker, J. R., Emerson, B., Hogenesch, J. B., Unterman, T., Young, R. A., and Montminy,
853 M. (2005). Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target
854 gene activation in human tissues. *Proceedings of the National Academy of Sciences of the United States of America*,
855 **102**(12), 4459–4464.

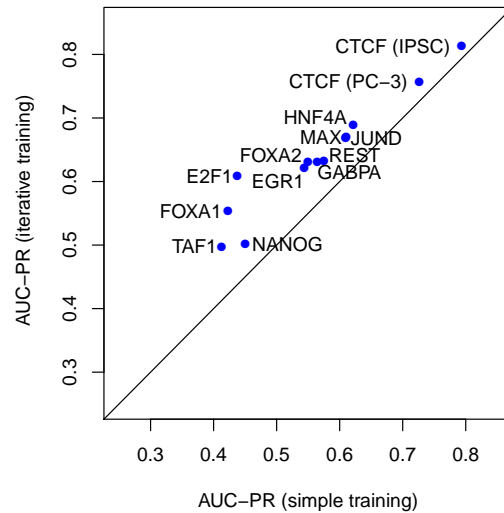
856 **Supplementary Tables and Figures**



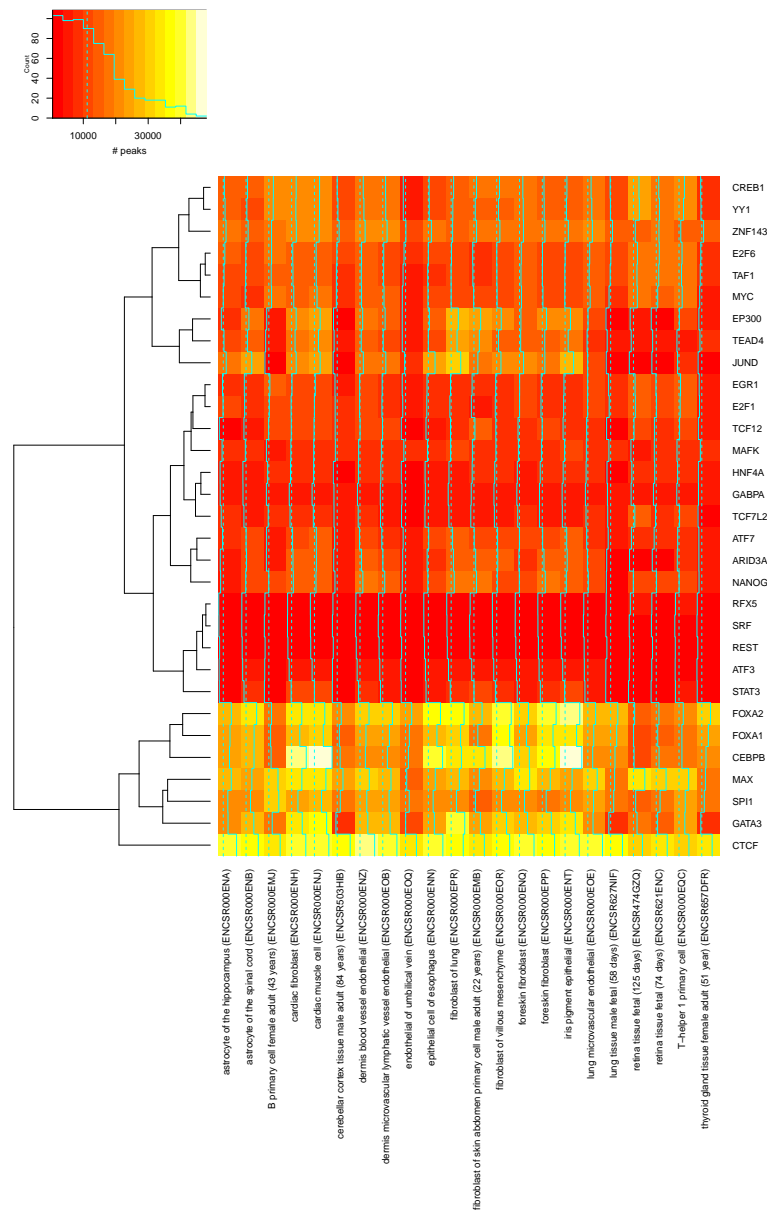
Supplementary Figure S1: Overview of the combinations of cell type and TF in the ENCODE-DREAM training, leaderboard, and final round sets.



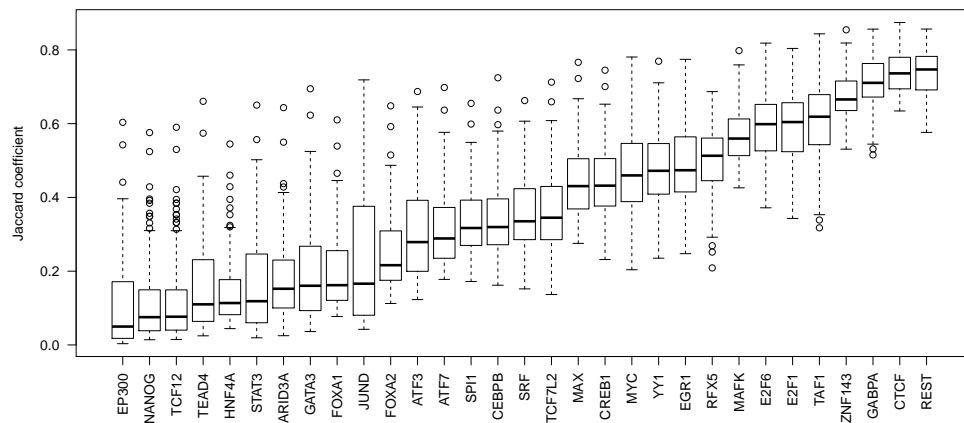
Supplementary Figure S2: Within cell type performance. For each of the 13 combinations of TF and cell type within the test data, we compute the prediction performance (AUC-PR) on the held-out chromosomes of classifiers i) using all features considered, ii) using only motif-based features, iii) using 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
CENTIPÈDE (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 CENTIPÈDE, 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”, “motif & 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	ENC SR000DSU	ENCF F312HCK	ENCDO224AAA	astrocyte of the spinal cord	relaxed
CTCF	ENC SR000DSU	ENCF F787GLH	ENCDO224AAA	astrocyte of the spinal cord	conservative
CTCF	ENC SR000DTI	ENCF F266GGD	ENCDO330AAA	cardiac muscle cell	relaxed
CTCF	ENC SR000DTI	ENCF F386NQE	ENCDO330AAA	cardiac muscle cell	conservative
CTCF	ENC SR000DTR	ENCF F528VFN	ENCDO104AAA	epithelial cell of esophagus	relaxed
CTCF	ENC SR000DTR	ENCF F373BXG	ENCDO104AAA	epithelial cell of esophagus	conservative
CTCF	ENC SR000DPM	ENCF F681OWQ	ENCDO001AAA	fibroblast of lung	relaxed
CTCF	ENC SR000DPM	ENCF F138PXI	ENCDO001AAA	fibroblast of lung	conservative
CTCF	ENC SR000DVQ	ENCF F738CXX	ENCDO253AAA	fibroblast of villous mesenchyme	relaxed
CTCF	ENC SR000DVQ	ENCF F199ZDU	ENCDO253AAA	fibroblast of villous mesenchyme	conservative
CTCF	ENC SR000DWQ	ENCF F337WIE	ENCDO191CQJ	foreskin fibroblast	relaxed
CTCF	ENC SR000DWQ	ENCF F275AVH	ENCDO191CQJ	foreskin fibroblast	conservative
CTCF	ENC SR000DLW	ENCF F002DBA	ENCDO000AAS	endothelial cell of umbilical vein	relaxed
CTCF	ENC SR000DWY	ENCF F002DDO	ENCDO269AAA	fibroblast of lung	relaxed
CTCF	ENC SR000DUH	ENCF F002DCY	ENCDO232AAA	foreskin fibroblast	relaxed
CTCF	ENC SR000DQI	ENCF F649IRT	ENCDO000AAG	foreskin fibroblast	relaxed
JUN	ENC SR000EFA	ENCF F002CVC	ENCDO000AAS	endothelial cell of umbilical vein	relaxed
MAX	ENC SR000EEZ	ENCF F002CVE	ENCDO000AAS	endothelial cell of umbilical vein	relaxed
MYC	ENC SR000DLU	ENCF F002DAZ	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.

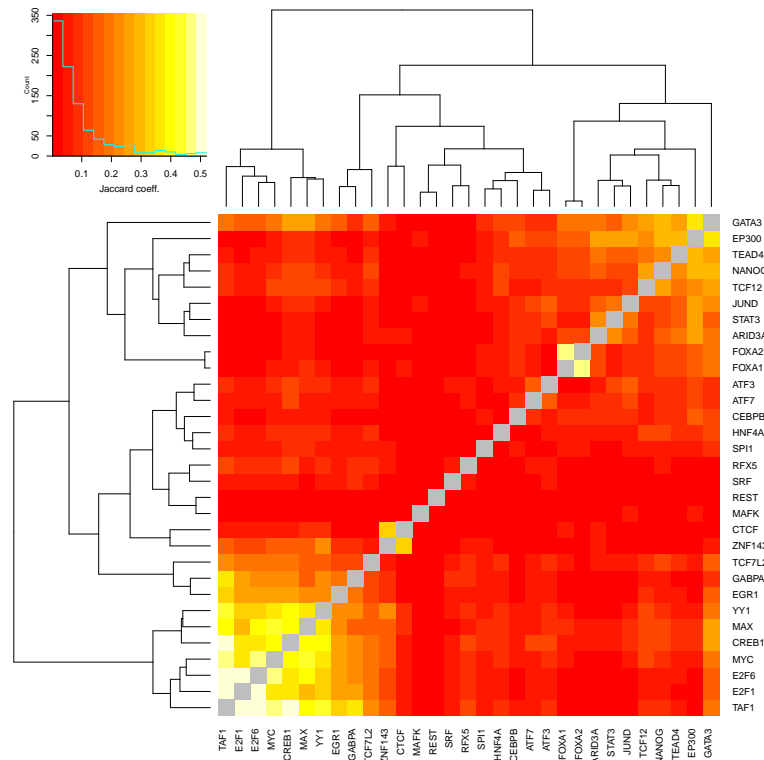
TF	DNase ID	Experiment ID	Matching donor	AUC-ROC	AUC-PR	recall @ 10% FDR	recall @ 50% FDR
CTCF	ENCSR000ENB	ENCSR000DSU	yes	0.9953	0.7895	0.5603	0.8240
CTCF	ENCSR000ENJ	ENCSR000DTI	yes	0.9950	0.8197	0.6316	0.8486
CTCF	ENCSR000ENN	ENCSR000DTR	yes	0.9932	0.7788	0.5621	0.8098
CTCF	ENCSR000EPP	ENCSR000DWQ	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. We find that many of the cell type-specific predictions for CTCF are more similar to the ChIP-seq peaks determined for “endothelial cells of umbilical vein” than to those of their cell type of origin according to the DNase-seq data. One reason might be that only for this experiment (ENCSR000DLW), peaks have not been called using the uniform ENCODE pipeline including SPP (Kharchenko *et al.*, 2008), but by another, “unknown” software. However, if we in turn ask for each experimentally determined peak list, which of the predicted peak lists is the most similar one, this picture becomes more encouraging. For 7 of the 8 cell types with matching donor between ChIP-seq and DNase-seq data, the most similar prediction is obtained for the true cell type, while in one case (“fibroblast of lung”), the most similar cell type is “foreskin fibroblast”.

astrocyte of the spinal cord (ENCSR000DSU)	0.5465	0.5555	0.5675	0.5659	0.5725	0.5537
cardiac muscle cell (ENCSR000DTI)	0.5852 (y)	0.5287	0.5408	0.5540	0.5517	0.5395
endothelial cell of umbilical vein (ENCSR000DLW)	0.5309	0.6849	0.6730	0.6924	0.6888	0.6664
epithelial cell of esophagus (ENCSR000DTR)	0.7090 (y)	0.5576 (y)	0.5123	0.5147	0.5214	0.5014
fibroblast of lung (ENCSR000DPM)	0.5361	0.5230	0.5257 (y)	0.5338	0.5383	0.5290
fibroblast of lung (ENCSR000DWM)	0.6451	0.5372	0.6391 (y)	0.6497	0.6530	0.6501
fibroblast of villous mesenchyme (ENCSR000DVQ)	0.5818	0.5977	0.5737	0.6084 (y)	0.5938	0.5758
foreskin fibroblast (ENCSR000DQI)	0.5752	0.5549	0.5723	0.6787	0.5920	0.5850
foreskin fibroblast (ENCSR000DUH)	0.6714	0.6447	0.6637	0.6787	0.6934 (y)	0.6770
foreskin fibroblast (ENCSR000DWQ)	0.5400	0.5086	0.5301	0.5418	0.5469	0.5680 (y)
astrocyte of the spinal cord (ENCSR000ENB)	0.5465	0.5555	0.5675	0.5659	0.5725	0.5537
cardiac muscle cell (ENCSR000ENJ)	0.5732 (y)	0.5287	0.5408	0.5540	0.5517	0.5395
endothelial cell of umbilical vein (ENCSR000EQ)	0.5309	0.6849	0.6730	0.6924	0.6888	0.6664
epithelial cell of esophagus (ENCSR000ENN)	0.7090 (y)	0.5576 (y)	0.5123	0.5147	0.5214	0.5014
fibroblast of lung (ENCSR000EPR)	0.5361	0.5230	0.5257 (y)	0.5338	0.5383	0.5290
fibroblast of lung (ENCSR000EOR)	0.6478	0.5372	0.6391 (y)	0.6497	0.6530	0.6501
fibroblast of villous mesenchyme (ENCSR000EOR)	0.5977	0.5663	0.5737	0.6084 (y)	0.5938	0.5758
foreskin fibroblast (ENCSR000ENQ)	0.5752	0.5549	0.5723	0.6787	0.5920	0.5850
foreskin fibroblast (ENCSR000ENR)	0.6714	0.6447	0.6637	0.6787	0.6934 (y)	0.6770
foreskin fibroblast (ENCSR000EPP)	0.5400	0.5086	0.5301	0.5418	0.5469	0.5680 (y)

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.

TF	cell type	DNase+motif			Catchitt			
		AUC-PR	Rank	#motifs	AUC-PR	Rank	#motifs	first iteration
CTCF	IPSC	0.810	4	177	0.776	6	13	0.760
NANOG	IPSC	0.489	2	127	0.404	4	13	0.366
FOXA1	liver	0.544	2	121	0.465	4	12	0.435
HNF4A	liver	0.671	4	123	0.615	6	10	0.597
TAF1	liver	0.473	4	160	0.412	6	12	0.400

Supplementary Table S8: Benchmark of the simplified open source implementation (Catchitt) of the presented approach compared with the challenge implementation using only DNase-based and motif-based features for the within cell type case (cf. Supplementary Figure S2). For each of the TFs considered, we report AUC-PR achieved by the challenge implementation using only DNase-based and motif-based features (“DNase+motif”), the open source Catchitt implementation. For Catchitt, we additionally consider using only the classifier of the first iteration (in analogy to the comparison in Supplementary Figure S3). We also list the number of motifs utilized in the respective runs for a specific TFs. For the Catchitt runs, we deliberately limited the number of motifs considered to approximate a real-world application of the software. We finally report the ranks among the challenge participants according to the results available at <https://www.synapse.org/#!Synapse:syn6131484/wiki/412905>.

857 **Supplementary Methods**

858 **Supplementary Text S1 – Tools for predicting in-vivo binding regions**

859 Most approaches (e.g., Pique-Regi *et al.* (2011); Natarajan *et al.* (2012); Piper *et al.*
860 (2013); Gusmao *et al.* (2014); Chen *et al.* (2017)) use binding motifs represented as
861 position weight matrix (PWM) models that have been obtained from databases like
862 TRANSFAC (Matys *et al.*, 2006), Jaspar (Mathelier *et al.*, 2016), UniProbe (Newburger
863 and Bulyk, 2009) or CisBP (Weirauch *et al.*, 2014), or from motif collections like Factor-
864 book (Wang *et al.*, 2012), the ENCODE-motif collection (Kheradpour and Kellis, 2014),
865 or Homer (Heinz *et al.*, 2010), while some perform de-novo motif discovery based on
866 k-mers (Arvey *et al.*, 2012) or as part of convolutional neural networks (Quang and Xie,
867 2017; Qin and Feng, 2017). Irrespective of the source of the motifs considered, three
868 general schemas are have been established for combining motif predictions with chro-
869 matin accessibility data. First, motif matches (i.e., predicted binding sites) may be used
870 as prior information and combined with DNase-seq data to distinguish functional from
871 non-functional binding sites (e.g., Pique-Regi *et al.* (2011); Jankowski *et al.* (2016); Raj
872 *et al.* (2015)), Second, TF footprints may be first identified from DNase-seq data and
873 then annotated with specific TFs based on motif matches afterwards (Gusmao *et al.*,
874 2014). Third, both sources of information are combined in a holistic approach (Quang
875 and Xie, 2017; Qin and Feng, 2017). DNase-seq (and ATAC-seq) data are employed
876 in different ways by existing approaches including i) binning of chromatin accessibility
877 statistics in larger genomic regions around putative binding sites (Luo and Hartemink,
878 2012), ii) association of chromatin accessibility with specific genes (Schmidt *et al.*, 2017),
879 or iii) high-resolution maps of DNase cut sites (Sherwood *et al.*, 2014; Raj *et al.*, 2015),
880 which may additionally be considered separately for each DNA strand (Piper *et al.*,
881 2013). On the methodological level, approaches either follow a supervised approach
882 based on training examples labeled as “bound” or “unbound”, typically derived from
883 TF ChIP-seq data (e.g., Arvey *et al.* (2012); Luo and Hartemink (2012); Kähärä and
884 Lähdesmäki (2015); Liu *et al.* (2017)), or an unsupervised approach clustering regions
885 into “bound” and “unbound” based on their experimental properties (e.g., DNase-seq
886 data or histone modifications (Pique-Regi *et al.*, 2011; Sherwood *et al.*, 2014; Gusmao
887 *et al.*, 2014)), while others base their predictions on statistical tests (Piper *et al.*, 2013)
888 or scores related to binding affinity predictions (Schmidt *et al.*, 2017). Supervised ap-
889 proaches use a variety of methods like support vector machines (Arvey *et al.*, 2012;
890 Kumar and Bucher, 2016), (sparse) logistic regression (Natarajan *et al.*, 2012; Luo and
891 Hartemink, 2012; Kähärä and Lähdesmäki, 2015; Chen *et al.*, 2017), random forests (Liu
892 *et al.*, 2017), or neural networks adapted by deep learning (Quang and Xie, 2017; Qin
893 and Feng, 2017). Unsupervised approaches use hierarchical mixture models (Pique-Regi
894 *et al.*, 2011), hierarchical multi-scale models (Raj *et al.*, 2015), hidden Markov mod-
895 els (Gusmao *et al.*, 2014), or other probabilistic models (Sherwood *et al.*, 2014). In some
896 approaches, sequence-based features besides motif matches (Kumar and Bucher, 2016;
897 Gusmao *et al.*, 2014; Chen *et al.*, 2017), sequence conservation (Kumar and Bucher,
898 2016; Liu *et al.*, 2017; Chen *et al.*, 2017), or additional experimental data like histone

899 modification (Pique-Regi *et al.*, 2011; Arvey *et al.*, 2012; Gusmao *et al.*, 2014) are in-
900 cluded into the model. Finally, a subset of approaches uses the prediction of TF binding
901 regions as an intermediate step for predicting gene regulation (Natarajan *et al.*, 2012)
902 or tissue-specific gene expression (Schmidt *et al.*, 2017).

903 **Supplementary Text S2 – Features**

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

909 **S2.1 Sequence-based features**

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

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

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

927 Finally, we consider the length of the longest poly-A or poly-T tract, the length of the
928 longest poly-C or poly-G tract, the length of the longest poly-A/T tract, and the length
929 of the longest poly-G/C tract in these three bins.

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

933 **S2.2 Annotation-based features**

934 Based on the Gencode v19 genome annotation of the *hg19* genome, we derive a set
935 of annotation-based features. First, we consider the distance of the current center bin
936 to the closest TSS annotation (regardless of its strand orientation), which might be
937 informative about core promoter regions. Second, we collect the binary information if

938 the current center bin overlaps with annotations of i) a CDS, ii) a UTR, iii) an exon,
939 iv) a transcript, or v) a TSS annotation, separately for each of the two possible strand
940 orientations. Like some of the previous features, this helps to identify coding, non-coding
941 but transcribed, core promoter, and intergenic regions. Again, these features are not TF
942 or cell type-specific, but model parameters may be adapted specifically for a TF or cell
943 type.

944 **S2.3 Motif-based features**

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

- 954 • PWM models from the “conservative” peak files for each training cell type,
- 955 • PWM models from the “relaxed” peak files complemented by negative regions se-
956 lected to be DNase positive (i.e., open chromatin) but ChIP-seq negative according
957 to the ChIP-seq and DNase-seq peak files provided with the challenge data,
- 958 • LSlim(3) models from the “conservative” peak files for each training cell type,
- 959 • LSlim(3) models from the “relaxed” peak files for each training cell type,
- 960 • LSlim(3) models from the “relaxed” peak files complemented by negative regions
961 selected to be DNase positive (i.e., open chromatin) but ChIP-seq negative accord-
962 ing to the ChIP-seq and DNase-seq peak files provided with the challenge data.

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

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

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

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

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

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

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

- 998 • the maximum log-probability of all sliding windows starting in the center bin,
- 999 • the logarithm of the sum of binding probabilities in all sliding windows starting in
1000 the center bin or its two adjacent bins, and
- 1001 • the logarithm of the sum of binding probabilities in all sliding windows starting in
1002 any of the bins considered.

1003 The first feature should capture the binding affinity at the strongest binding site around
1004 the peak summit, while the latter two features represent the general binding affinity of
1005 a region with different levels of resolution.

1006 For all of the remaining motifs, we consider the maximum of the bin-wise logarithm
1007 of the sum of binding probabilities over all bins considered (see section Binning the
1008 genome), as this reduces memory requirements as well as model complexity and this
1009 level of detail might be sufficient to capture TF interactions.

1010 **S2.4 DNase-based features**

1011 For the DNase-seq data, the challenge provided tracks with a “fold-enrichment coverage”
1012 track, peak files, and the original BAM files from mapping the DNase-seq reads, of which

1013 we consider only the former two. From the fold-enrichment coverage track, we compute
1014 the following statistics:

- 1015 • the minimum value across the center bin and its two adjacent bins,
- 1016 • the minimum of the maximum value within each bin considered,
- 1017 • the minimum of the 25% percentile within each bin considered, and
- 1018 • the median values of all the bins considered.

1019 After extracting those feature values for all genomic bins, we quantile normalize each
1020 of the features independently across the challenge cell types. Before normalization, we
1021 randomize the order of values to avoid systematic effects due to genomic order, which
1022 might especially occur for the large number of very low values. For the additional,
1023 primary cell types, we do not perform an independent quantile normalization but instead
1024 map the DNase-seq features (according to their numerical order) to the corresponding,
1025 quantile normalized values of the challenge cell types.

1026 In addition to these short-range DNase features, we also determine a set of long-range
1027 features, which are computed from i) 10 bins ii) 20 bins, and iii) 40 bins preceding and
1028 succeeding the current center bin. These features are

- 1029 • the minimum value across all bins,
- 1030 • the maximum value across all bins,
- 1031 • the minimum value across the bins preceding the center bin,
- 1032 • the minimum value across the bins succeeding the center bin,
- 1033 • the maximum value across the bins preceding the center bin, and
- 1034 • the maximum value across the bins succeeding the center bin.

1035 Together, these features capture chromatin accessibility on a short and long range level
1036 with reasonable resolution, which should be highly informative with regard to the general
1037 TF-binding potential. Model parameters should then be able to adapt for TF-specific
1038 preferences of chromatin accessibility.

1039 For the current center bin, we additionally determine features of stability across the
1040 different cell types, namely

- 1041 • the ratio of the minimum value in the current cell type divided by the average of
1042 the minimum values across all cell types,
- 1043 • the ratio of the maximum value in the current cell type divided by the average of
1044 the maximum values across all cell types,
- 1045 • the coefficient of variation (standard deviation divided by mean) of the minimum
1046 values across all cell types, and

1047 • the coefficient of variation of the maximum values across all cell types,

1048 where the latter two features are identical for all cell types by design.

1049 We also determine several features that represent the monotonicity/stability of these
1050 DNase-seq signals. Specifically, these features are

1051 • the number of steps (increasing or decreasing) in the track profile in a 450 bp
1052 interval centered at the center bin,

1053 • the longest strictly monotonically increasing stretch in the four bins preceding the
1054 center bin,

1055 • the longest strictly monotonically decreasing stretch in the four bins preceding the
1056 center bin,

1057 • the longest strictly monotonically increasing stretch in the four bins succeeding
1058 the center bin, and

1059 • the longest strictly monotonically decreasing stretch in the four bins succeeding
1060 the center bin.

1061 The first of these features has been inspired by the “orange” feature coined by team
1062 autosome.ru in the challenge.

1063 Finally, we define further features based on the “conservative” and “relaxed” DNase-
1064 seq peak files as provided with the challenge data. These are

1065 • the distance of the center bin to the summit of the closest conservative peak,

1066 • the distance of the center bin to the summit of the closest relaxed peak,

1067 • the peak statistic of a conservative peak overlapping the center bin (or zero if no
1068 such overlapping peak exists) multiplied by the length of the overlap,

1069 • the peak statistic of a relaxed peak overlapping the center bin (or zero if no such
1070 overlapping peak exists) multiplied by the length of the overlap,

1071 • the maximum of the q-values of an overlapping conservative peak (or zero if no
1072 such overlapping peak exists) multiplied by the length of the overlap across the
1073 five central bins,

1074 • the maximum of the q-values of an overlapping relaxed peak (or zero if no such
1075 overlapping peak exists) multiplied by the length of the overlap across the five
1076 central bins.

1077 **S2.5 RNA-seq-based features**

1078 The RNA-seq data provided with the challenge data included the TPM values of genes
1079 according to the gencode v19 genome annotation. TPM values are also quantile normal-
1080 ized across the cell types. As features, we consider

- 1081 • the maximum TPM value (averaged over the two bio-replicates per cell type) of
 1082 genes in at most 2.5 kb distance
- 1083 • the coefficient of variation of the bio-replicated of the corresponding gene,
- 1084 • the relative difference (difference of values in bio-replicated divided by their mean
 1085 value) of the corresponding gene.

1086 In analogy to the DNase-based features, we computed from the first feature as measures
 1087 of stability across the different cell types

- 1088 • the ratio of the maximum TPM value in the current cell type divided by the average
 1089 of the maximum values across all cell types, and
- 1090 • the coefficient of variation of the maximum TPM values across all cell types.

1091 **Supplementary Text S3 – Model & learning principle**

1092 For numerical features x , we use independent Gaussian densities parameterized as

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

1093 which allows for unconstrained numerical optimization of both, λ and μ .

1094 For features y with K possible discrete values v_1, \dots, v_K , we use (unnormalized) multi-
 1095 nomial distributions with parameters $\beta = (\beta_1, \dots, \beta_K)$ defined as

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

1096 The multinomial coefficient is neglected in this case, since it only depends on the in-
 1097 put data but not on the model parameters. In case of binary features, i.e., $K=2$, this
 1098 corresponds to an (unnormalized) binomial distribution.

1099 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
 1100 homogeneous Markov model of order 3 parameterized as

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

1101 where $\beta_{h,a|\mathbf{b}}$, $a \in \Sigma$, $\mathbf{b} \in \Sigma^3$ are the homogeneous parameters and

1102 $\beta_s = (\beta_{1,A}, \dots, \beta_{1,T}, \beta_{2,A|A}, \dots, \beta_{2,T|T}, \beta_{3,A|AA}, \dots, \beta_{3,T|TT}, \beta_{h,A|AAA}, \dots, \beta_{h,T|TTT})$ de-
 1103 notes the vector of all model parameters.

1104 Let $\mathbf{x} = (x_1, \dots, x_N)$ denote the vector of all numerical features, $\mathbf{y} = (y_1, \dots, y_M)$ de-
 1105 note the vector of all discrete features, and \mathbf{s} denote the raw sequence of one region repre-
 1106 sented by its feature values $\mathbf{z} = (\mathbf{x}, \mathbf{y}, \mathbf{s})$. Let $\theta = (\lambda_1, \dots, \lambda_N, \mu_1, \dots, \mu_N, \beta_1, \dots, \beta_M, \beta_s)$

1107 denote the set of all model parameters. We compute the likelihood of \mathbf{z} as an independent
1108 product of the terms for the individual features, i.e.,

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

1109 For modeling the distribution in the positive (foreground) and negative (background)
1110 class, we use likelihoods $P(\mathbf{z}|\boldsymbol{\theta}_{fg})$ and $P(\mathbf{z}|\boldsymbol{\theta}_{bg})$ with independent sets of parameters
1111 $\boldsymbol{\theta}_{fg}$ and $\boldsymbol{\theta}_{bg}$, respectively. In addition, we define the a-priori class probabilities as
1112 $P(fg|\gamma_1, \gamma_2) := \frac{\exp(\gamma_1)}{\exp(\gamma_1) + \exp(\gamma_2)}$ and $P(bg|\gamma_1, \gamma_2) = \frac{\exp(\gamma_2)}{\exp(\gamma_1) + \exp(\gamma_2)}$.

1113 Based on these definitions, we may compute the a-posteriori class probability of the
1114 positive class as

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

1115 and the a-posteriori class probability of the negative class in complete analogy.

1116 Using the discriminative maximum conditional likelihood principle (Roos *et al.*, 2005),
1117 the parameters are optimized such that the a-posteriori probabilities of the correct class
1118 labels given data and parameters are maximized. Here, we use a variant (Grau, 2010)
1119 of the maximum conditional likelihood principle that incorporates weights. Let $\mathbf{F} =$
1120 $(\mathbf{z}_1, \dots, \mathbf{z}_I)$ denote the set of positive examples and let $\mathbf{B} = (\mathbf{z}_{I+1}, \dots, \mathbf{z}_J)$ denote the
1121 set of negative examples, where \mathbf{z}_i is assigned weight w_i . The parameters are then
1122 optimized with regard to

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

1123 **Supplementary Text S4 – Sampling of DNase-matched negative regions**

1124 We sample negative regions with chromatin accessibility values matched to the positive
1125 regions (following an idea related to importance sampling) as explained in the following.
1126 We consider the center bins of all positive regions, collect the corresponding DNase-
1127 seq median feature values (see Supplementary Text S2) of those bins, and determine a
1128 histogram of the collected values. The histogram is composed of 20 equally sized bins
1129 between the observed maximum and minimum values of the DNase-seq median values.
1130 This histogram represents an approximation of the distribution of DNase-seq median
1131 values in the positive regions. As we expect DNase-seq values to be highly informative
1132 about TF binding, we aim at sampling a representative set of negative regions that
1133 exhibit similar DNase-seq values but might be distinguished from positive regions by
1134 other features.

1135 To this end, we assign each of the negative regions to the same histogram bins based
1136 on their respective DNase-seq median values at their center bins. This also yields an
1137 analogous histogram of the DNase-seq median values for the negative regions, which will
1138 usually be different from the histogram for the positive regions.

1139 Within each histogram bin, we then draw a subset of the negative regions assigned to
1140 that bin by i) drawing a subset of these regions four times as large as the corresponding
1141 positive set, and ii) weighting the drawn negative regions such that the sum of weights
1142 matches the relative abundance of that histogram bin in the histogram on all negative
1143 region.

1144 Conceptually, this procedure yields an over-sampling of negative regions with large
1145 DNase-seq median features, which is adjusted for by down-weighting such examples to
1146 the corresponding frequency on the chromosome level. This is especially important as
1147 these will be regions that are hard to classify using DNase-seq based features but are
1148 only lowly represented by the uniform sampling schema.