

1 **Genome-wide Prediction of DNase I Hypersensitivity Using Gene Expression**

2 Weiqiang Zhou¹, Ben Sherwood¹, Zhicheng Ji¹, Fang Du¹, Jiawei Bai¹, Hongkai Ji^{1,*}

3 ¹Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, 615 North
4 Wolfe Street, Baltimore, MD 21205, USA

5 * To whom correspondence should be addressed: hji@jhu.edu

6

7 Corresponding author:

8 Hongkai Ji, Ph.D.

9 Department of Biostatistics

10 Johns Hopkins Bloomberg School of Public Health

11 615 N Wolfe Street, Rm E3638

12 Baltimore, MD 21205, USA

13 Email: hji@jhu.edu

14 Phone: 410-955-3517

15

16 Running title:

17 Genome-wide Prediction of DNase I Hypersensitivity

18

19 Keywords:

20 DNase I hypersensitivity, Gene expression, Gene regulation, Big data regression, DNase-seq

21

22

23

24

25

26

27

28 **ABSTRACT**

29 We evaluate the feasibility of using a biological sample's transcriptome to predict its genome-wide
30 regulatory element activities measured by DNase I hypersensitivity (DH). We develop BIRD, Big Data
31 Regression for predicting DH, to handle this high-dimensional problem. Applying BIRD to the
32 Encyclopedia of DNA Element (ENCODE) data, we found that gene expression to a large extent predicts
33 DH, and information useful for prediction is contained in the whole transcriptome rather than limited to
34 a regulatory element's neighboring genes. We show that the predicted DH predicts transcription factor
35 binding sites (TFBSs), prediction models trained using ENCODE data can be applied to gene expression
36 samples in Gene Expression Omnibus (GEO) to predict regulome, and one can use predictions as
37 pseudo-replicates to improve the analysis of high-throughput regulome profiling data. Besides
38 improving our understanding of the regulome-transcriptome relationship, this study suggests that
39 transcriptome-based prediction can provide a useful new approach for regulome mapping.

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55 **INTRODUCTION**

56 A fundamental question in functional genomics is how genes' activities are controlled temporally and
57 spatially. To answer this question, it is crucial to comprehensively map activities of all genomic
58 regulatory elements (i.e., regulome) and understand the complex interplay between the regulome and
59 transcriptome (i.e., transcriptional activities of all genes). Regulome mapping has been accelerated by
60 high-throughput technologies such as chromatin immunoprecipitation coupled with high-throughput
61 sequencing (Johnson et al. 2007) (ChIP-seq) and sequencing of chromatin accessibility (e.g., DNase-seq
62 (Crawford et al. 2006) for DNase I hypersensitivity, FAIRE-seq (Giresi et al. 2007) for Formaldehyde-
63 Assisted Isolation of Regulatory Elements, and ATAC-seq (Buenrostro et al. 2013) for Assaying
64 Transposase-Accessible Chromatin). So far these technologies have only been applied to interrogate a
65 small subset of all possible biological contexts defined by different combinations of cell or tissue type,
66 disease state, time, environmental stimuli, and other factors. A major limitation of current high-
67 throughput technologies is the difficulty to simultaneously analyze a large number of different biological
68 contexts. This limitation along with various practical constraints such as lack of materials, antibodies,
69 resources, or expertise has hindered their application by the vast majority of biomedical investigators
70 from small laboratories.

71

72 For the study of regulome-transcriptome relationship, numerous researchers have examined how genes'
73 transcriptional activities can be predicted using activities of their associated regulatory elements
74 (Natarajan et al. 2012; Cheng et al. 2012; Kumar et al. 2013). However, the interplay between regulome
75 and transcriptome is bidirectional due to presence of feedback (Neph et al. 2012; Voss and Hager 2014).
76 A systematic understanding of this relationship in the reverse direction -- to what extent regulatory
77 elements' activities can be predicted by transcriptome -- is still lacking. We investigate this reverse
78 prediction problem using DNase I hypersensitivity (DH) and gene expression data generated by the
79 Encyclopedia of DNA Elements (ENCODE) Project (ENCODE Project Consortium 2012). Besides creating a
80 more complete picture of the regulome-transcriptome relationship, this investigation also has important
81 practical implications for regulome mapping. Gene expression is the most widely measured data type in

82 high-throughput functional genomics. Measuring expression does not require large amounts of
83 materials and complex protocols, and technologies for expression profiling are relatively mature. As a
84 result, expression data are routinely collected even when other functional genomic data types are
85 difficult to generate due to technical or resource constraints. Today, the Gene Expression Omnibus (GEO)
86 database (Edgar et al. 2002) contains 200,000+ human gene expression samples from a broad spectrum
87 of biological contexts, as compared to only ~7000 human ChIP-seq, DNase-seq, FAIRE-seq and ATAC-seq
88 samples available in GEO. We reasoned that if one can use the ENCODE data to build prediction models
89 and apply these models to existing and new transcriptome data to predict regulome, the catalog of
90 biological contexts with regulome information may be quickly expanded (**Fig. 1a**). This will provide a
91 useful approach for regulome mapping that is complementary to existing experimental methods. It will
92 also allow researchers to more effectively use expression data to study gene regulation. Unlike a recent
93 study that imputes one functional genomic data type based on multiple other data types which are non-
94 trivial to collect (Ernst and Kellis 2015), prediction in this study is based on one single but widely
95 available data type and hence can have a substantially broader range of applications. During our
96 investigation, we develop a big data regression approach, BIRD, to handle the prediction problem where
97 both predictors (i.e., transcriptome) and responses (i.e., regulome) are ultra-high-dimensional, which is
98 an emerging problem in the analysis of big data.

99

100 RESULTS

101 Big Data Regression for Predicting DNase I hypersensitivity (BIRD)

102 We obtained DNase-seq and exon array (i.e., gene expression) data for 57 distinct human cell types with
103 normal karyotype from ENCODE (**Supplementary Table 1**). The 57 cell types were randomly partitioned
104 into a training dataset (40 cell types) and a test dataset (17 cell types). After filtering out genomic
105 regions with weak or no DH signal across all 40 training cell types, 912,886 genomic loci (also referred to
106 as “DNase I hypersensitive sites” or “DHSs”) with unambiguous DNase-seq signal in at least one training
107 cell type were retained for subsequent analyses (**Methods**).

108

109 Our goal is to use gene expression to predict DH. This can be formulated as a problem of fitting millions
110 of regression models, one per genomic locus, to describe the relationship between DH (response) and
111 gene expression (predictor). The regression for each locus can be constructed using either its
112 neighboring genes or all genes as predictors (**Supplementary Fig. 1**). We tested both strategies (see
113 **Methods and Supplementary Figs. 2-4** for details). The latter strategy requires dealing with a
114 challenging big data regression problem which involves fitting about 1 million high-dimensional
115 regressions, each with a large number (18,000+) of predictors and small sample size. To cope with the
116 high dimensionality and heavy computation, we developed the BIRD algorithm (**Fig. 1b, Methods**). The
117 elementary BIRD model ($\text{BIRD}_{\bar{X},Y}$) groups correlated predictors into clusters and transforms each
118 cluster into one predictor. A prediction model for each genomic locus is then constructed using the
119 transformed predictors. Clustering reduces the predictor dimension, mitigates co-linearity, makes the
120 predictors less sensitive to measurement noise, and improves prediction accuracy. A variant of the
121 elementary model, $\text{BIRD}_{\bar{X},\bar{Y}}$, further clusters co-activated DHSs (i.e., correlated responses) and predicts
122 the mean DH level of each cluster. Finally, the compound BIRD model (BIRD) integrates the locus-level
123 predictions from $\text{BIRD}_{\bar{X},Y}$ and the cluster-level predictions from $\text{BIRD}_{\bar{X},\bar{Y}}$ via model averaging to better
124 balance the prediction bias and variance. A systematic benchmark analysis shows that BIRD not only
125 offers the computational efficiency suitable for big data regression, but also had the best prediction
126 performance in our problem compared to other methods (**Supplementary Methods, Supplementary Fig.**
127 **4**).

128

129 **Predicting DNase I Hypersensitivity Based on Gene Expression**

130 We applied BIRD to the 40 training cell types to build prediction models, and evaluated their prediction
131 performance in the 17 test cell types using three types of statistics: (1) the Pearson correlation between
132 the predicted and true DH values (or P-T correlation) across different genomic loci within each cell type
133 (r_L), (2) the P-T correlation across different cell types at each genomic locus (r_C), and (3) the total squared
134 prediction error scaled by the total DH data variance (τ) (**Fig. 1c, Methods**). These analyses led to the
135 following findings.

136

137 *Gene expression provides valuable information for predicting DH.* **Figure 1d-g** compares r_L , r_C and τ from
138 different methods (**Methods**). These plots show that the elementary BIRD model ($BIRD_{\bar{X},Y}$) significantly
139 increased the P-T correlation (r_L and r_C) and substantially decreased the squared prediction error (τ)
140 compared to random prediction models (BIRD-Permute). The random prediction models were obtained
141 by applying the same $BIRD_{\bar{X},Y}$ analysis after permuting the link between DNase-seq and gene expression
142 data in the training dataset.

143

144 *Prediction based on the whole transcriptome substantially improves prediction based on a genomic locus'*
145 *neighboring genes.* We tested the neighboring gene approach by gradually increasing the number of
146 neighboring genes and identified the optimal performance (**Methods, Supplementary Fig. 2**). Compared
147 to the best prediction performance by the neighboring gene approach, $BIRD_{\bar{X},Y}$ substantially increased
148 the prediction accuracy (**Fig. 1d-g**), indicating that not all information useful for prediction is contained
149 in neighboring genes. This is plausible biologically because DH of a locus may be correlated *in trans* with
150 expression of TFs that bind to the locus, genes that co-express with these TFs, and genes that co-express
151 with the target gene controlled *in cis* by the locus. Moreover, since cell-type-specific transcription of a
152 gene may be controlled by multiple *cis*-regulatory elements, DH of a particular regulatory element may
153 not always correlate well with its neighboring gene expression.

154

155 *Clustering correlated predictors (i.e., co-expressed genes) helps prediction.* In $BIRD_{\bar{X},Y}$, correlated
156 predictors are consolidated by clustering. $BIRD_{X,Y}$ is a special case of $BIRD_{\bar{X},Y}$ in which predictors are
157 not clustered whereas all subsequent predictor selection and model fitting procedures remain the same
158 (**Methods**). Compared to $BIRD_{X,Y}$, $BIRD_{\bar{X},Y}$ produced higher prediction accuracy (**Fig. 1d-g**,
159 **Supplementary Fig. 3b**). This shows that in a high-dimensional regression setting where predictors far
160 outnumber the sample size, clustering correlated predictors before variable selection and model fitting,
161 a technique not widely used in high-dimensional regression literature, can improve the model compared

162 to conventional techniques (Tibshirani 1996; Fan and Lv 2008) that directly apply variable selection to
163 reduce the predictor dimension.

164
165 *DH variation across different genomic loci within a cell type can be accurately predicted.* In the 17 test
166 cell types, the mean cross-locus P-T correlation r_L of $\text{BIRD}_{\bar{X},Y}$ was 0.81 (**Fig. 1d**). Interestingly, random
167 prediction models were also able to produce large r_L (**Fig. 1d**, mean = 0.65). This is because different loci
168 have different DH propensity, consistent with observations in a previous study (Ernst and Kellis 2015).
169 For instance, some loci tend to show higher DH signal than other loci in most cell types (**Supplementary**
170 **Fig. 5**). As a result, using the average DH profile of all training cell types can predict the cross-locus DH
171 variation in a new cell type with good accuracy (Ernst and Kellis 2015), even though such predictions are
172 cell-type-independent and remain the same for all new cell types. Our random prediction models were
173 generated by permutations that did not perturb the locus-specific DH propensity. Therefore, their r_L was
174 large. Since $\text{BIRD}_{\bar{X},Y}$ uses cell-type-dependent information carried by transcriptome, its predictions are
175 more accurate (**Fig. 1d**).

176
177 *DH variation across cell type can be predicted, although it is more challenging than predicting cross-locus*
178 *variation.* **Figure 2a** shows an example demonstrating that the true cross-cell-type DH variation
179 measured by DNase-seq can be captured by BIRD predictions, but not by the mean DH profile of all
180 training cell types. Comparing the cross-locus P-T correlation (r_L) in **Figure 1d** with the cross-cell-type P-T
181 correlation (r_C) in **Figure 1e**, r_L on average was substantially larger than r_C (e.g., 0.81 vs. 0.48 for
182 $\text{BIRD}_{\bar{X},Y}$). Unlike r_L , the distribution of r_C for random prediction models was centered around zero (**Fig.**
183 **1e**, mean = -0.03) because the cross-cell-type prediction accuracy was evaluated within each locus and
184 hence not affected by locus effects. Compared to random prediction models, $\text{BIRD}_{\bar{X},Y}$ significantly
185 increased r_C (**Fig. 1e,g**).

186
187 *Cross-cell-type DH variation of regulatory element pathways can be predicted with substantially higher*
188 *accuracy than that of individual loci.* This can be illustrated by comparing $\text{BIRD}_{\bar{X},Y}$ with $\text{BIRD}_{\bar{X},\bar{Y}}$. In

189 BIRD_{X,Y}, we first grouped correlated genomic loci into 1,000 clusters using the training data (**Methods**).
190 Loci within each cluster share similar cross-cell-type DH variation pattern and hence can be viewed as a
191 “pathway” consisting of co-activated regulatory elements (Thurman et al. 2012; Sheffield et al. 2013).
192 BIRD_{X,Y} predicted the mean DH level of each cluster in each test cell type. The cross-cell-type P-T
193 correlation r_c for the cluster-level prediction was substantially higher than r_c for the locus-level
194 prediction (**Fig. 2b**, mean r_c for BIRD_{X,Y₁₀₀₀} vs. BIRD_{X,Y} = 0.71 vs. 0.48). When genomic loci were
195 grouped into 2,000 or 5,000 clusters, we obtained similar results (**Fig. 2b**). Thus, similar to gene set
196 analysis (Subramanian et al. 2005), the overall cross-cell-type activity of a pathway of co-activated
197 regulatory elements can be more reliably studied through prediction than that of individual loci. From a
198 statistical perspective, the cluster mean can reduce the variance of random noise by averaging many
199 measurements. Therefore, it can provide a cleaner signal that is easier to predict.

200
201 *The compound BIRD model improves locus-level prediction.* The compound BIRD model (denoted as BIRD)
202 combines the locus-level prediction by BIRD_{X,Y} and cluster-level prediction by BIRD_{X,Y} to balance the
203 prediction bias and variance. As a result, it increased the locus-level prediction accuracy compared to
204 BIRD_{X,Y} (**Fig. 1d-g, Methods**).

205
206 *Cross-cell-type prediction accuracy varies greatly among different loci.* For the compound BIRD model, r_c
207 of different genomic loci varied substantially (**Fig. 1e**, mean = 0.5). For 6% of loci, $r_c < 0$ (i.e., prediction
208 did not help). On the other hand, 56% and 20% of loci had $r_c > 0.5$ and >0.75 respectively, indicating that
209 DH could be predicted with moderate to high accuracy for a substantial fraction of loci. For each locus,
210 we computed the coefficient of variation (CV) to characterize the variability of the predicted DH across
211 the test cell types (**Methods**). We found that loci with poor cross-cell-type prediction accuracy (i.e.,
212 small r_c) also tend to be less variable (i.e., had small CV) in the test cell types (**Fig. 2c,d**). Computing CV
213 using the true DNase-seq data instead of the predicted DH yielded qualitatively similar results
214 (**Supplementary Fig. 6**). One possible explanation for this phenomenon is that, compared to a highly
215 variable locus, DH variation observed at a lowly variable locus is more likely due to random noise rather

216 than true biological signals, and the correlation between predictions and random noise is expected to be
217 zero. The CV of predicted DH provides a way to screen for loci whose cross-cell-type prediction is likely
218 to be accurate. For instance, if we were to focus on loci with $CV > 0.4$ rather than all loci, the mean r_c
219 would increase from 0.5 to 0.61, and 74% and 37% of loci would have $r_c > 0.5$ and > 0.75 respectively (**Fig.**
220 **2e**). Compared to the locus-level prediction by BIRD, cross-cell-type prediction accuracy by $BIRD_{\bar{X}, \bar{Y}}$ at
221 the cluster-level was both more accurate and less variable (**Fig. 2b**). For $BIRD_{\bar{X}, \bar{Y}_{1000}}$, 84% and 55%
222 clusters had $r_c > 0.5$ and > 0.75 respectively. The results were similar for $BIRD_{\bar{X}, \bar{Y}_{2000}}$ and $BIRD_{\bar{X}, \bar{Y}_{5000}}$.

223
224 *Comparisons of BIRD and ChromImpute.* ChromImpute is a recently developed method for imputing one
225 functional genomic data type using multiple other data types (Ernst and Kellis 2015). We compared DH
226 predictions by BIRD using only gene expression data with DH predictions by ChromImpute using multiple
227 functional genomic data types (**Supplementary Methods**). Among 10 tested cell types, BIRD and
228 ChromImpute showed comparable prediction performance. Neither method consistently outperformed
229 the other (**Fig. 2f, Supplementary Fig. 7**). However, ChromImpute used ChIP-seq data for multiple
230 histone modifications as predictors (these are the best predictors selected by ChromImpute for imputing
231 DH (Ernst and Kellis 2015)), which are non-trivial to generate. By contrast, BIRD was based on gene
232 expression data alone which are easier to generate and widely available.

233
234 *Robustness analysis.* The conclusions above do not depend on how the 57 cell types are partitioned into
235 the training and testing data. We repeated the same analyses on four other random partitions (**Methods**,
236 **Supplementary Table 1**), and similar results were obtained. For instance, **Supplementary Figure 8** shows
237 that r_L , r_c and τ for BIRD from different partitions were similar.

238
239 **Predicting Transcription Factor Binding Sites Based on Gene Expression**
240 We asked whether the predicted DH at DNA motif sites can predict transcription factor binding sites
241 (TFBSs). Using BIRD models based on the 40 training cell types, we predicted TFBSs for 9 TFs in GM12878
242 cell line which was not in the training data. The predictions were evaluated using the corresponding TF

243 ChIP-seq data from ENCODE in the same cell line. As a comparison, we also predicted TFBSs using true
244 DNase-seq data (positive control) and using the mean DH profile of the training cell types (negative
245 control). **Figure 3a-b** and **Supplementary Figure 9a-g** show how sensitivity of detecting motif-containing
246 ChIP-seq binding sites changed with increasing number of predictions. For example, for TF ELF1 in
247 GM12878, top 15000 BIRD (UW) predictions gave a sensitivity of 0.76 at an estimated false discovery
248 rate (FDR, measured using *q*-value) of 0.05 (**Fig. 3b, Supplementary Methods**). As expected, true DNase-
249 seq data predicted TFBSs better than BIRD. However, BIRD substantially improved the prediction based
250 on the mean DH profile. For BIRD, predictions were made using exon array data generated by three
251 different laboratories. The lab difference turned out to be smaller than the differences between
252 prediction methods (**Fig. 3a-b, Supplementary Fig. 9a-g**). A similar analysis for 3 other TFs in K562 cell
253 line yielded similar results (**Supplementary Methods, Fig. 3c, Supplementary Fig. 9h-i**).

254
255 To further demonstrate BIRD in a realistic setting, we retrained BIRD using all 57 cell types for 1,108,603
256 loci with DH signal in at least one cell type. We then applied it to exon array data for P493-6 B cell
257 lymphoma (a non-ENCODE cell line) generated by a non-ENCODE lab (Ji et al. 2011). We predicted MYC
258 binding sites by identifying and ranking E-box motif sites CACGTG based on the predicted DH signal
259 (**Supplementary Methods**). The predictions were evaluated using MYC ChIP-seq data (Sabò et al. 2014)
260 in P493-6 cells (**Supplementary Methods**), from which 12,484 MYC binding peaks (FDR<0.01)
261 overlapping E-box motif sites were discovered and served as the gold standard. **Figure 3d** shows the
262 prediction performance. Among the top 20,000 predicted MYC binding sites (*q*-value < 0.073), 10,866
263 (54%) were indeed bound by MYC according to MYC ChIP-seq. The remaining 46% may represent a
264 mixture of noise and true binding sites of other TFs since the E-box motif can also be recognized by
265 multiple other TFs. In terms of sensitivity, 8,338 (67%) MYC binding peaks were overlapped with the
266 predicted MYC binding sites (one peak may overlap with >1 DHSs). Thus, despite the fact that the
267 training and test data have different lab origins, one can discover a substantial fraction of true MYC
268 binding sites. The predicted DH also showed strong correlation with the true ChIP-seq signal (**Fig. 3e,g**).
269 By contrast, predictions based on the mean DH profile of the 57 training cell types had substantially

270 lower prediction accuracy (**Fig. 3d,f-g**). This demonstrates that in the absence of ChIP-seq data, one may
271 use gene expression to predict TFBSS to identify promising follow-up targets.

272

273 **Regulome Prediction Based on 2000 Public Gene Expression Samples in GEO**

274 The vast amounts of gene expression data from diverse biological contexts in GEO represent a resource
275 that no single laboratory can generate. As a proof-of-principle test, we collected 2,000 human exon
276 array samples from GEO and applied BIRD trained using all 57 ENCODE cell types for 1,108,603 loci to
277 these samples to predict regulome. These predictions are made available as a web resource PDDB
278 (Predicted DNase I hypersensitivity database). A user interface is provided for data query, display and
279 download (**Fig. 4a-c, Methods, Supplementary Methods**).

280

281 Researchers can use PDDB to explore regulatory element activities in biological contexts for which they
282 do not have available regulome data. As a feasibility test, we first queried predicted DH for three genes
283 FBL, LIN28A and BLMH in P493-6 B cell lymphoma (for which no public DNase-seq data are available)
284 and H9 human embryonic stem cells. Promoters of these genes are known to be bound by MYC in a cell
285 type dependent fashion (Ji et al. 2011). FBL is bound in both P493-6 and H9, LIN28A is bound in H9 but
286 not in P493-6, and BLMH is bound in P493-6 but not in H9 (Koh et al. 2011; Chang et al. 2009; Ji et al.
287 2011). PDDB successfully predicted these known cell-type-dependent binding patterns (**Fig. 5a-c,**
288 **Supplementary Fig. 10**).

289

290 Next, we obtained a list of SOX2 binding sites in human embryonic stem cells from a published ChIP-seq
291 study (Watanabe et al. 2014) (**Supplementary Methods**). **Figure 5d** shows the predicted DH at these
292 sites across the 2,000 GEO samples. The samples were ordered based on the overall DH enrichment
293 level at all SOX2 binding sites relative to random genomic sites (**Supplementary Methods, Fig. 5e**).
294 Samples with strong predicted DH at SOX2 binding sites include stem cells (green bar in **Fig. 5d**) and
295 brain (brown bar), consistent with known roles of SOX2 in these sample types (Chambers and Tomlinson
296 2009; Takahashi and Yamanaka 2006; Ferri et al. 2004; Phi et al. 2008). Interestingly, PDDB contained

297 differentiating H7 embryonic stem cells collected at day 2, 5 and 9 after initiation of differentiation. Our
298 57 training cell types contained undifferentiated H7 cells and H7 cells at differentiating day 14. Together,
299 these samples formed a time course. Examination of the predicted DH for day 2, 5, and 9 along with the
300 true DH for day 0 and 14 shows that the predicted DH at SOX2 binding sites decreased as the
301 differentiation progressed (**Fig. 5f-g**), consistent with the known role of SOX2 for maintaining the
302 undifferentiated status of stem cells (Takahashi and Yamanaka 2006; Chambers and Tomlinson 2009).
303 Thus, the dynamic changes of SOX2 binding activities were correctly predicted in PDDB.

304
305 The above examples show that expression samples in GEO can be used to meaningfully predict DH. With
306 ChIP-seq data for a TF from one biological context, one may also use PDDB to systematically explore in
307 what other biological contexts each binding site might be active, and group TFBSs into functionally
308 related subclasses accordingly. For instance, we obtained MEF2A ChIP-seq binding sites in GM12878
309 lymphoblastoid cells from ENCODE. MEF2A is a TF involved in muscle development (Edmondson et al.
310 1994) and neuronal differentiation (Flavell et al. 2008). Using PDDB (**Supplementary Methods, Fig. 5h-i**,
311 **Supplementary Fig. 11, Supplementary Tables 5-6**), we first clustered samples and MEF2A binding sites
312 into different groups and performed functional annotation analysis on each group using the Database
313 for Annotation, Visualization and Integrated Discovery (DAVID) (Huang et al. 2009; Huang et al. 2008). A
314 group of MEF2A binding sites associated with genes involved in cell motion, cell migration and
315 regulation of metabolic processes was found to be more active in muscle related samples (including
316 coronary artery smooth muscle and cardiac precursor cell which are not covered by ENCODE) than in
317 lymphoblastoid (**Fig. 5h-i**). Another group of sites associated with neuron differentiation and
318 neurogenesis genes was found to be more active in neuron and brain related samples (including non-
319 ENCODE sample types such as entorhinal cortex and motor neuron) (**Fig. 5h-i**). This demonstrates how
320 PDDB can provide a more detailed view of TFBSs not offered by the original experiment in GM12878,
321 and how PDDB can be used to investigate many biological contexts not covered by ENCODE.

322

323

324 **Predictions as Pseudo-Replicates to Improve Analyses of DNase-seq and ChIP-seq Data**

325 In applications of high-throughput regulome profiling technologies, it is common to encounter data with
326 low signal-to-noise ratio or small replicate number. Both can lead to low signal detection power.
327 However, if one has gene expression data, BIRD predictions may be used as pseudo-replicates to
328 enhance the signal. As a test, we analyzed DNase-seq data for GM12878 generated by ENCODE. The
329 data had two replicates. We reserved one replicate as “truth” and used the other one as the “observed”
330 data. Applying the BIRD prediction models trained earlier using the 40 training cell types (GM12878 not
331 included), we predicted DH in GM12878 and treated the prediction as a pseudo-replicate. We then
332 estimated “true” DH using either the “observed” data alone (obs-only) or the average of the “observed”
333 data and pseudo-replicate (BIRD+obs). After adding the pseudo-replicate, the correlation between the
334 predicted and true DH increased (**Fig. 6a-b**, r_L for BIRD+obs vs. obs-only = 0.82 vs. 0.77). Replacing BIRD
335 predictions with the mean DH profile of 40 training cell types in this analysis (Mean+obs) did not yield
336 similar increase in the P-T correlation (r_L = 0.76). We carried out the same analyses on 16 test cell types,
337 and BIRD predictions improved signal in 12 of them (**Fig. 6c, Supplementary Methods**).
338

339 Similarly, we tested if the predicted DH can boost ChIP-seq signals using ChIP-seq data for 9 TFs in
340 GM12878 and 3 TFs in K562 (**Supplementary Methods**). Similar results were observed (**Fig. 6d-f**).
341 BIRD+obs outperformed obs-only in nearly all test cases (11 out of 12 TFs). Together, these results show
342 that predictions can serve as a bridge to integrate expression and regulome data so that one can more
343 effectively use available information to improve data analysis.

344

345 **DISCUSSION**

346 In summary, this study for the first time examined systematically to what extent regulatory element
347 activities can be predicted by gene expression alone. We developed BIRD for big data prediction. The
348 study also demonstrates the feasibility of using gene expression to predict TFBSSs, applying BIRD to GEO
349 to expand the current regulome catalog, and using predictions to facilitate data integration. BIRD is a
350 novel approach to extract information from gene expression data to study regulome. In the absence of

351 experimental regulome data (e.g., ChIP-seq or DNase-seq data), BIRD predictions can provide valuable
352 information to guide hypothesis generation, target prioritization, and design of follow-up experiments.
353 When experimental regulome data are available, BIRD predictions can also serve as pseudo-replicate to
354 improve the data analysis. In a companion study, we show that BIRD can also predict DH using RNA-seq
355 and in samples with small number of cells, and it can outperform state-of-the-art technologies for
356 mapping regulome in small-cell-number samples (Zhou et al. submitted).

357
358 Our results have important practical implications for the analysis of existing and future gene expression
359 data. Conventionally, gene expression data are mainly collected to study transcriptome. The method
360 and software developed in this study now allow one to conveniently utilize such data to study gene
361 regulation. By adding a new component to the standard analysis pipeline of expression data, expression-
362 based regulome prediction can bring added value to an enormous number of new and existing gene
363 expression experiments. Given the wide application of gene expression profiling, this will greatly impact
364 how expression data are most effectively used.

365
366 Compared to conventional regulome mapping technologies, BIRD also has its unique advantages. Since
367 gene expression profiling experiments are more widely conducted than regulome mapping experiments,
368 the number of biological contexts with gene expression data is orders of magnitude larger than the
369 number of contexts with experimental regulome data. BIRD can be readily applied to massive amounts
370 of existing and new gene expression data to generate regulome information for a large number of
371 biological contexts without experimental regulome data. In the near future, no other experimental
372 regulome mapping technology can achieve similar level of comprehensiveness in terms of biological
373 context coverage.

374
375 Our current study may be extended in multiple directions in the future. For instance, it is important to
376 extend BIRD to other gene expression platforms. It also remains to be answered whether gene
377 expression can be similarly used to predict other functional genomic data types.

378 **METHODS**

379 **DNase-seq data processing**

380 The bowtie (Langmead et al. 2009) aligned (alignment based on hg19) DNase-seq data for 57 human cell
381 types with normal karyotype were downloaded from the ENCODE in bam format (download link:
382 <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase>). The human
383 genome was divided into 200 base pair (bp) non-overlapping bins. The number of reads falling into each
384 bin was counted for each DNase-seq sample. To adjust for different sequencing depths, bin read counts
385 for each sample i were first divided by the sample's total read count N_i and then scaled by multiplying a
386 constant N ($N = \min_i\{N_i\} = 17,002,867$, which is the minimum sample read count of all samples). After
387 this procedure, the raw read count n_{li} for bin l and sample i was converted into a normalized read
388 count $\tilde{n}_{li} = Nn_{li}/N_i$. The normalized read counts from replicate samples were averaged to characterize
389 the DH level for each bin in each cell type. The DH level was then log2 transformed after adding a
390 pseudocount 1. The transformed data were used for training and testing prediction models, treating
391 each bin as a genomic locus. Since chromosome Y was not present in all samples, we excluded this
392 chromosome from our subsequent analyses.

393

394 **Gene expression data processing**

395 The Affymetrix Human Exon 1.0 ST Array (i.e. exon array) data for the same 57 ENCODE cell types were
396 downloaded from GEO (GEO accession number: GSE19090). Additionally, we downloaded 2000 exon
397 array samples from GEO for constructing the PDDB database (GEO accession numbers for these samples
398 are available at PDDB). All samples were processed using the GeneBASE (Kapur et al. 2007) software to
399 compute gene-level expression. The output of GeneBASE was expression levels of 18,524 genes in each
400 sample. The GeneBASE gene expression levels were log2 transformed after adding a pseudocount 1 and
401 then quantile normalized (Bolstad 2015) across samples. For the 57 ENCODE cell types, replicate
402 samples within each cell type were averaged and the averaged mean expression profile of each cell type
403 was used for training and testing the prediction models.

404

405 **Training-test data partitioning and genomic loci filtering**

406 The 57 ENCODE cell types were randomly partitioned into a training dataset with 40 cell types and a test
407 dataset with 17 cell types (**Supplementary Table 1**, partition # 1). Since not all genomic loci are
408 regulatory elements, we first screened for genomic loci with unambiguous DH signal in at least one cell
409 type in the training data as follows. Genomic bins with normalized read count >10 in at least one cell
410 type were identified and retained, and the other genomic bins were excluded. Among the retained loci,
411 bins with normalized read count >10,000 in any cell type were considered abnormal and these bins were
412 also excluded from subsequent analyses. Finally, for each remaining bin, a signal-to-noise ratio (SNR)
413 was computed in each cell type, and bins with small SNR in all cell types were filtered out. To compute
414 SNR of a genomic bin in a cell type, we first collected 500 bins in the neighborhood of the bin in question.
415 Then, we computed the average DH level of these bins. Next, the DH level was log2 transformed after
416 adding a pseudocount 1 to serve as the background. The $\log_2(\text{SNR})$ was defined as the difference
417 between the normalized and log2 transformed DH level of the bin in question and the background.
418 Genomic bins with $\log_2(\text{SNR}) > 2$ in at least one cell type were identified and retained for subsequent
419 analyses, and the other genomic bins were excluded. After applying this filtering procedure to the 40
420 training cell types, 912,886 genomic bins were retained and used for training and testing prediction
421 models in **Figures 1** and **2**. Bins selected by this procedure were referred to as DNase I hypersensitive
422 sites (DHSs) in this article. We note that the above filtering procedure only uses the training cell types.
423 This allows one to objectively evaluate the prediction performance in real applications where models
424 trained using the training cell types are applied to make predictions in new cell types for which DNase-
425 seq data are not available.

426

427 In order to evaluate the robustness of our conclusions, we repeated the same random partitioning
428 procedure five times, resulting in five different training-test data partitions (**Supplementary Table 1**).
429 For each partition, genomic loci were filtered using the same protocol described above, and the retained
430 loci (which depend on the training data and therefore are different for different partitions) were used to

431 train and test BIRD. Results from the first partition were presented in the main article, and results from
432 the other four random partitions were similar (**Supplementary Fig. 8**).

433
434 For predicting TFBSSs in K562 and P493-6 B cell lymphoma and the analyses of 2000 GEO exon array
435 samples used for constructing PDDB, prediction models were retrained using all 57 ENCODE cell types as
436 training data. Applying the genomic loci filtering protocol described above to these 57 cell types resulted
437 in 1,108,603 genomic bins for which prediction models were constructed and evaluated.

438
439 **Notations and problem formulation**
440 For a biological sample, let Y_l be the DH level of genomic locus l ($=1, \dots, L$), and let X_g be the expression
441 level of gene g ($=1, \dots, G$). The genome-wide DH profile and gene expression profile are represented by
442 two vectors $\mathbf{Y} = (Y_1, \dots, Y_L)^T$ and $\mathbf{X} = (X_1, \dots, X_G)^T$ respectively. Here, the superscript T indicates
443 matrix or vector transpose. Both the DH and gene expression profiles are assumed to be normalized and
444 at log2 scale. Our goal is to use \mathbf{X} to predict \mathbf{Y} . This can be formulated as a problem of building a
445 regression $Y_l = f_l(\mathbf{X}) + \epsilon_l$ for each genomic locus. Here ϵ_l represents random noise, and $f_l(\cdot)$ is the
446 function that describes the systematic relationship between the DH level of locus l (i.e., Y_l) and the gene
447 expression profile (i.e., \mathbf{X}).

448
449 The function $f_l(\mathbf{X})$ is unknown. We train it using \mathbf{X} and \mathbf{Y} observed from a number of different cell types.
450 The training data are organized into two matrices: a gene expression matrix $\mathbb{X} = (x_{gc})_{G \times C}$ and a DH
451 matrix $\mathbb{Y} = (y_{lc})_{L \times C}$. Rows in these matrices are genes and genomic loci respectively. Columns in these
452 matrices are cell types. C is the number of training cell types. Each column of \mathbb{X} and \mathbb{Y} is a realization of
453 the random vector \mathbf{X} and \mathbf{Y} in a specific cell type. Building the prediction model for each locus l is a
454 challenging high-dimensional regression problem since the dimensionality of the predictor \mathbf{X} is much
455 bigger than the sample size of the training data (i.e., $G \gg C$). What makes this problem even more
456 challenging than the conventional high-dimensional problems in statistics is that one needs to solve a
457 massive number of such high-dimensional regression problems (one for each locus) simultaneously.

458 Thus it is important to consider both statistical efficiency and computational efficiency when developing
459 solutions.

460
461 In subsequent sections, various methods for training $f_l(\mathbf{X})$ will be described. Each method has a training
462 component and a prediction component. Before training prediction models, we standardize each row of
463 \mathbb{X} and \mathbb{Y} in the training data to have zero mean and unit standard deviation (SD). More precisely, each
464 DH value in \mathbb{Y} is standardized using $\tilde{y}_{lc} = (y_{lc} - a_l^y)/s_l^y$ where a_l^y and s_l^y are the mean and SD of the
465 DH signals at locus l (i.e., row l of \mathbb{Y}). Similarly, each expression value in \mathbb{X} is standardized using $\tilde{x}_{gc} =$
466 $(x_{gc} - a_g^x)/s_g^x$ where a_g^x and s_g^x are the mean and SD of the gene expression for gene g (i.e., row g of
467 \mathbb{X}). The prediction models are then constructed using the standardized values $\tilde{\mathbb{X}}$ and $\tilde{\mathbb{Y}}$.

468
469 Once the models are constructed using the training data, they can be applied to new samples to make
470 predictions. To do so, the expression profile \mathbf{X} of the new sample is first quantile normalized to the
471 quantiles of the training exon array data. The log2-transformed expression value of each gene X_g in the
472 new sample is then standardized using $\tilde{X}_g = (X_g - a_g^x)/s_g^x$, where a_g^x and s_g^x are the pre-computed
473 mean and SD of the gene expression for gene g in the training data. After applying the trained model to
474 the standardized gene expression profile $\tilde{\mathbf{X}}$ to make predictions, the predicted DH value for each locus,
475 \tilde{Y}_l , is transformed back using $\hat{Y}_l = s_l^y * \tilde{Y}_l + a_l^y$, where a_l^y and s_l^y are the pre-computed mean and SD of
476 the DH signals for locus l in the training data. The unstandardized \hat{Y}_l gives the prediction for Y_l , the DH
477 level of genomic locus l in the new sample.

478
479 **Measures for method evaluation**
480 In order to evaluate prediction performance of a prediction method, the method can be applied to a
481 number of test cell types to predict their DH profiles based on their gene expression profiles. Let \hat{y}_{lm} be
482 the predicted DH level of locus l in test cell type m ($=1, \dots, M$), and let y_{lm} be the true DH level
483 measured by DNase-seq (both are at log2 scale). Three performance statistics were used in this study
484 (**Fig. 1c**):

485

486 (1) Cross-locus correlation (r_L). This is the Pearson's correlation between the predicted signals $\hat{\mathbf{y}}_{*m} =$
487 $(\hat{y}_{1m}, \dots, \hat{y}_{Lm})^T$ and the true signals $\mathbf{y}_{*m} = (y_{1m}, \dots, y_{Lm})^T$ across different loci for each test cell type m .
488 The cross-locus correlation measures the extent to which the DH signal within each cell type can be
489 predicted.

490

491 (2) Cross-cell-type correlation (r_C). This is the Pearson's correlation between the predicted signals $\hat{\mathbf{y}}_{l*} =$
492 $(\hat{y}_{l1}, \dots, \hat{y}_{lM})$ and the true signals $\mathbf{y}_{l*} = (y_{l1}, \dots, y_{lM})$ across different cell types for each locus l . The
493 cross-cell-type correlation measures how much of the DH variation across cell type can be predicted.

494

495 (3) Squared prediction error (τ). This is measured by the total squared prediction error scaled by the
496 total DH data variance in the test dataset: $\tau = \frac{\sum_l \sum_m (y_{lm} - \hat{y}_{lm})^2}{\sum_l \sum_m (y_{lm} - \bar{y})^2}$, where \bar{y} is the mean of y_{lm} across all
497 DHSs and test cell types.

498

499 **Prediction based on neighboring genes**

500 For each genomic locus l , N closest genes were identified (gene annotation based on RefSeq genes of
501 human genome hg19 downloaded from UCSC genome browser: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refFlat.txt.gz>). The closeness was defined by the distance between the
502 gene's transcription start site and the locus center. Using the selected genes $(\tilde{X}_{l_1}, \dots, \tilde{X}_{l_N})$ as predictors,
503 a multiple linear regression $\tilde{Y}_l = \beta_{l0} + \beta_{l1}\tilde{X}_{l_1} + \dots + \beta_{lN}\tilde{X}_{l_N} + \epsilon_l$ is fitted. Based on the fitted model, the
504 standardized DH level of locus l in a new sample is predicted using $\tilde{Y}_l = f_l(\tilde{\mathbf{X}}) = \beta_{l0} + \beta_{l1}\tilde{X}_{l_1} + \dots +$
505 $\beta_{lN}\tilde{X}_{l_N}$. We tested different values of N (= 1, 2, ..., 20) on a randomly selected set of DHSs (n=9,128; ~1%
506 of the 912,886 DHSs obtained from the 40 training cell types). The performance for the neighboring
507 gene approach shown in **Figure 1d-g** was based on the performance achieved at the optimal N . For
508 instance, **Supplementary Figure 2a** shows the r_C distribution for different N based on the 9,128 DHSs. At
509 $N=15$, the mean r_C reached its maximum. Correspondingly, the r_C distribution shown in **Figure 1e** was
510 based on $N=15$.

512 We also tested whether nonlinear regression can improve the prediction. Generalized additive model
513 with smoothing spline (GAM) was applied (using R package “gam” (Hastie 2015)) to the same 1% of
514 DHSs. However, the best prediction performance of GAM was worse than the best prediction
515 performance of the linear regression (**Supplementary Fig. 2a**, see the best performance of GAM
516 achieved at $N = 17$ vs. the best performance of linear model achieved at $N = 15$). This indicates that
517 using non-linear model did not improve prediction accuracy. Moreover, the computational time
518 required by GAM was substantially longer than linear regression (**Supplementary Fig. 2b**), making it
519 difficult to apply to the whole genome. Based on this, linear regression was used to perform our
520 genome-wide analysis.

521

522 **BIRD $_{\bar{X},Y}$ – The elementary BIRD model**

523 BIRD $_{\bar{X},Y}$ is the basic building block of BIRD. This approach first groups correlated genes into clusters.
524 This is achieved by clustering rows of the standardized training data matrix $\tilde{\mathbf{X}}$ into K clusters using k-
525 means clustering (Hartigan and Wong 1979) (Euclidean distance used as similarity measure). Based on
526 the clustering result, the gene expression profile $\tilde{\mathbf{X}}$ of each sample is converted into a lower dimensional
527 vector $\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_K)$, where \bar{X}_k is the mean expression level of genes in cluster k . BIRD will use gene
528 clusters’ mean expression $\bar{\mathbf{X}}$ instead of the expression of individual genes $\tilde{\mathbf{X}}$ as predictors to build
529 prediction models. Clustering serves multiple purposes. It reduces the dimension of the predictor space.
530 By combining correlated genes, it also reduces the co-linearity among predictors. Additionally, cluster
531 mean is less sensitive to measurement noise and therefore can reduce the impact of measurement error
532 of a gene on the prediction.

533

534 After clustering, the $G \times C$ matrix $\tilde{\mathbf{X}}$ is converted into a $K \times C$ matrix $\bar{\mathbf{X}}$ ($G \approx 10^4$, $K \approx 10^2 \sim 10^3$). The
535 predictor dimension is reduced, but it is still high compared to sample size. Borrowing the idea from the
536 recent high-dimensional regression literature (Fan and Lv 2008), we further reduce the predictor
537 dimension using a fast variable screening procedure: for each DHS locus l , the Pearson’s correlation
538 between its DH signal (i.e., row l of $\tilde{\mathbf{Y}}$) and the expression of each gene cluster k (i.e., row k of $\bar{\mathbf{X}}$) across

539 the training cell types is computed, and the top N ($\approx 10^1$) clusters with the largest correlation
540 coefficients are selected. Using the selected clusters ($\bar{X}_{l_1}, \dots, \bar{X}_{l_N}$) as predictors, a multiple linear
541 regression $\tilde{Y}_l = \beta_{l0} + \beta_{l1}\bar{X}_{l_1} + \dots + \beta_{lN}\bar{X}_{l_N} + \epsilon_l$ is then fitted. Based on the fitted model, the
542 standardized DH level of locus l in a new sample is predicted by $\tilde{Y}_l = f_l(\bar{\mathbf{X}}) = \beta_{l0} + \beta_{l1}\bar{X}_{l_1} + \dots +$
543 $\beta_{lN}\bar{X}_{l_N}$. Of note, although each regression model only contains a small number of predictors, these
544 predictors are selected after examining information from all genes. Therefore, training the prediction
545 model utilizes information from all genes.

546

547 The elementary BIRD model has two parameters: the cluster number K and the predictor number N . In
548 this study, we set $K=1500$ and $N=7$. These parameters were chosen based on testing different values of
549 K and N ($K=100, 200, 500, 1000, 1500, 2000$; $N=1, 2, 3, 4, 5, 6, 7, 8$) using a 5-fold cross-validation
550 conducted within the 40 training cell types (i.e., the same training cell types used for **Figs. 1 and 2**) on a
551 random subset of genomic loci (1% of all DHSs). Since cross-cell-type prediction is more difficult than
552 cross-locus prediction, we identified the optimal parameter combination as the one that maximizes the
553 mean cross-cell-type correlation r_C . **Supplementary Figure 3a** shows that the optimal combination was
554 $K=1500$ and $N=7$. This parameter combination was then used in all subsequent $BIRD_{\bar{X}, Y}$, $BIRD_{\bar{X}, \bar{Y}}$, and
555 compound BIRD models throughout this study.

556

557 In **Supplementary Methods and Supplementary Figure 4**, we compared the elementary BIRD model
558 $BIRD_{\bar{X}, Y}$ with a number of alternative prediction methods including Lasso (Tibshirani 1996), linear
559 regression with stepwise predictor selection (Hocking 1976) (SPS), k-nearest neighbors (Altman 1992)
560 (KNN) and random forest (Breiman 2001) (RF) using 1% of the DHSs obtained from the 40 training cell
561 types. This benchmark analysis shows that the elementary BIRD model not only offers the best
562 prediction accuracy but also is computationally efficient. Based on this result, $BIRD_{\bar{X}, Y}$ was used as the
563 basic building block for subsequent modeling.

564

565 **BIRD_{X,Y} model**

566 If one does not cluster co-expressed genes in the elementary BIRD model, $\text{BIRD}_{\bar{X},Y}$ reduces to $\text{BIRD}_{X,Y}$.
567 In other words, $\text{BIRD}_{X,Y}$ is a special case of $\text{BIRD}_{\bar{X},Y}$ when the gene cluster number K is equal to the
568 gene number G . $\text{BIRD}_{X,Y}$ is not used in the final BIRD compound model. However, in **Figure 1d-f**,
569 $\text{BIRD}_{X,Y}$ and $\text{BIRD}_{\bar{X},Y}$ were compared to study the effect of gene clustering on prediction. $\text{BIRD}_{X,Y}$ only
570 has one parameter: the number of predictors N . Based on 5-fold cross-validation performed on the 40
571 training cell types using 1% of all DHSs from these training cell types, we identified $N = 5$ as the optimal
572 value for $\text{BIRD}_{X,Y}$ (**Supplementary Fig. 3a,b**). $\text{BIRD}_{X,Y}$ based on this optimal N ($N = 5$) was compared to
573 $\text{BIRD}_{\bar{X},Y}$ ($K=1500$ and $N=7$) in **Figure 1d-g**. In **Supplementary Figure 3b**, $\text{BIRD}_{X,Y}$ and $\text{BIRD}_{\bar{X},Y}$ ($K=1500$)
574 were also compared when both methods used the same N . In both comparisons, $\text{BIRD}_{\bar{X},Y}$ consistently
575 outperformed $\text{BIRD}_{X,Y}$.

576

577 **$\text{BIRD}_{\bar{X},Y}$ model**

578 In addition to clustering co-expressed genes, $\text{BIRD}_{\bar{X},Y}$ also groups genomic loci with similar DH patterns
579 into clusters. This is done by clustering rows of the standardized matrix $\tilde{\mathbb{Y}}$ into H clusters using k-means
580 clustering (Euclidean distance used as similarity measure). Based on the clustering result, the DH profile
581 $\tilde{\mathbb{Y}}$ of each sample can be converted into a lower dimensional vector $\bar{\mathbb{Y}} = (\bar{Y}_1, \dots, \bar{Y}_H)$, where \bar{Y}_h is the
582 mean DH level of DHSs in cluster h . Instead of predicting the DH level $\tilde{\mathbb{Y}}$ of individual loci, $\text{BIRD}_{\bar{X},Y}$ uses
583 the cluster-level gene expression \bar{X} to predict cluster-level DH $\bar{\mathbb{Y}}$. The prediction models are constructed
584 using linear regression in a way similar to how the regression models are constructed in $\text{BIRD}_{\bar{X},Y}$. In
585 **Figure 2b**, comparisons between $\text{BIRD}_{\bar{X},Y}$ and $\text{BIRD}_{\bar{X},Y}$ was used to illustrate cluster-level DH can be
586 predicted with higher accuracy than DH at individual genomic loci. The same parameter combination
587 $K=1500$ and $N=7$ was set for both $\text{BIRD}_{\bar{X},Y}$ and $\text{BIRD}_{\bar{X},Y}$. For $\text{BIRD}_{\bar{X},Y}$, H was set to 1000, 2000 and 5000
588 respectively.

589

590 **BIRD – The compound BIRD model**

591 $\text{BIRD}_{\bar{X},Y}$ is a special case of $\text{BIRD}_{\bar{X},Y}$ when DHSs are not clustered (i.e., $H = L$). Compared to $\text{BIRD}_{\bar{X},Y}$,
592 the increased accuracy of cluster-level prediction by $\text{BIRD}_{\bar{X},Y}$ is partly because a cluster's mean DH is

593 usually associated with smaller variance of measurement noise than the DH level of individual loci. In
594 BIRD_{X̄,Ȳ}, one may use the predicted cluster mean as the predicted DH level of each individual locus
595 within the cluster. This will also generate a prediction for each locus. This locus-level prediction may be
596 biased, but it is usually associated with smaller variance. By contrast, predictions by BIRD_{X̄,Y} for each
597 locus may be less biased but has larger variance. This motivates the compound BIRD model.

598
599 In the compound BIRD model, multiple BIRD_{X̄,Ȳ} models with different H values are combined through
600 model averaging, a useful technique to improve prediction accuracy by balancing the variance and bias.
601 Consider making predictions for a sample. Let \mathcal{H} be the set of H values used by BIRD_{X̄,Ȳ}. $\mathcal{H} =$
602 $\{1000, 2000, 5000, L\}$ in this study. For each DHS locus l , let $\hat{Y}_l^{(H)}$ denote the locus-level DH predicted
603 by BIRD_{X̄,Ȳ} using cluster number H . $\hat{Y}_l^{(L)}$ is the locus-level DH predicted by BIRD_{X̄,Y}. The compound
604 BIRD model predicts the locus-level DH for locus l using a weighted average

$$\frac{\sum_{H \in \mathcal{H}} d_l^H \hat{Y}_l^{(H)}}{\sum_{H \in \mathcal{H}} d_l^H}$$

605 where d_l^H is the weight. For a given cluster number H , the weight d_l^H is determined using training data
606 as follows. Let $\tilde{\mathbf{y}}_l = (\tilde{y}_{l1}, \dots, \tilde{y}_{lM})$ be the standardized locus-level DH for locus l observed in M training
607 cell types. Each locus l is associated with a cluster. Let $\tilde{\mathbf{y}}_l^{(H)} = (\tilde{y}_{l1}^{(H)}, \dots, \tilde{y}_{lM}^{(H)})$ represent the average of
608 the standardized DH level of all loci within the cluster corresponding to locus l in the M training cell
609 types. Define d_l^H as the Pearson's correlation between the two vectors $\tilde{\mathbf{y}}_l^{(H)}$ and $\tilde{\mathbf{y}}_l$. Note that when
610 $H = L$, BIRD_{X̄,Ȳ} reduces to BIRD_{X̄,Y}, and we have $\tilde{\mathbf{y}}_l^{(L)} = \tilde{\mathbf{y}}_l$ and $d_l^L = 1$. Thus the weight for BIRD_{X̄,Y}
611 is 1.
612

613
614 Comparisons between the compound BIRD model (referred to as "BIRD") and BIRD_{X̄,Y} in **Figure 1d-g**
615 show that BIRD outperforms BIRD_{X̄,Y}. Therefore, the compound BIRD model was used as our final
616 prediction model, and it was used for predicting TFBS, constructing PDDB, and improving DNase-seq and
617 ChIP-seq data analyses.

618

619 **Random prediction models by permutation**

620 To construct random prediction models, we permuted the cell type labels of DNase-seq data in the
621 training dataset. This permutation broke the connection between DNase-seq and gene expression data.
622 BIRD was then trained using the permuted training dataset, and the trained model was applied to
623 predict DH in the test dataset. The permutation was performed 10 times. The prediction performance r_L ,
624 r_C and τ were computed for each permutation. The average values of these three statistics from the 10
625 permutations were used to represent the prediction performance of random prediction models.

626

627 **Wilcoxon signed-rank test for comparing different methods**

628 In order to generate **Figure 1g**, two-sided Wilcoxon signed-rank test was performed to obtain p -values
629 for comparing prediction accuracy of each pair of methods. For instance, in order to test whether two
630 methods A and B perform equally in terms of r_L , the paired r_L values from these two methods for each
631 cell type was obtained. Then the r_L pairs from all cell types are used for the Wilcoxon signed-rank test.
632 Similarly, to compare methods A and B in terms of r_C , the paired r_C values for each locus was obtained,
633 and r_C pairs from all genomic loci were used for the Wilcoxon signed-rank test.

634

635 **Categorization of test genomic loci when studying cross-cell-type correlation**

636 When studying the cross-cell-type prediction performance (i.e., r_C) of BIRD in **Figure 2c-e**, genomic loci
637 were grouped into different categories based on their DH profile in the test cell types. First, because
638 test cell types were not used to select genomic loci, a subset of selected genomic loci may not contain
639 strong or meaningful DH signal in any test cell type. For such loci, the cross-cell-type correlation
640 between the predicted and true DH signals (which are essentially noise) is expected to be low. For this
641 reason, we identified DHSs with predicted DH level (log2 transformed) smaller than 2 in all 17 test cell
642 types and labeled them as “noisy loci” (**Fig. 2c**). After excluding the noisy loci, the other loci were then
643 categorized based on the coefficient of variation (CV) of the cross-cell-type DH values. For each locus, CV
644 was calculated as the ratio of the standard deviation to mean of the predicted DH at this locus across all
645 test cell types. Loci were divided into three categories: $CV \leq 0.2$, $0.2 < CV \leq 0.4$, $CV > 0.4$ (**Fig. 2c**). A large CV

646 indicates that the DH of a locus has more variation across cell types. **Figure 2c** shows the distribution of
647 r_c . Genomic loci are grouped into bins based on r_c values. For each bin, the number of loci in different CV
648 categories is shown. **Figure 2d** shows the percentage of loci in different CV categories for each r_c bin.
649 **Figure 2e** shows distribution of r_c values for each CV category.

650
651 We also computed CV using the true DH values from the test DNase-seq data rather than predicted DH
652 values. The results that loci with large r_c also tend to have large CV remain qualitatively the same
653 (**Supplementary Fig. 6**). In practice, however, since BIRD is typically used when DNase-seq data are not
654 available, one can only use CV based on predicted DH values.

655
656 **The Predicted DNase I hypersensitivity database (PDDB)**
657 PDDB is available at <http://jilab.biostat.jhsph.edu/~bsherwo2/bird/index.php>. Details on database
658 construction and use are provided in **Supplementary Methods**.

659
660 **Software**
661 BIRD software is available at <https://github.com/WeiqiangZhou/BIRD>. Models trained using the 57
662 ENCODE cell types have been stored in the software package. With these pre-compiled prediction
663 models, making predictions on new samples provided by users is computationally fast. On a computer
664 with 2.5 GHz CPU and 10Gb RAM, it took less than 2 minutes to make predictions for ~1 million DHSs in
665 100 samples.

666
667 **Other data analysis protocols**
668 Procedures for comparing BIRD with other prediction methods, TFBS prediction, MYC, SOX2 and MEF2A
669 analyses using PDDB, and improving DNase-seq and ChIP-seq signals are provided in **Supplementary**
670 **Methods.**

671
672

673 **ACKNOWLEDGMENTS**

674 The authors would like to thank Drs. X. Shirley Liu and Yingying Wei for insightful discussions. This
675 research is supported by grants from the Maryland Stem Cell Research Fund (2012-MSCRFE-0135-00)
676 and the National Institutes of Health (R01HG006282 and R01HG006841).

677

678 **FIGURE LEGENDS**

679 **Figure 1.** BIRD – concepts and evaluation.

680 **(a)** Outline of the study. ENCODE DNase-seq and exon array data are used to train BIRD. Users can apply
681 BIRD to new or existing gene expression samples to predict DH. The predicted DH can be used to predict
682 TFBSSs, convert expression samples in GEO into a regulome database (PDDB), and improve DNase-seq
683 and ChIP-seq data analyses.

684 **(b)** Overview of BIRD. Instead of using expression of individual genes as predictors to predict DH at each
685 locus ($BIRD_{X,Y}$), BIRD first groups co-expressed genes into clusters (i.e., gene-cluster) and uses the
686 clusters' mean expression levels as predictors to predict the DH level at each genomic locus ($BIRD_{\bar{X},Y}$).
687 Additionally, BIRD also groups correlated loci (i.e., loci with co-varying DH) into different levels of
688 clusters (i.e., DHS-cluster) and predicts the DH level for each cluster ($BIRD_{\bar{X},\bar{Y}}$). Finally, BIRD combines
689 the locus-level and cluster-level predictions via model averaging.

690 **(c)** Statistics used to evaluate prediction accuracy.

691 **(d)** Cross-locus P-T correlation (r_L) for different methods. For each method, the boxplot and number
692 show the distribution and mean of r_L from the 17 test cell types.

693 **(e)** Cross-cell-type P-T correlation (r_C) for different methods. For each method, the distribution and
694 mean of r_C from the 912,886 genomic loci are shown.

695 **(f)** Squared prediction error (τ) for different methods.

696 **(g)** Two-sided Wilcoxon signed-rank test p-values for comparing prediction performance (r_L or r_C) of
697 different methods. We did not perform similar test for the squared prediction error (τ) since there is
698 only one τ for each method.

699

700 **Figure 2.** Cross-cell-type prediction performance and a comparison with ChromImpute.

701 (a) An example of true and predicted DNase-seq signals for five different cell types. “True”: true DNase-
702 seq bin read count; “BIRD”: DH signal predicted by BIRD; “Mean”: the average DH signal of training cell
703 types. For “BIRD” and “Mean”, signals are transformed back from the log-scale to the original scale.

704 (b) Comparison between locus-level and cluster-level predictions in terms of cross-cell-type prediction
705 accuracy. For each method, distribution and mean of r_c of all genomic loci or pathways (i.e., DHS clusters)
706 are shown. DHSs were clustered into 1000, 2000 and 5000 clusters in $\text{BIRD}_{\bar{X}, \bar{Y}_{1000}}$, $\text{BIRD}_{\bar{X}, \bar{Y}_{2000}}$ and
707 $\text{BIRD}_{\bar{X}, \bar{Y}_{5000}}$ respectively.

708 (c) Locus-level cross-cell-type prediction accuracy by BIRD. Genomic loci were grouped into four
709 categories based on the coefficient of variation (CV) of the predicted DH values in 17 test cell types. The
710 histogram shows the distribution of r_c of all loci, stratified based on the four CV categories.

711 (d) Loci are grouped into bins based on the cross-cell-type prediction accuracy r_c . For each r_c bin, the
712 percentage of loci in each CV category is shown.

713 (e) Distribution of r_c for loci in each CV category.

714 (f) Comparison between BIRD and ChromImpute. Cross-locus P-T correlation r_L in 10 test cell types
715 analyzed by both methods are shown. As a baseline, predictions based on the mean DH profile of
716 training cell types are also shown.

717

718 **Figure 3.** Predicting transcription factor binding sites.

719 (a)-(b) Sensitivity-rank curve for predicting MAX and ELF1 binding sites in GM12878 using three different
720 methods: true DNase-seq data (“True”), BIRD, and mean DH profile of training cell types (“Mean”). For
721 BIRD, “BIRD(UW)”, “BIRD(Duke)” and “BIRD(Chicago)” denote predictions made using exon arrays
722 generated by three different labs. For each method, the sensitivity-rank curve shows the percentage of
723 true TFBSS that were discovered by top predicted motif sites. The q -values corresponding to top 5000,
724 15000, and 25000 predictions are shown on top of each plot. q -values from BIRD(UW) are shown, and q -
725 values from the other two labs were similar and therefore not displayed for clarity.

726 (c) Sensitivity-rank curve for predicting GABPA binding sites in K562. BIRD predictions were generated
727 using exon arrays from three different labs. *q*-values from BIRD(UW) are shown.
728 (d) Sensitivity-rank curve for predicting MYC binding sites in P493-6 using BIRD and the mean DH profile
729 of training cell types. *q*-values from BIRD are shown.
730 (e) True MYC ChIP-seq signal (log2 read count) in P493-6 versus BIRD predicted DH at all E-box motif
731 sites. The correlation was high ($r_L=0.70$).
732 (f) True MYC ChIP-seq signal in P493-6 versus mean DH of training cell types at all E-box motif sites. The
733 correlation was low ($r_L=0.48$).
734 (g) Examples showing true MYC ChIP-seq signal (read count, blue) in P493-6 and the predicted DH signal
735 by BIRD (red) and “Mean” (brown). BIRD more accurately captured the true signal than “Mean”
736 (highlighted with red boxes).

737
738 **Figure 4.** The predicted DNase I hypersensitivity database (PDDB).
739 (a) Flowchart illustrating how to use PDDB. Step 1: provide a list of genomic loci of interest. Step 2:
740 provide keywords in one or multiple annotation fields (e.g., type “stem cell” in the “Cell Type” column)
741 to search for samples of interest. PDDB will return predicted DH for the queried loci and samples along
742 with sample annotation and data for visualization.
743 (b) Web interface of PDDB. Users can download the predicted DH data by clicking the “Download
744 DNase-seq” button. The sample annotation data can be downloaded by clicking the “Get Annotation
745 Data” button.
746 (c) By clicking the “Visualization of Predicted DNase-Seq data in UCSC Browser” link in the PDDB web
747 interface (red circle in b), one can display the predicted DH signal in the UCSC genome browser.

748
749 **Figure 5.** Predicting regulome using PDDB.
750 (a)-(c) Predicted DH in promoter regions of FBL (a), LIN28A (b) and BLMH (c) in P493-6 B cell lymphoma
751 and H9 embryonic stem cells. For H9, PDDB contains multiple replicate samples which produced similar
752 results. One replicate is shown here and the other replicates are shown in **Supplementary Fig. 10**.

753 (d) Predicted DH at SOX2 binding sites in 2,000 PDDB samples. Each column is a sample, and each row is
754 a binding site. Values within each row are standardized to have zero mean and unit SD before
755 visualization.

756 (e) Relative DH enrichment level when comparing SOX2 binding sites with random sites (**Supplementary**
757 **Methods**).

758 (f) Predicted DH at SOX2 binding sites in H7 stem cells after 2, 5 and 9 days of differentiation. True DH
759 from undifferentiated H7 cells and cells at differentiating day 14 in the training data are also shown.
760 Rows are SOX2 sites and columns are time points. Values within each row are standardized before
761 visualization.

762 (g) Predicted DH at SOX2 binding sites are compared with predicted DH at random DHSs. At each time
763 point, DH values from all sites are displayed using a boxplot.

764 (h) Predicted DH at 2,011 MEF2A binding sites in 1,061 MEF2A-expressing PDDB samples
765 (**Supplementary Methods**). Each column is a sample. Each row is a MEF2A binding site. Values within
766 each row are standardized before visualization. Samples and DHSs were clustered.

767 (i) The highlighted region in (h) which shows DHS-clusters with increased DH in muscle, lymphoblastoid,
768 and brain related samples, respectively.

769

770 **Figure 6.** BIRD predictions used as pseudo-replicates to improve DNase-seq and ChIP-seq data analyses.
771 The observed signal from one sample (“obs-only”) in a test cell type was combined with BIRD predictions
772 to produce the integrated signal (“BIRD+obs”). Signals before and after integration are compared with
773 the observed signal from another sample from the same cell type (“truth”).

774 (a)-(c) DNase-seq.

775 (a) Correlation (r) between the “truth” and “BIRD+obs” (i.e., the integrated signal) in GM12878. Each dot
776 is a genomic locus.

777 (b) Correlation between the “truth” and “obs-only” (i.e., the original signal without integrating BIRD) in
778 GM12878.

779 (c) The same analyses were done for 16 test cell types. Red dots in the scatterplot compares the P-T
780 correlation r for BIRD+obs vs. r for obs-only in the 16 cell types. BIRD+obs outperformed obs-only in 12
781 of 16 test cell types. As a control, BIRD was replaced by the mean DH profile of training cell types. Blue
782 dots show the P-T correlation r for Mean+obs vs. r for obs-only in the 16 test cell types. Mean+obs did
783 not improve over obs-only.

784 (d)-(f) ChIP-seq.

785 (d) Correlation between the “truth” and “BIRD+obs” for SP1 in GM12878 at SP1 motif sites.

786 (e) Correlation between the “truth” and “obs-only” for SP1 in GM12878 at SP1 motif sites.

787 (f) The same analyses were done for 9 TFs in GM12878 (circles) and 3 TFs in K562 (triangles). Once again,
788 BIRD+obs outperformed obs-only in 11 of 12 cases (red and green), but Mean+obs did not improve over
789 obs-only (blue and yellow)

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

Figure 1. BIRD – concepts and evaluation.

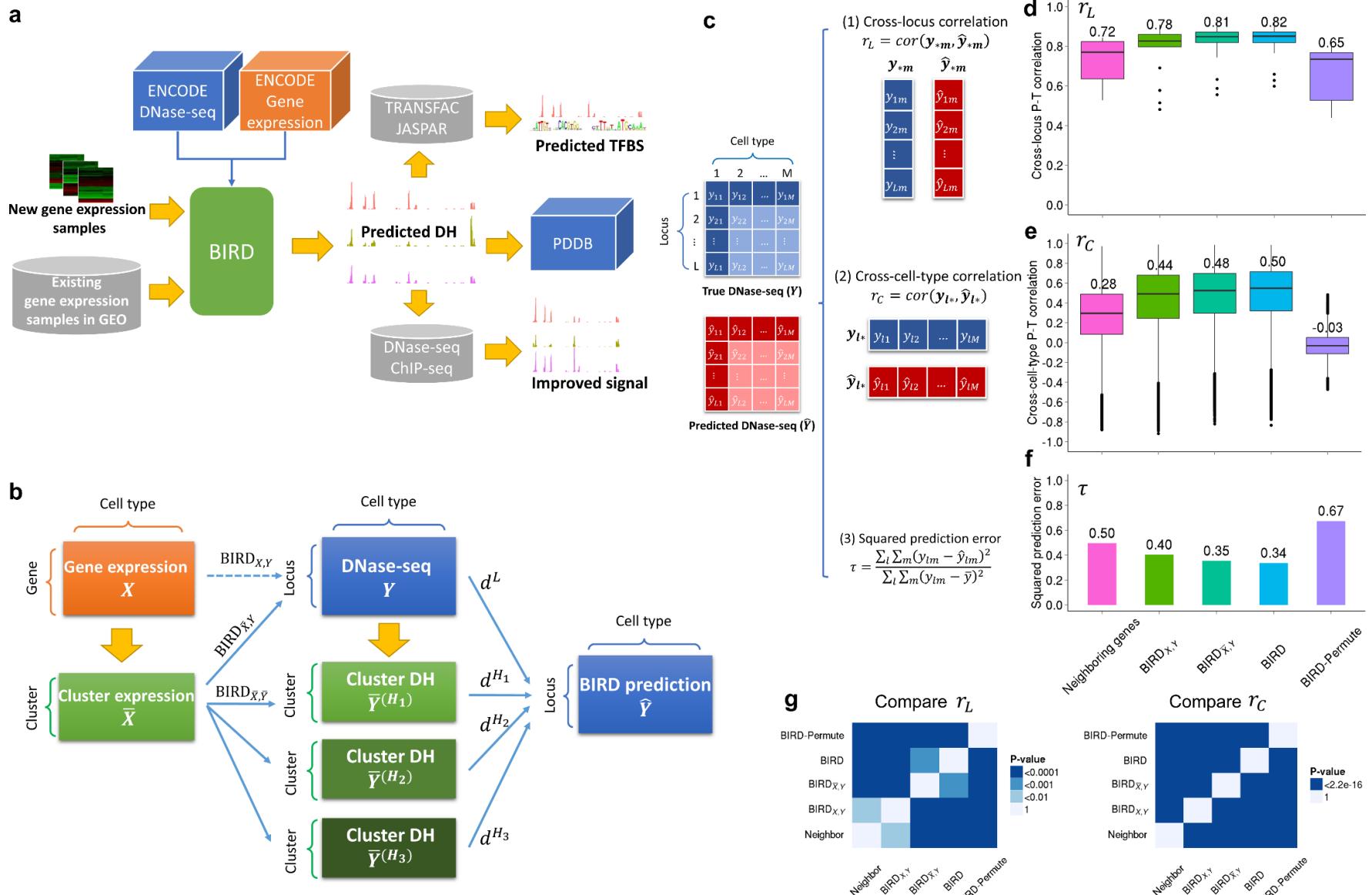


Figure 2. Cross-cell-type prediction performance and a comparison with ChromImpute.

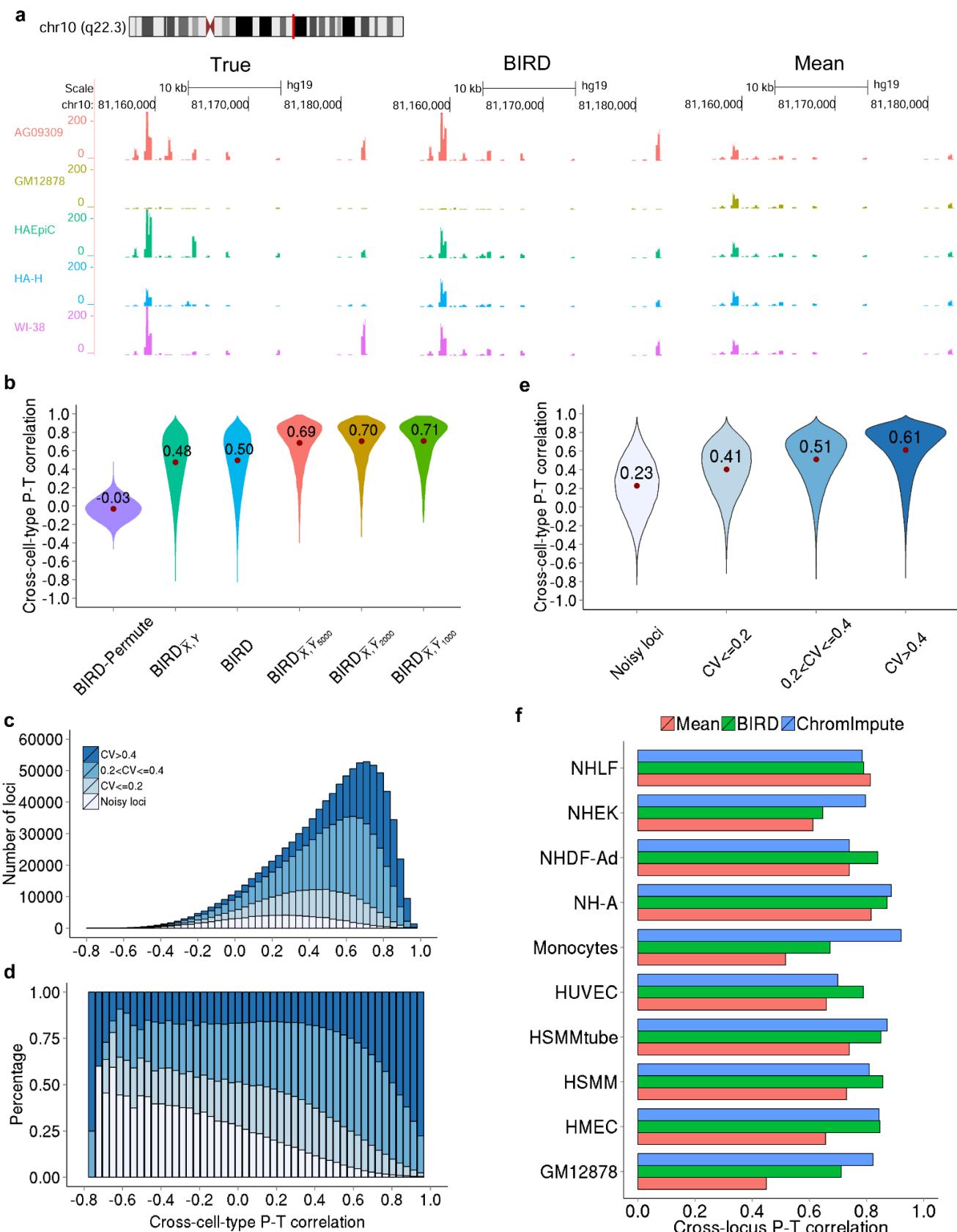


Figure 3. Predicting transcription factor binding sites.

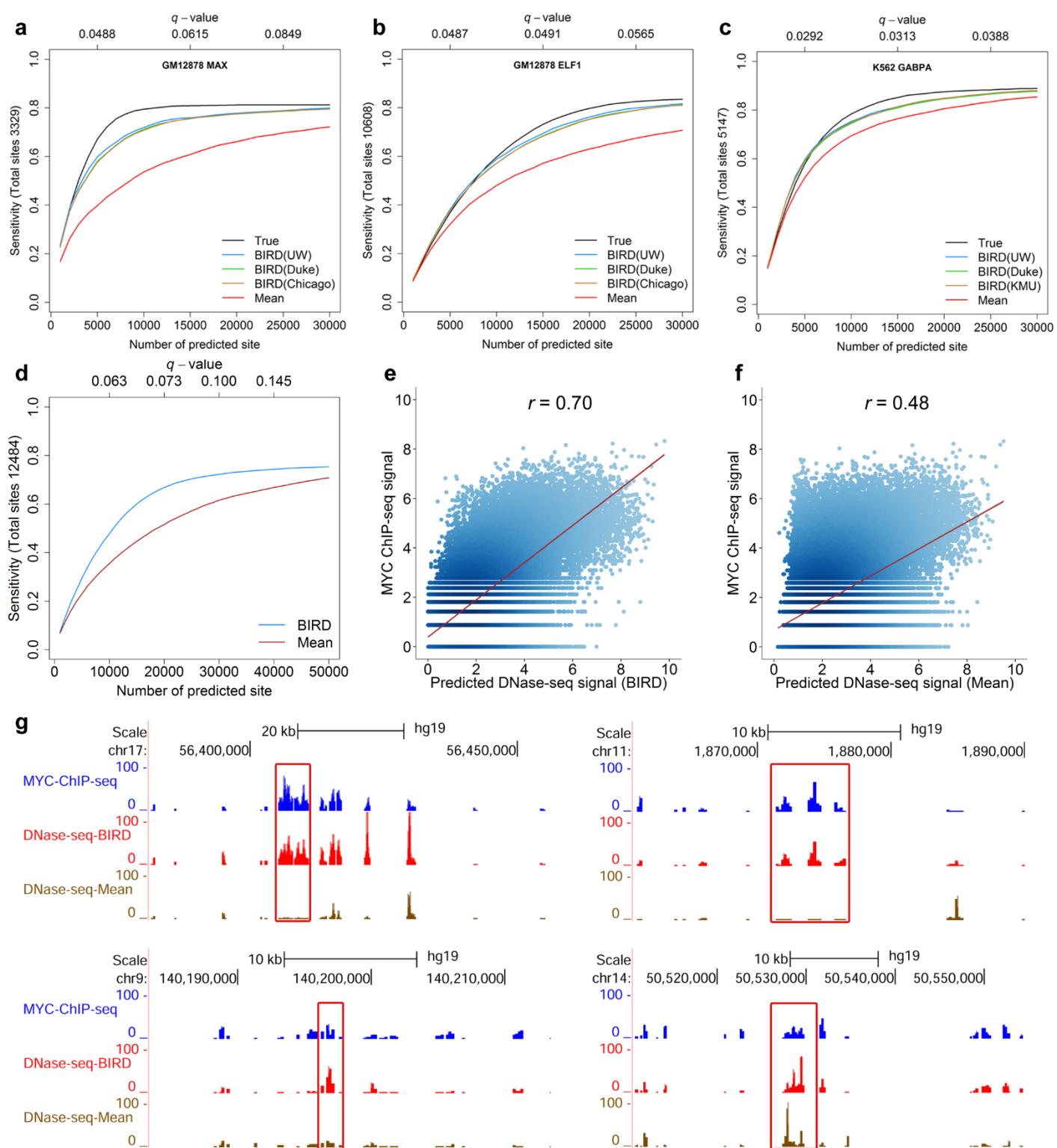


Figure 4. The predicted DNase I hypersensitivity database (PDDB).

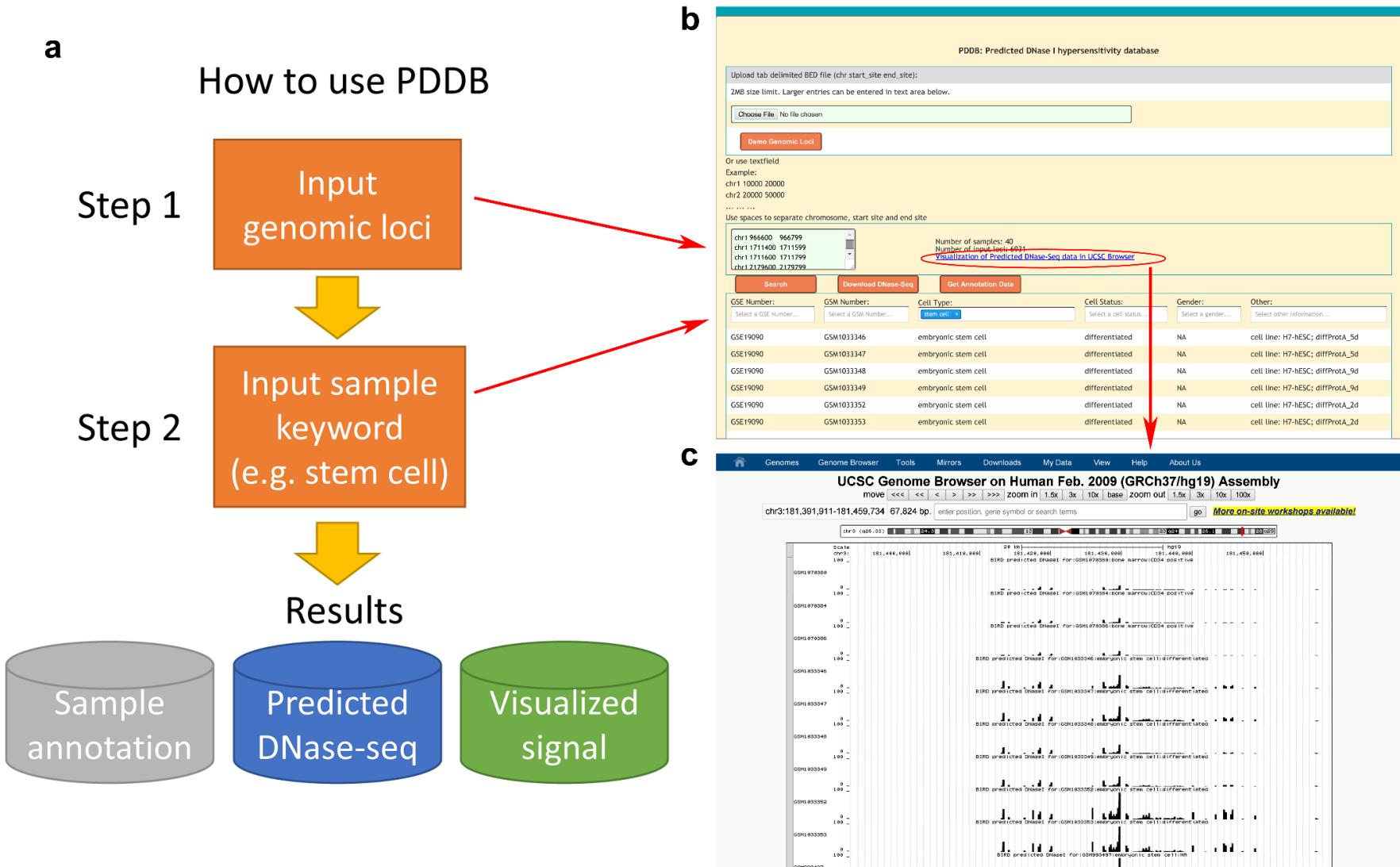


Figure 5. Predicting regulome using PDDB.

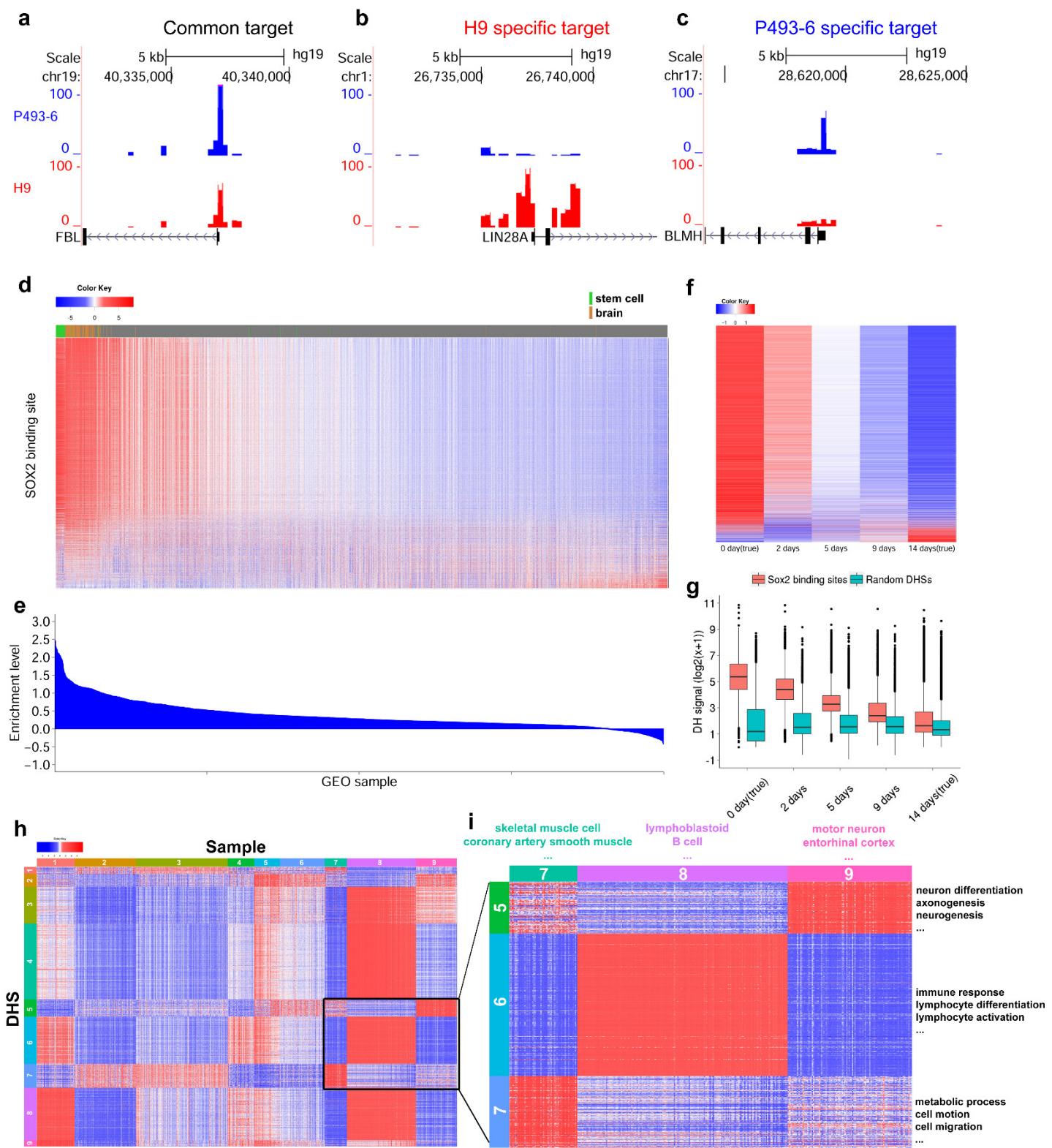
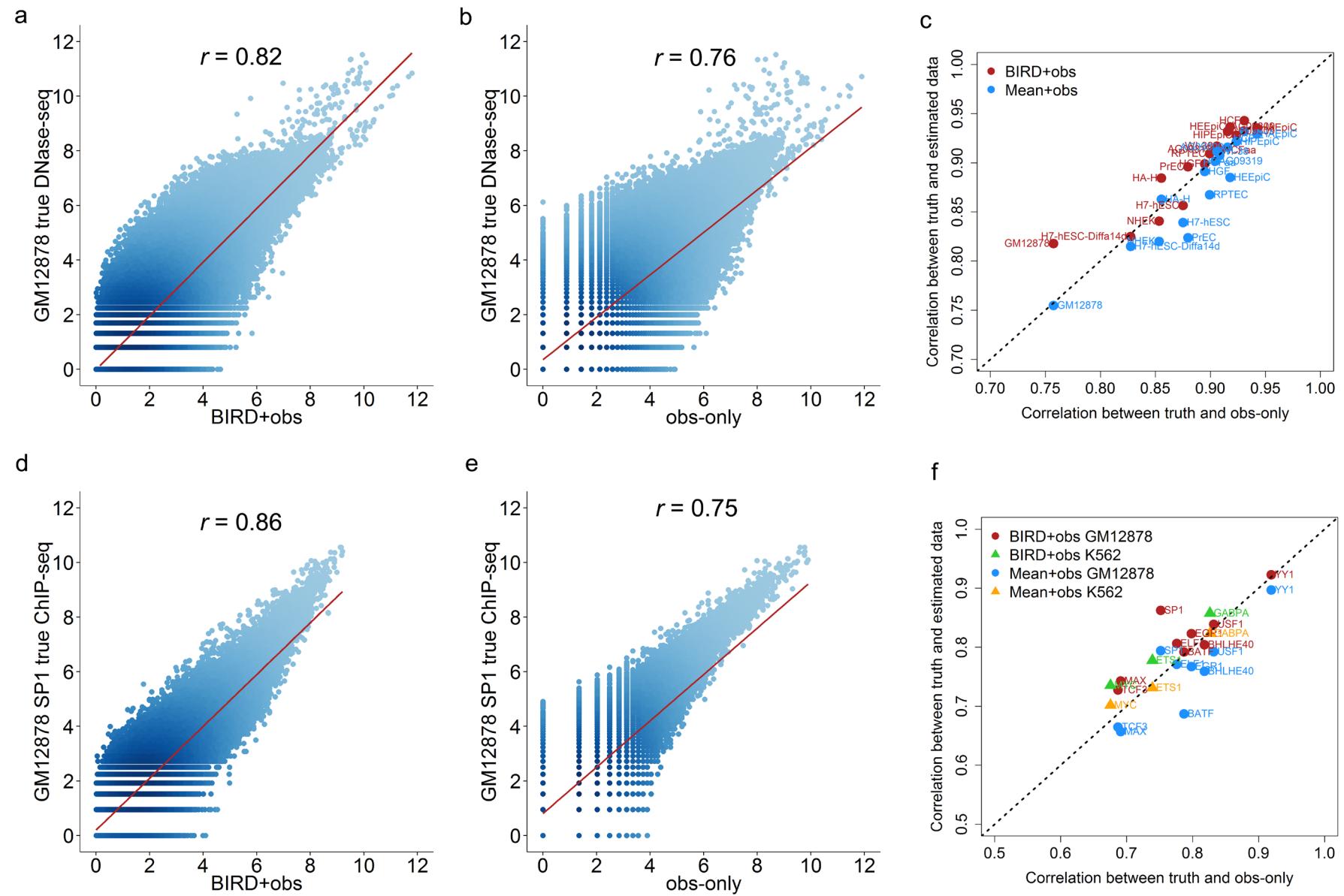


Figure 6. BIRD predictions used as pseudo-replicates to improve DNase-seq and ChIP-seq data analyses.



REFERENCES

- Altman NS. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46**: 175-185.
- Bolstad BM. 2015. preprocessCore: A collection of pre-processing functions. *R package version 1.28.0*.
- Breiman L. 2001. Random forests. *Mach Learning* **45**: 5-32.
- Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**: 1213-1218.
- Chambers I, Tomlinson SR. 2009. The transcriptional foundation of pluripotency. *Development* **136**: 2311-2322.
- Chang TC, Zeitels LR, Hwang HW, Chivukula RR, Wentzel EA, Dews M, Jung J, Gao P, Dang CV, Beer MA, et al. 2009. Lin-28B transactivation is necessary for Myc-mediated let-7 repression and proliferation. *Proc Natl Acad Sci U S A* **106**: 3384-3389.
- Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan KK, Dong X, Djebali S, Ruan Y, et al. 2012. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res* **22**: 1658-1667.
- Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**: 123-131.
- Dabney A, Storey JD. . qvalue: Q-value estimation for false discovery rate control. *R package version 1.40.0*.
- Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**: 207-210.
- Edmondson DG, Lyons GE, Martin JF, Olson EN. 1994. Mef2 gene expression marks the cardiac and skeletal muscle lineages during mouse embryogenesis. *Development* **120**: 1251-1263.
- ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

- Ernst J, Kellis M. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol.*
- Fan J, Lv J. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**: 849-911.
- Ferri AL, Cavallaro M, Braida D, Di Cristofano A, Canta A, Vezzani A, Ottolenghi S, Pandolfi PP, Sala M, DeBiasi S, et al. 2004. Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. *Development* **131**: 3805-3819.
- Flavell SW, Kim T, Gray JM, Harmin DA, Hemberg M, Hong EJ, Markenscoff-Papadimitriou E, Bear DM, Greenberg ME. 2008. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron* **60**: 1022-1038.
- Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**: 877-885.
- Hartigan JA, Wong MA. 1979. Algorithm AS 136: A k-means clustering algorithm. *Applied statistics*: 100-108.
- Hastie T. 2015. gam: Generalized Additive Models. *R package version 1.12*.
- Hocking RR. 1976. A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*: 1-49.
- Huang DW, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**: 1-13.
- Huang DW, Sherman BT, Lempicki RA. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**: 44-57.
- Hystad ME, Myklebust JH, Bo TH, Sivertsen EA, Rian E, Forfang L, Munthe E, Rosenwald A, Chiorazzi M, Jonassen I, et al. 2007. Characterization of early stages of human B cell development by gene expression profiling. *J Immunol* **179**: 3662-3671.

- Ji H, Wu G, Zhan X, Nolan A, Koh C, De Marzo A, Doan HM, Fan J, Cheadle C, Fallahi M. 2011. Cell-type independent MYC target genes reveal a primordial signature involved in biomass accumulation. *PloS one* **6**: e26057.
- Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* **26**: 1293-1300.
- Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497-1502.
- Kapur K, Xing Y, Ouyang Z, Wong WH. 2007. Exon arrays provide accurate assessments of gene expression. *Genome Biol* **8**: R82.
- Koh CM, Gurel B, Sutcliffe S, Aryee MJ, Schultz D, Iwata T, Uemura M, Zeller KI, Anele U, Zheng Q. 2011. Alterations in nucleolar structure and gene expression programs in prostatic neoplasia are driven by the MYC oncogene. *The American journal of pathology* **178**: 1824-1834.
- Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, Prabhakar S. 2013. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotechnol* **31**: 615-622.
- Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ilenasescu H, et al. 2014. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**: D142-7.
- Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**: D108-10.
- Natarajan A, Yardımcı GG, Sheffield NC, Crawford GE, Ohler U. 2012. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res* **22**: 1711-1722.

- Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. 2012. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**: 1274-1286.
- Phi JH, Park SH, Kim SK, Paek SH, Kim JH, Lee YJ, Cho BK, Park CK, Lee DH, Wang KC. 2008. Sox2 expression in brain tumors: a reflection of the neuroglial differentiation pathway. *Am J Surg Pathol* **32**: 103-112.
- Potthoff MJ, Olson EN. 2007. MEF2: a central regulator of diverse developmental programs. *Development* **134**: 4131-4140.
- Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, et al. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**: 1003-1005.
- Sabò A, Kress TR, Pelizzola M, de Pretis S, Gorski MM, Tesi A, Morelli MJ, Bora P, Doni M, Verrecchia A. 2014. Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. *Nature* **511**: 488-492.
- Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* **23**: 777-788.
- Storey JD. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**: 479-498.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**: 15545-15550.
- Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**: 663-676.
- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75-82.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.Series B (Methodological)*: 267-288.

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

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh T, Peng W, Zhang MQ. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40**: 897-903.

Watanabe H, Ma Q, Peng S, Adelman G, Swain D, Song W, Fox C, Francis JM, Pedamallu CS, DeLuca DS. 2014. SOX2 and p63 colocalize at genetic loci in squamous cell carcinomas. *J Clin Invest* **124**: 0-0.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137-2008-9-9-r137. Epub 2008 Sep 17.

Zhou W, Ji Z, Ji H. submitted. Global Prediction of Chromatin Accessibility Using RNA-seq from Small Number of Cells.