

1           **Global Prediction of Chromatin Accessibility Using RNA-seq from**  
2           **Small Number of Cells**

3                           Weiqiang Zhou<sup>1</sup>, Zhicheng Ji<sup>1</sup>, Hongkai Ji<sup>1,\*</sup>

4           <sup>1</sup>Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, 615 North  
5           Wolfe Street, Baltimore, MD 21205, USA

6           \*To whom correspondence should be addressed: [hji@jhu.edu](mailto:hji@jhu.edu)

7

8           Corresponding author:

9           Hongkai Ji, Ph.D.

10          Department of Biostatistics

11          Johns Hopkins Bloomberg School of Public Health

12          615 N Wolfe Street, Rm E3638

13          Baltimore, MD 21205, USA

14          Email: [hji@jhu.edu](mailto:hji@jhu.edu)

15          Phone: 410-955-3517

16

17          Running title:

18          Predicting DH with small number of cells

19

20          Keywords:

21          DNase I hypersensitivity, Gene expression, single cell, RNA-seq, DNase-seq

22

23

24

25

26

27

28    **ABSTRACT**

29    Conventional high-throughput technologies for mapping regulatory element activities such as ChIP-seq,  
30    DNase-seq and FAIRE-seq cannot analyze samples with small number of cells. The recently developed  
31    ATAC-seq allows regulome mapping in small-cell-number samples, but its signal in single cell or samples  
32    with  $\leq 500$  cells remains discrete or noisy. Compared to these technologies, measuring transcriptome by  
33    RNA-seq in single-cell and small-cell-number samples is more mature. Here we show that one can  
34    globally predict chromatin accessibility and infer regulome using RNA-seq. Genome-wide chromatin  
35    accessibility predicted by RNA-seq from 30 cells is comparable with ATAC-seq from 500 cells. Predictions  
36    based on single-cell RNA-seq can more accurately reconstruct bulk chromatin accessibility than using  
37    single-cell ATAC-seq by pooling the same number of cells. Integrating ATAC-seq with predictions from  
38    RNA-seq increases power of both methods. Thus, transcriptome-based prediction can provide a new  
39    tool for decoding gene regulatory programs in small-cell-number samples.

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55 **INTRODUCTION**

56 Decoding gene regulatory network in developmental systems, precious clinical samples, and purified  
57 cells often requires measuring transcriptome (i.e., genes' transcriptional activities) and regulome (i.e.,  
58 regulatory element activities) in samples with small number of cells. While significant progress has been  
59 made to measure transcriptome in single cell (Tang et al. 2010; Ramsköld et al. 2012) and in small-cell-  
60 number (Marinov et al. 2014) samples using RNA sequencing (RNA-seq), accurately measuring regulome  
61 in single-cell and small-cell-number samples remains a challenge. Conventional high-throughput  
62 technologies such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Johnson et al.  
63 2007), sequencing of DNase I hypersensitive sites (DNase-seq) (Crawford et al. 2006), and  
64 Formaldehyde-Assisted Isolation of Regulatory Elements coupled with sequencing (FAIRE-seq) (Giresi et  
65 al. 2007) require large amounts of input material ( $\sim 10^6$  cells). They cannot analyze samples with small  
66 number of cells. The state-of-the-art technology ATAC-seq – assay for transposase-accessible chromatin  
67 using sequencing – can analyze chromatin accessibility in bulk samples with 500-50,000 cells  
68 (Buenrostro et al. 2013). However, ATAC-seq data are noisy when the cell number is small (e.g.,  $\leq 500$ ).  
69 Most recently, single-cell ATAC-seq (Cusanovich et al. 2015; Buenrostro et al. 2015) (scATAC-seq) has  
70 been invented to analyze individual cells. Nevertheless, signals from scATAC-seq are intrinsically discrete  
71 since each genomic locus only has up to two copies of chromatin that can be assayed within a cell, and  
72 scATAC-seq only provides a snapshot of chromatin accessibility of a cell at the time when it is assayed  
73 and destroyed. As a surrogate for regulatory element activity, chromatin accessibility is arguably a  
74 continuous signal. This is because molecular events such as transcription-factor-DNA binding and  
75 dissociation are stochastic over time, and the overall activity of a regulatory element in a cell is  
76 determined by the probability – a continuous measure – that such stochastic events occur if one were to  
77 repeatedly observe the same cell at random time points. The discrete signal measured by scATAC-seq at  
78 a single time point cannot accurately describe this continuum of chromatin accessibility (**Supplementary**  
79 **Fig. 1**). Parallel to ATAC-seq, microfluidic oscillatory washing-based ChIP-seq (MOWChIP-seq) is a  
80 recently developed method for measuring histone modifications in small-cell-number samples (100-600  
81 cells) (Cao et al. 2015). Similar to ATAC-seq, MOWChIP-seq remains noisy when the cell number is small.

82

83 In a companion study, we found that chromatin accessibility measured by DNase I hypersensitivity (DH)  
84 in a bulk sample can be predicted with good accuracy using the sample's gene expression profile  
85 measured by Affymetrix exon array (Zhou et al. submitted). We also developed a computational method  
86 BIRD to handle this big data prediction problem (**Supplementary Methods, Supplementary Fig. 2**). Here  
87 we investigate whether one can use a similar approach to predict regulome based on RNA-seq, and  
88 importantly, whether this approach allows one to use small-cell-number and single-cell RNA-seq (scRNA-  
89 seq) to predict regulome in samples with limited amounts of materials (**Fig. 1a**).

90

## 91 RESULTS

### 92 Predicting Chromatin Accessibility Using Bulk RNA-seq

93 We begin with evaluating the feasibility of using bulk RNA-seq to predict DH. We downloaded DNase-  
94 seq and matching RNA-seq data for 70 human samples representing 30 different cell types  
95 (**Supplementary Table 1**) from the Roadmap Epigenomics project (Kundaje et al. 2015) (also called  
96 Epigenome Roadmap below). After preprocessing and normalization, 37,335 transcripts with expression  
97 measurements from RNA-seq and 1,136,465 genomic loci with DH measurements from DNase-seq were  
98 obtained and served as predictors and responses respectively (**Methods**). Our goal is to predict DH at  
99 these 1,136,465 loci using the 37,335 predictors. We evaluate the prediction using leave-one-out cross-  
100 validation. In each fold of the cross-validation, the 30 cell types were partitioned into a training dataset  
101 consisting of 29 cell types and a test dataset consisting of 1 cell type. BIRD prediction models were  
102 trained using samples in the training data and then applied to RNA-seq samples in the test data to  
103 predict DH (**Methods**). Prediction performance was evaluated using true DNase-seq signals in the test  
104 data and the following statistics (**Fig. 1c**): (1) Pearson correlation between the predicted and true DH  
105 values across all genomic loci within each sample ("cross-locus correlation"  $r_L$ ), (2) Pearson correlation  
106 between the predicted and true DH values across all samples at each genomic locus ("cross-sample  
107 correlation"  $r_C$ ), and (3) total squared prediction error scaled by the total DH data variance ( $\tau$ ). As a  
108 control, we also constructed random prediction models ("BIRD-Permute") by permuting the link

109 between the DNase-seq and RNA-seq samples in the training data and then applied them to the test  
110 data.

111

112 Based on  $r_L$ , RNA-seq was able to accurately predict how DH varied across different genomic loci (mean  
113  $r_L=0.87$ ), and the prediction accuracy of BIRD was significantly higher than random expectation (**Fig. 1d**,  
114 BIRD vs. BIRD-Permute: two-sided Wilcoxon signed-rank test  $p$ -value =  $3.6 \times 10^{-13}$ ). Of note, BIRD-Permute  
115 also explained a large amount of cross-locus DH variation (**Fig. 1d**, mean  $r_L = 0.70$ ). This was caused by  
116 strong locus-dependent DH propensities not perturbed by permutation (**Supplementary Fig. 3**,  
117 **Methods**). Due to these locus effects, simply using the mean DH profile across training samples can  
118 predict cross-locus DH variation to certain extent (**Supplementary Fig. 4**), although such prediction is  
119 independent of test sample and therefore less accurate compared to BIRD predictions which utilize test-  
120 sample-dependent transcriptome information.

121

122 Based on  $r_C$ , RNA-seq was also able to predict how DH varied across samples with substantially higher  
123 accuracy than random expectation (**Fig. 1e**, mean  $r_C$  of BIRD vs. BIRD-Permute = 0.51 vs. -0.15, two-  
124 sided Wilcoxon signed-rank test  $p$ -value <  $2.2 \times 10^{-16}$ ). **Figure 1b** shows two examples of such prediction.  
125 Prediction of cross-sample variation was less accurate than prediction of cross-locus variation (**Fig. 1d-e**,  
126 BIRD mean  $r_C$  vs. mean  $r_L = 0.51$  vs.  $0.87$ ), because cross-sample prediction performance was evaluated  
127 within each locus and not affected by the locus effects. Cross-sample prediction accuracy varied  
128 substantially across loci (**Fig. 1e**). A large proportion of loci can be predicted with good accuracy: 57%  
129 and 23% of loci had  $r_C > 0.5$  and  $> 0.75$ , respectively. For each locus, the coefficient of variation (CV) of the  
130 predicted DH values across samples was computed to characterize its cross-sample DH variability  
131 (**Methods**). It was observed that loci with smaller  $r_C$  also tend to have smaller CV (**Fig. 1g**). On average,  
132 prediction of cross-sample variability was more accurate for loci with higher variability (**Fig. 1h**). For  
133 instance, for loci with  $CV > 0.4$ , the mean  $r_C$  was 0.69 ( $> 0.51$ , the mean  $r_C$  of all loci), and 84% and 47% of  
134 such loci had  $r_C > 0.5$  and  $> 0.75$  respectively.

135

136 By grouping genomic loci with similar cross-sample DH variation patterns into clusters and treating each  
137 cluster as a “pathway” of co-activated regulatory elements (**Methods**), cross-sample variation of the  
138 pathway activity (i.e., mean DH of all loci in each pathway) can be predicted more accurately (mean  
139  $r_c=0.71$ , 85% and 55% of pathways had  $r_c > 0.5$  and  $>0.75$ ) than predicting cross-sample variability of  
140 individual loci (**Fig. 1i**).

141  
142 BIRD prediction also substantially reduced the squared prediction error compared to random  
143 expectation (**Fig. 1f**, BIRD vs. BIRD-Permute;  $\tau=0.24$  vs. 0.55). Together, the above results are consistent  
144 with the results in the companion study where DH was predicted using exon arrays (Zhou et al.  
145 submitted).

146  
147 **Predicting Transcription Factor Binding Sites Using Bulk RNA-seq**  
148 We tested if the predicted DH at DNA motif sites can predict transcription factor (TF) binding sites  
149 (TFBSs) by analyzing 34 TFs in GM12878 and 25 TFs in K562 cells. BIRD models trained using the  
150 Epigenome Roadmap data (70 samples, GM12878 and K562 were not part of the 70 samples) were  
151 applied to predict DH in GM12878 and K562 using RNA-seq. The DNA motif of each TF was mapped to  
152 the genome, and motif sites with high predicted DH were identified and ranked as predicted TFBSs. For  
153 each TF and cell type, the corresponding ChIP-seq data were obtained from ENCODE (ENCODE Project  
154 Consortium 2012). Motif-containing ChIP-seq peaks were used as gold standard to evaluate the  
155 prediction accuracy (**Methods**). **Figure 1j-k** and **Supplementary Figures 5-6** show the percentage of gold  
156 standard TFBSs that were discovered by the top predicted sites. For comparison, we also predicted  
157 TFBSs using the true DNase-seq data (positive control) and the mean DH profile of the training samples  
158 (negative control). The results show that BIRD predictions based on RNA-seq were able to discover a  
159 substantial proportion of the true TFBSs. For instance, the top 15,000 predictions for YY1 in GM12878  
160 ( $q$ -value = 0.01) covered 76% of the gold standard YY1 binding sites (**Fig. 1j**). As expected, predictions  
161 based on true DNase-seq were more accurate than BIRD predictions. However, compared to the cell-

162 type-independent prediction based on the mean DH profile, BIRD predictions were substantially better  
163 because BIRD used cell-type-specific information contained in the transcriptome.

164

### 165 **Predicting Chromatin Accessibility Using Small-cell-number RNA-seq**

166 Our next question is whether BIRD trained using bulk RNA-seq data from the Epigenome Roadmap can  
167 be applied to RNA-seq generated using small-cell-number samples to predict DH. We obtained published  
168 bulk RNA-seq and RNA-seq from samples with 10, 30 and 100 cells for GM12878 (Marinov et al. 2014).  
169 BIRD models trained using the Epigenome Roadmap data were applied to each sample. For evaluation,  
170 true chromatin accessibility profiles for these small-cell-number samples are not available. However,  
171 according to statistical theory, if cells in a small-cell-number sample are randomly drawn from a bulk cell  
172 population, the mean DH profile of the small-cell-number sample and that of the bulk sample should  
173 have the same expectation. Therefore, one can use the bulk DNase-seq data for GM12878 (ENCODE  
174 Project Consortium 2012) from the ENCODE as the “truth”. Based on this gold standard, we compared  
175 BIRD predictions with GM12878 ATAC-seq from 500 and 50,000 cells. Our evaluation was primarily  
176 based on cross-locus correlation  $r_L$ , because reliably estimating cross-sample correlation  $r_C$  requires a  
177 large number of test cell types which were not available here. It turns out that ATAC-seq from 50,000  
178 cells (“ATAC-b50k”) showed the highest cross-locus correlation with the true DNase-seq signal (**Fig. 2a-b**,  
179  $r_L=0.76$ ). Surprisingly, however, BIRD-predicted DH signals from 30 and 100 cells consistently predicted  
180 the truth better than ATAC-seq from 500 cells (**Fig. 2a-b**,  $r_L=0.63$ , 0.69 and 0.69 for “ATAC-b500”, “BIRD-  
181 b30” and “BIRD-b100”). Of note, using the mean DH profile from the training data alone was able to  
182 predict DH to certain degree (**Fig. 2a-b**,  $r_L=0.56$  for “Mean”). The prediction accuracy of BIRD increased  
183 with increasing cell number. BIRD predictions based on  $\geq 30$  cells were almost as accurate as predictions  
184 based on bulk RNA-seq (**Fig. 2a-b**,  $r_L=0.70$  for “BIRD-bulk”). **Figure 2b** provides an example illustrating  
185 signals from different methods.

186

187 Interestingly, combining the ATAC-seq signal from 500 cells and the BIRD-predicted DH from 30 cells by  
188 average (530 cells used in total) allowed one to better predict the gold standard DNase-seq signal (**Fig.**

189 **2a-b**, “BIRD-b30+ATAC-b500”,  $r_L=0.76$ ). The combined signal achieved the same accuracy as ATAC-seq  
190 using 50,000 cells ( $r_L=0.76$ ) and was better than using BIRD-b30 ( $r_L=0.69$ ) or ATAC-b500 ( $r_L=0.63$ ) alone  
191 (**Fig. 2c-f**). Similarly, by averaging ATAC-seq from 50,000 cells and BIRD predictions from 30 cells, we  
192 were able to predict the gold standard better than ATAC-b50k (**Fig. 2a-b**,  $r_L=0.80$  for “BIRD-b30+ATAC-  
193 b50k”). The same improvement was not observed when the BIRD prediction was replaced by the  
194 prediction based on the mean DH profile (**Fig. 2a-b**,  $r_L=0.69$  and 0.75 for “Mean+ATAC-b500” and  
195 “Mean+ATAC-b50k”). These results show that DH predicted from small-cell-number RNA-seq can be  
196 integrated with small-cell-number ATAC-seq data (BIRD+ATAC-seq) to obtain better signal.

197  
198 We repeated the above evaluation by using ATAC-seq from 50,000 cells to replace bulk DNase-seq to  
199 serve as gold standard. Similar conclusions were obtained (**Methods, Supplementary Fig. 7**). Unlike the  
200 DNase-seq gold standard which came from a study different from the studies that generated the test  
201 ATAC-seq and RNA-seq data, the ATAC-50k gold standard was collected from the same study as ATAC-  
202 b500 (RNA-seq was from a different study). Thus, the ATAC-50k gold standard should intrinsically favor  
203 ATAC-b500 over BIRD due to potential lab effects. Despite this, BIRD predictions based on  $\geq 30$  cells  
204 performed close to ATAC-b500 in this comparison, and BIRD-b30+ATAC-b500 outperformed ATAC-b500  
205 (**Supplementary Fig. 7**).

206  
207 **Predicting Transcription Factor Binding Sites Using Small-cell-number RNA-seq**  
208 We further evaluated whether DH predicted using small-cell-number RNA-seq coupled with DNA motif  
209 information can predict TFBSSs. Similar to the analyses performed for the bulk RNA-seq, we predicted  
210 TFBSSs for 34 TFs in GM12878 using BIRD-b30, BIRD-hybrid (i.e. BIRD-b30+ATAC-b500), ATAC-b50k,  
211 ATAC-b500, true DNase-seq (“True”), and mean DH of training samples (“Mean”). **Figure 3a-f** and  
212 **Supplementary Figure 8** show the performance curves. To facilitate method comparison, we calculated  
213 the area under the curve (AUC) for each method, normalized by dividing the AUC of the “True” DNase-  
214 seq (**Fig. 3g, Supplementary Table 3, Methods**). Comparison of the normalized AUC shows that BIRD-  
215 b30 outperformed mean DH in all 34 tested TFs. Furthermore, BIRD-b30 outperformed ATAC-b500 in 23

216 of 34 TFs (**Fig. 3g, Supplementary Fig. 8**). Interestingly, BIRD-hybrid (BIRD-b30+ATAC-b500)  
217 outperformed ATAC-b500 in 32 of 34 TFs, and outperformed ATAC-b50k in 21 of 34 TFs. Thus, DH  
218 predicted by BIRD from 30 cells more accurately predicted TFBSSs than ATAC-seq from 500 cells, and  
219 combining BIRD predictions with ATAC-seq from small number of cells better predicted TFBSSs than bulk  
220 ATAC-seq.

221

## 222 **A Comparison of BIRD, ATAC-seq and MOWChIP-seq for Small-cell-number Samples**

223 Next, we compared DH predicted by BIRD using 30 cells, ATAC-seq, and histone modification H3K27ac  
224 and H3K4me3 profiles measured by MOWChIP-seq using 100 and 600 GM12878 cells. Since the genomic  
225 distribution of histone modification signal is different from that of chromatin accessibility due to  
226 nucleosome displacement around TFBSSs (He et al. 2010), we first optimized the parameter for analyzing  
227 MOWChIP-seq data (**Methods, Supplementary Fig. 9**). The comparisons below are based on the optimal  
228 MOWChIP-seq performance. It was observed that predictions or measurements for each data type  
229 correlated better with the bulk data from the same data type than the bulk data from other data types  
230 (**Supplementary Fig. 10**). For instance, H3K27ac MOWChIP-seq using 100 and 600 cells (H3K27ac-b100  
231 and H3K27ac-b600) performed better than BIRD-b30 when H3K27ac bulk ChIP-seq was used as gold  
232 standard for evaluation, but the same MOWChIP-seq data performed worse than BIRD-30 when bulk  
233 DNase-seq was used as gold standard (**Supplementary Fig. 10, Fig. 2a,b,g,h**). This suggests that there  
234 were substantial differences among data types, making a fair comparison difficult. For predicting TFBSSs,  
235 however, both BIRD-b30 and ATAC-b500 substantially outperformed MOWChIP-seq based on the overall  
236 performance in all 34 tested TFs (**Fig. 3a-f, Supplementary Fig. 8**). Among the MOWChIP-seq data,  
237 H3K27ac-b600 had the best overall performance for predicting TFBSSs (**Fig. 3g, Supplementary Table 3**).  
238 BIRD-b30 outperformed H3K27ac-b600 MOWChIP-seq in 27 of 34 tested TFs. ATAC-b500 outperformed  
239 H3K27ac-b600 in 31 of 34 tested TFs. Finally, BIRD-hybrid (BIRD-b30+ATAC-b500) outperformed  
240 H3K27ac-b600 in 33 out of 34 TFs.

241

242

243 **Predicting Chromatin Accessibility and TFBSS Using Single-cell RNA-seq**

244 We proceeded to investigate whether one can use single-cell RNA-seq data to predict DH. We analyzed a  
245 single-cell RNA-seq dataset with 28 single cells for GM12878 (Marinov et al. 2014). After calculating  
246 gene expression for each cell, we pooled  $k$  ( $k = 1, 5, 10, 20, 28$ ) cells randomly drawn from the dataset  
247 together and used their average expression profile to predict DH based on BIRD models trained from the  
248 Epigenome Roadmap bulk RNA-seq data. For comparison, we analyzed published single-cell ATAC-seq  
249 data in GM12878 generated by two different protocols ("ATAC1" (Cusanovich et al. 2015): 222 cells;  
250 "ATAC2" (Buenrostro et al. 2015): 340 cells). We computed average scATAC-seq profile for  $k$  ( $k = 1, 5, 10,$   
251  $20, 28, 50, 100, 222$  and 340) cells randomly drawn from each dataset respectively. **Figure 4** shows the  
252 performance of different methods evaluated using bulk DNase-seq as gold standard. Holding the cell  
253 number the same, BIRD based on pooled scRNA-seq was consistently better than pooled scATAC-seq for  
254 predicting bulk DNase-seq (**Fig. 4b,c**). BIRD predictions based on a single cell and pooled scATAC-seq  
255 using  $\leq 50$  cells from ATAC1 or  $\leq 20$  cells from ATAC2 were less accurate than predictions based on the  
256 mean DH profile (**Fig. 4b**). However, prediction accuracy increased as more cells were pooled together.  
257 BIRD with 10 cells performed better than the mean DH profile, and it was comparable to pooling 100  
258 cells from ATAC1 or pooling 50 cells from ATAC2. The results remained similar when the gold standard  
259 was changed to bulk ATAC-seq data from 50,000 or 500 cells (**Supplementary Fig. 11**).  
260

261 We also combined BIRD predictions based on pooling scRNA-seq from 28 cells with the pooled scATAC-  
262 seq profile from  $x$  cells ( $x = 22, 72, 194$  and 312) by taking the average of the two profiles ("BIRD-  
263 hybrid"). We then compared BIRD-hybrid with pooled scATAC-seq data using the same number of cells  
264 (i.e.,  $k = 28 + x = 50, 100, 222, 340$ ). BIRD-hybrid also outperformed pooled scATAC-seq (**Fig. 4a-c**,  
265 **Supplementary Fig. 11**).

266  
267 To test whether predictions from scRNA-seq can predict TFBSS in a similar fashion as small-cell-number  
268 RNA-seq, we again analyzed 34 TFs in GM12878 (**Fig. 5a-f, Supplementary Figs. 12-15, Supplementary**  
269 **Table 4**). Once again, BIRD and BIRD-hybrid performed better than pooled scATAC-seq. For instance,

270 when pooling 10 cells, BIRD prediction outperformed ATAC1 and ATAC2 in all 34 TFs, and it  
271 outperformed mean DH in 33 of 34 TFs. Using 28 cells, BIRD outperformed ATAC1, ATAC2 and mean DH  
272 in 34, 32 and 33 out of 34 tested TFs respectively. Using 222 single cells, BIRD-hybrid outperformed  
273 ATAC1, ATAC2 and mean DH in 33, 33 and 31 out of the 34 tested TFs respectively (**Supplementary Fig.**  
274 **15, Supplementary Table 4**).

275

276 We applied BIRD to another scRNA-seq dataset (Trapnell et al. 2014) (69 cells) from human skeletal  
277 muscle myoblasts (HSMM) (**Fig. 5g**, “approach 1”, pooling  $k=1, 5, 10, 20, 30, 40, 50$  and 69 cells). For  
278 this dataset, scATAC-seq was not available and therefore not compared. We used bulk DNase-seq data  
279 in HSMM as gold standard for evaluation. The prediction accuracy of BIRD by pooling  $\geq 5$  cells was better  
280 than the accuracy based on the mean DH profile, and the accuracy of BIRD by pooling  $\geq 30$  cells was  
281 comparable to BIRD predictions based on bulk RNA-seq (**Fig. 5g**). This further demonstrates that one can  
282 predict DH from scRNA-seq by pooling a small number of cells.

283

284 When applying BIRD to scRNA-seq data, it is important to pool RNA-seq data from multiple cells first and  
285 then make predictions based on the pooled gene expression profile. When we tried to first predict DH  
286 based on each single cell and then average the predictions from multiple cells, the prediction  
287 performance was substantially worse for both the GM12878 and HSMM data (**Fig. 5g-h**, “approach 2”).  
288 This is because expression measurements from scRNA-seq have substantial biases (e.g., zero-inflation by  
289 dropout events (Kharchenko et al. 2014)) that cannot be removed by the usual normalization. Impacts  
290 on prediction by such bias can be reduced when multiple cells are pooled together to measure gene  
291 expression and the pooled expression profile is then normalized against the bulk RNA-seq data in the  
292 training dataset.

293

## 294 **DISCUSSION**

295 To summarize, our analyses demonstrate that predicting chromatin accessibility using RNA-seq can  
296 provide a new approach for regulome mapping both in bulk samples and in samples with small number

297 of cells. The study compared multiple state-of-the-art technologies for mapping regulome in small-cell-  
298 number samples including ATAC-seq, scATAC-seq, MOWChIP-seq and BIRD. Our results show that for  
299 analyzing small-cell-number samples, BIRD can offer competitive performance compared to ATAC-seq  
300 and scATAC-seq. In particular, using 5-10 folds fewer cells, BIRD reached the same accuracy as ATAC-seq  
301 and pooled scATAC-seq for predicting bulk chromatin accessibility. Also, BIRD based on scRNA-seq more  
302 accurately predicted bulk chromatin accessibility than using scATAC-seq by pooling the same number of  
303 cells. Besides ATAC-seq, BIRD based on fewer cells also offered competitive or better performance  
304 compared to MOWChIP-seq using more cells.

305  
306 Based on our analyses, the minimum number of cells required by the current technology to recover  
307 chromatin accessibility in a bulk sample is approximately 10 cells. This was achieved by BIRD. Averaging  
308 single-cell ATAC-seq from 10 cells predicted bulk chromatin accessibility worse than the trivial prediction  
309 based on the mean DH profile. By contrast, BIRD predictions based on pooling scRNA-seq from 10 cells  
310 were better than predictions based on the mean DH profile. This highlights the limitation of scATAC-seq  
311 due to its intrinsic discreteness (**Supplementary Fig. 1**). Compared to scATAC-seq, scRNA-seq data are  
312 less discrete since each gene can have more than two copies of transcripts in a cell.

313  
314 Most recently, single-cell ChIP-seq (Drop-ChIP) for histone modifications and single-cell DNase-seq  
315 (scDNase-seq) have been reported (Rotem et al. 2015; Jin et al. 2015). Since the current Drop-ChIP and  
316 scDNase-seq data for single cells are in mouse and we do not have enough training samples in mouse to  
317 build BIRD models, we were unable to directly compare BIRD with Drop-ChIP and scDNase-seq here. Of  
318 note, Drop-ChIP data are highly discrete, with 500~10,000 reads and an average of ~800 peaks detected  
319 per cell. Our results on scATAC-seq (ATAC1 and ATAC2 had an average of ~2,700 and ~14,000 reads per  
320 cell respectively) suggest that discreteness of the signal will remain a problem for Drop-ChIP. Although  
321 scDNase-seq has also been applied to pooled human cells dissected from formalin-fixed paraffin-  
322 embedded tissues, there is no gold standard available for a direct comparison between BIRD and

323 scDNase-seq for that application. In the future, it will be interesting to compare BIRD with Drop-ChIP  
324 and scDNase-seq when appropriate test and benchmark data become available.

325

326 Our study has important practical relevance on future data analyses. It shows that transcriptome-based  
327 regulome prediction can greatly increase the value of current and future bulk, small-cell-number and  
328 single-cell RNA-seq experiments. By adding a new component to the standard RNA-seq analysis pipeline,  
329 this approach allows one to use RNA-seq not only for studying transcriptome but also for studying  
330 regulome. This can greatly impact how to most effectively use existing and future RNA-seq data, which is  
331 particularly relevant given that enormous amounts of RNA-seq data will be generated in the years to  
332 come.

333

334 Our study also has important implications for future experiment design. When a sample contains only a  
335 very limited number of cells, researchers have to decide how these cells should be wisely used. For  
336 example, should one use all cells for transcriptome profiling by RNA-seq or regulome mapping by ATAC-  
337 seq? Results from this study show that one may divide the samples into two parts, one for RNA-seq or  
338 scRNA-seq, and one for ATAC-seq or scATAC-seq. This strategy has two advantages. First, one can obtain  
339 information for two different data types instead of only one data type. Second, by spending some cells  
340 on RNA-seq, BIRD-hybrid allows one to combine the two data types to produce comparable or better  
341 regulome mapping than spending all cells on ATAC-seq. This study also shows that if one decides to use  
342 all cells for RNA-seq, one can still obtain information on regulome through prediction. Thus, it is also  
343 possible to analyze transcriptome and regulome simultaneously in a small-cell-number sample by  
344 measuring only transcriptome.

345

346 Currently, BIRD predictions based on RNA-seq from a single cell were less accurate than the mean DH  
347 profile for predicting the bulk chromatin accessibility. One possible reason is that technical biases in the  
348 single-cell RNA-seq data (e.g., excessive zeros in the data) cannot be easily removed by normalization  
349 when there is only one cell, making the prediction inaccurate. Another possible reason is that the small

350 sample size ( $n=1$  cell) is not sufficient to overcome the random variation in single-cell expression to  
351 recover the behavior of a bulk sample. Naturally, an important question for future research is whether  
352 one can develop methods insensitive to the biases in scRNA-seq to improve predictions in a single cell.  
353 More generally, there is still great demand for new experimental or computational methods for single-  
354 cell regulome mapping, particularly in the context that the discrete signals generated at one random  
355 time point by current experimental technologies such as scATAC-seq may not adequately describe the  
356 average steady-state behavior of a cell over time.

357

358 As a proof-of-concept, this study shows that predicting chromatin accessibility using bulk and small-cell-  
359 number RNA-seq is feasible. Another important next step is to explore whether other functional  
360 genomic data types can be predicted in a similar fashion in small-cell-number samples.

361

362

## 363 METHODS

### 364 DNase-seq data processing

365 The aligned DNase-seq data (alignment based on hg19) from 70 samples were downloaded from the  
366 Roadmap Epigenomics project (Kundaje et al. 2015) (<ftp://ftp.genboree.org/EpigenomeAtlas/Current->  
367 [Release/experiment-sample/Chromatin\\_Accessibility/](#)). The analyses in this study were focused on  
368 chromosomes 1 to X. Excluding chromosome Y, the genome was divided into 200 base pair (bp) non-  
369 overlapping bins. The number of reads mapped to each bin was counted for each DNase-seq sample. To  
370 adjust for different sequencing depths, bin read counts for each sample  $i$  were first divided by the  
371 sample's total read count  $N_i$  and then scaled by multiplying a constant  $N$  ( $N = \min_i\{N_i\} = 12,422,306$ ,  
372 which is the minimum sample read count of all samples). The normalized read counts were then log2  
373 transformed after adding a pseudocount 1. The normalized and log2-transformed read counts were  
374 used to represent DH levels of genomic bins.

375

376 DNase-seq data for GM12878, K562 and HSMM were downloaded from the ENCODE project (ENCODE  
377 Project Consortium 2012). The data were aligned to human genome hg19 using bowtie (Langmead et al.  
378 2009) (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase>). The  
379 aligned reads were processed in the same way as the Epigenome Roadmap data to derive DH levels. Of  
380 note, the ENCODE data contained replicate samples for each cell type. The normalized read counts from  
381 replicate samples were first averaged to characterize the DH level for each bin in each cell type. The DH  
382 level was then log2 transformed after adding a pseudocount 1.

383

#### 384 **Genomic loci filtering**

385 Since most genomic loci are noise rather than regulatory elements, we filtered genomic loci to exclude  
386 those without strong DH signal in any Epigenome Roadmap DNase-seq sample. The filtering was done in  
387 three steps. First, genomic bins with normalized read count  $\leq 8$  in all samples were excluded. Second,  
388 bins with normalized read count larger than 10,000 in  $\geq 1$  sample were considered abnormal and  
389 therefore also excluded. Third, a signal-to-noise ratio (SNR) was computed for each bin in each sample,  
390 and bins with SNR  $\leq 2$  in all samples were considered as noise and filtered out. In order to compute SNR  
391 of a genomic bin in a sample, we first collected 500 bins in the neighborhood of the bin in question. The  
392 average DH level of these bins was computed and then log2 transformed after adding a pseudocount 1  
393 to serve as the background. The  $\log_2(\text{SNR})$  was defined as the difference between the normalized and  
394 log2 transformed DH level of the bin in question and the background. SNR  $\leq 2$  is equivalent to  $\log_2(\text{SNR})$   
395  $\leq 1$ .

396

397 After filtering, 1,136,465 genomic bins (called DNase I hypersensitive sites, or DHSs, hereinafter) with  
398 unambiguous DNase-seq signal in at least one sample were identified. All analyses in this study were  
399 performed on these genomic loci, except for the leave-one-out cross-validation analysis in **Figure 1d-i**  
400 which will be described in a separate section below.

401

402

403 **Bulk RNA-seq data processing**

404 The aligned RNA-seq data (alignment based on hg19) for the same 70 Epigenome Roadmap samples  
405 were downloaded from <ftp://ftp.genboree.org/EpigenomeAtlas/Current-Release/experiment-sample/mRNA-Seq/>. Cufflinks (Trapnell et al. 2010) was used to compute the expression values (i.e.,  
406 FPKM: fragments per kilobase of exon per million mapped fragments) using gene annotations in  
407 GENCODE (Harrow et al. 2012) (Release 19 (GRCh37.p13)). 37,335 transcripts (called “genes” hereinafter  
408 for simplicity) with FPKM > 1 in at least one sample were identified. These FPKM values were log2  
409 transformed after adding a pseudocount 1 and then quantile normalized across samples. After  
410 normalization, the quantiles of the Epignome Roadmap training data were stored for future use. When  
411 new RNA-seq samples need to be analyzed, they will be quantile normalized against these stored  
412 quantiles.

414

415 For evaluation, we downloaded the following data from GEO: (1) GM12878 and K562 bulk RNA-seq data  
416 (GSM958728, GSM958729), (2) GM12878 RNA-seq data from small-cell-number samples with 10, 30 and  
417 100 cells (GSM1087860, GSM1087861, GSM1087858, GSM1087859, GSM1087856, GSM1087857). For  
418 these samples, reads were mapped to human genome hg19 using Tophat (Kim et al. 2013). Gene  
419 expression values were then computed using Cufflinks in the same way as how we processed the  
420 Epigenome Roadmap RNA-seq data. Finally the gene expression values were quantile normalized with  
421 the Epigenome Roadmap RNA-seq data using the stored quantiles.

422

423 **BIRD Model**

424 The BIRD (Big data Regression for predicting DNase I hypersensitivity) algorithm is described and  
425 systematically evaluated in a companion article. For readers’ convenience, we review its workflow in  
426 **Supplementary Methods**. Readers are referred to Zhou *et al.* (*Zhou et al. submitted*) for more details.  
427 BIRD software is available at <https://github.com/WeiqiangZhou/BIRD>. Models trained using the 70  
428 Epigenome Roadmap samples have been stored in the software package.

429

430 **Prediction performance evaluation**

431 Three statistics were used in this article for evaluating prediction accuracy in different analyses. Let  $\hat{y}_{lm}$   
432 be the predicted DH level of locus  $l$  ( $=1, \dots, L$ ) in test sample  $m$  ( $=1, \dots, M$ ), and let  $y_{lm}$  be the true DH  
433 level measured by DNase-seq. The three statistics include:

434

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

439

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

443

444 (3) Squared prediction error ( $\tau$ ). This is measured by the total squared prediction error scaled by the  
445 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  
446 DHSs and test samples.

447

448 **Leave-one-out cross-validation**

449 Leave-one-out cross-validation was used to evaluate BIRD prediction accuracy when bulk RNA-seq data  
450 were used as predictors. In each fold of the cross-validation, the 30 Epigenome Roadmap cell types  
451 (consisting of 70 samples) were partitioned into a training dataset with 29 cell types and a test dataset  
452 with 1 cell type. In other words, all samples from one cell type were used as test data, and all samples  
453 from the remaining 29 cell types were used as the training data. BIRD was then trained using all samples  
454 in the training dataset and applied to predict DH for all samples in the test dataset.

455

456 To ensure that the test data are not used in the construction of prediction models, the predictor and  
457 genomic loci filtering procedure was applied to each fold of cross-validation by using the training data  
458 only. For instance, we identified genes with FPKM > 1 in at least one RNA-seq sample in the training data  
459 as predictors. The identified predictors were a subset of the 37,335 genes described before (note: the  
460 37,335 genes were identified using all 70 samples rather than using only the training samples). For  
461 different folds of cross-validation, a slightly different set of predictors was identified. Similarly, genomic  
462 loci (i.e. DHSs) to be predicted were selected by applying the previously described filtering protocol to  
463 the training data only: (1) normalized bin read count  $\geq 8$  in at least one sample; (2) normalized bin read  
464 count  $< 10,000$  in all samples; (3) SNR  $\geq 2$  in at least one sample. Prediction models were constructed for  
465 the identified genomic loci. These loci varied from fold to fold, and they were also slightly different from  
466 the 1,136,465 genomic loci derived from all 70 samples. Of note, the parameters of BIRD (i.e.  $K$  and  $N$  in  
467 **Supplementary Methods**) were selected following the same procedure described in **Supplementary**  
468 **Methods** using 1% loci randomly chosen from the training dataset (i.e., samples from 29 cell types in  
469 each fold). Since the training data were different in each fold, the parameters also varied from fold to  
470 fold.

471  
472 After predictions were made for all samples,  $r_L$ ,  $r_C$  and  $\tau$  were calculated between the true and predicted  
473 DH profiles. Conceptually, one can organize the predicted values into a matrix. Rows of the matrix  
474 correspond to genomic loci, and columns of the matrix correspond to samples. The matrix has missing  
475 values as not all genomic loci have prediction models in all samples. This is because genomic loci filtering  
476 was dependent on the training data. As a result, in each fold of cross-validation, prediction models were  
477 built for a slightly different set of genomic loci. To compute  $r_L$ ,  $r_C$  and  $\tau$ , missing data points in the  
478 prediction matrix were excluded, and only data points with predicted DH values were used. This  
479 produces **Figure 1d-f**.

480

481 **Random prediction models by permutation**

482 To construct random prediction models, sample labels of the DNase-seq data in the training dataset  
483 were shuffled. This permutation broke the connection between DNase-seq and RNA-seq samples. Then,  
484 BIRD was trained by the permuted training dataset and applied to predict DH in the test dataset. The  
485 permutation was performed in each fold of the leave-one-out cross-validation and the prediction  
486 performance was then evaluated by  $r_L$ ,  $r_C$  and  $\tau$ . Of note, our permutation here did not perturb the  
487 locus effects of DH profile. Therefore, predictions from the random prediction models mostly captured  
488 the average DH level of each genomic locus in the training dataset.

489

490 **Wilcoxon signed-rank test**

491 Two-sided Wilcoxon signed-rank test (Wilcoxon 1945) was performed to obtain  $p$ -values for comparing  
492 prediction accuracy of BIRD and random prediction models. In order to test whether two methods  
493 perform equally in terms of  $r_L$ , the paired  $r_L$  values from these two methods for each sample was  
494 obtained. Then the  $r_L$  pairs from all samples are used for Wilcoxon signed-rank test. Similarly, to  
495 compare two methods in terms of  $r_C$ , the paired  $r_C$  values for each locus were obtained, and  $r_C$  pairs  
496 from all genomic loci were used for the Wilcoxon signed-rank test.

497

498 **Categorization of genomic loci based on cross-sample variability**

499 When studying the cross-sample prediction performance (i.e.,  $r_C$ ) in **Figure 1g-h**, genomic loci were  
500 grouped into different categories based on their cross-sample variability of the predicted DH profile.  
501 First, loci with predicted DH value (at log2 scale) smaller than 2 across all cell types were treated as noisy  
502 loci (**Fig. 1g-h**, indicated by “Noisy loci”). For such loci, the observed DH level may contain substantial  
503 noise, and the cross-sample correlation between the predicted and the true DH is expected to be low  
504 (since the correlation between random noise and another independent random variable is expected to  
505 be zero). After excluding the noisy loci, the other loci were then categorized based on the coefficient of  
506 variation (CV) of the cross-sample DH values. For each locus, CV was calculated as the ratio of the  
507 standard deviation to mean of the predicted DH at this locus across all samples. Loci were divided into

508 three categories:  $CV \leq 0.2$ ,  $0.2 < CV \leq 0.4$ ,  $CV > 0.4$  (**Fig. 1g-h**). A large CV indicates that the DH of a locus has  
509 more variation across samples. **Figure 1g** shows the distribution of  $r_c$ . Genomic loci are grouped into bins  
510 based on  $r_c$  values. For each bin, the number of loci in different CV categories is shown. **Figure 1h** shows  
511 the distribution of  $r_c$  in each CV category.

512

### 513 **Chromatin accessibility prediction for clusters of co-activated DHSs**

514 A transcriptional regulation process often involves co-activation of multiple *cis*-regulatory elements.  
515 Such co-activated regulatory elements can be viewed as regulatory “pathways”. Previously, DHSs  
516 discovered from the ENCODE DNase-seq data have been clustered into 2500 clusters based on their  
517 cross-sample co-variation patterns (Sheffield et al. 2013). Using these pre-defined clusters as  
518 “pathways”, we investigated how accurate the pathway activity can be predicted using bulk RNA-seq. To  
519 do so, we first identified the cluster membership of the 1,136,465 genomic loci studied here based on  
520 the clustering results provided by Sheffield et al. (2013) (obtained from  
521 <http://big.databio.org/RED/TableS03-dhs-to-cluster.txt.tar.gz>, the “original cluster” assignment was  
522 used). For each cluster, we then computed its mean DH level (missing values were excluded) in each  
523 sample using all DHSs in the cluster. Next, we built prediction models to predict the mean DH level of  
524 each cluster (i.e., “pathway activity”) in the same way as  $BIRD_{\bar{X},\bar{Y}}$ . The prediction accuracy was  
525 evaluated using leave-one-out cross-validation (i.e., using 29 cell types as training and 1 cell type as test  
526 data). After cross-validation, the average DH level for each DHS cluster and each sample was obtained,  
527 and the cross-sample correlation was calculated between the true and predicted mean DH level for each  
528 cluster (**Fig. 1i**).

529

### 530 **Transcription factor binding site prediction**

531  $BIRD$  models trained using the 70 Epigenome Roadmap samples were applied to predict binding sites for  
532 34 TFs in GM12878 cells and 25 TFs in K562 cells. For each TF, DNA motif obtained from TRANSFAC  
533 (Matys et al. 2006) and JASPAR (Mathelier et al. 2014) (**Supplementary Table 2**) was computationally  
534 mapped to the human genome using CisGenome (Ji et al. 2008) (using default likelihood ratio  $\geq 500$

535 cutoff). DHSs (i.e., the 1,136,465 genomic bins) that overlapped with motif sites were retained for  
536 subsequent analyses. These motif-containing DHSs were ranked in decreasing order based on the  
537 predicted DH level to serve as the predicted TFBSSs. As a comparison, DHSs were also ranked based on  
538 two other methods: the true DH level at each DHS from the corresponding DNase-seq data (“True”) and  
539 the DH level predicted based on the mean DH profile of the 70 training samples (“Mean”).

540

541 To evaluate the prediction performance of different methods, the transcription factor ChIP-seq uniform  
542 peaks data for the 34 TFs in GM12878 and 25 TFs in K562 were downloaded from the ENCODE project to  
543 serve as the truth (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/>). ChIP-seq peaks overlapped with motif sites were used as the gold standard. The percentage  
544 of these gold standard peaks that were recovered by the top ranked predicted TFBSSs was computed to  
545 measure the sensitivity of each prediction method. Different methods were compared by plotting the  
546 sensitivity as a function of the number of predicted TFBSSs (Fig. 1j-k, Supplementary Figs. 5-6).

548

549 To evaluate statistical significance of the predicted TFBSSs, the same BIRD models were applied to a set of  
550 randomly sampled genomic bins (n= 984,213, sampled from non-repeat genomic regions) to make  
551 predictions. Using the predicted DH values in the random genomic loci as the null distribution, a *p*-value  
552 was computed for each studied DHS to evaluate the significance of its predicted DH level (*p*-value of the  
553 predicted DH level at a DHS = [no. of random loci with equal or larger predicted DH levels] / [the total no.  
554 of random loci]). To adjust for multiple testing, the *p*-values were converted to *q*-values based on the  
555 previously described method (Dabney and Storey). *q*-values for BIRD predictions were labeled on top of  
556 each sensitivity-rank plot (e.g., Fig. 1j-k).

557

558 To generate **Figure 3g** and **Supplementary Figure 15** that compare different TFBSS prediction methods  
559 using small-cell-number or single-cell RNA-seq data, we computed the area under the curve (AUC) for  
560 each method using the sensitivity rank curves in **Figure 3a-f**, **Supplementary Figure 8**, **Figure 5a-f**, and  
561 **Supplementary Figures 12-14**. The AUC of each method was then scaled by (i.e., divided by) the AUC

562 obtained using the true DNase-seq data (**Supplementary Tables 3-4**). To show a clear comparison of  
563 different methods for predicting binding sites of each TF, colors in the heatmap (**Fig. 3g** and  
564 **Supplementary Fig. 15**) reflect the transformed AUC values. For instance, values within each row were  
565 transformed to the range between 0 and 1 by:  $[\text{AUC value} - \min(\text{value})]/[\max(\text{value}) - \min(\text{value})]$ , here  
566  $\min(\text{value})$  and  $\max(\text{value})$  represent the minimum and maximum AUC value within each row. Of note,  
567 within each TF, the minimum AUC was transformed to 0 and the maximum AUC was transformed to 1.  
568 As a reference, the untransformed minimum AUC value from all methods was shown for each TF using a  
569 blue bar beside the TF name in **Figure 3g** and **Supplementary Figure 15**.

570  
571 To measure the overall prediction performance of each method, we calculated the average rank score  
572 (shown using red bars under the name of each method in **Fig. 3g** and **Supplementary Fig. 15**) across all  
573 34 test TFs. First, for each TF, different methods were ranked according to their AUC values. For instance,  
574 in **Figure 3g**, the best performing method has rank 1 and the worst performing method has rank 9. Then,  
575 we calculated the average rank across all test TFs for each method. Smaller average rank indicates  
576 better overall prediction performance.

577  
578 **Bulk ATAC-seq data processing**  
579 ATAC-seq data for GM12878 with 50,000 and 500 cells were obtained from GEO (GSE47753). The  
580 paired-end reads were aligned to human genome hg19 using bowtie (Langmead et al. 2009) with  
581 parameters (-X2000 -m 1) which specify that paired reads (a pair of reads was referred to as a fragment)  
582 with insertion up to 2,000 base pair (bp) were allowed to align and only uniquely aligned fragments  
583 were retained. Then, PCR duplicates (i.e. fragments that aligned to exactly the same genomic location)  
584 were determined using Picard (<http://broadinstitute.github.io/picard/>) where only one fragment was  
585 kept and the others were removed. Next, we measured the bin-level fragment coverage by counting  
586 how many fragments covered each 200bp genomic bin. Similar to the DNase-seq data, bin-level  
587 fragment coverage for each sample was first divided by the sample's whole-genome fragment coverage  
588 (i.e., sum of bin-level fragment coverage across the genome) and then scaled by a constant  $N$

589 (=12,422,306, to be consistent with the DNase-seq data). Finally, the normalized bin fragment coverage  
590 from different replicate samples were averaged and log2 transformed after adding a pseudocount 1.

591

## 592 **Histone modification ChIP-seq and MOWChIP-seq data processing**

593 H3K27ac and H3K4me3 MOWChIP-seq data for GM12878 with 100 and 600 cells were obtained from  
594 GEO (GSE65516: GSM1666202, GSM1666203, GSM1666204, GSM1666205, GSM1666206, GSM1666207,  
595 GSM1666208, GSM1666209). For both histone marks, the processed signal files provided by the  
596 MOWChIP-seq authors were downloaded from GEO. These files contained normalized read counts for  
597 the whole genome divided by 100 bp bins (Cao et al. 2015). The data were converted to 200 bp  
598 resolution by merging adjacent two 100 bp bins (i.e., adding the read counts of the two 100 bp bins).

599

600 Due to nucleosome displacement, the spatial distribution of histone modification signal surrounding  
601 each regulatory element (e.g., transcription factor binding site) may differ from the peak of the DNase-  
602 seq and ATAC-seq signal (He et al. 2010). Therefore we first explored different ways to summarize the  
603 histone modification signal in order to maximize its correlation with the bulk DNase-seq data. To do so,  
604 we considered a  $W$ -bp long window centered at each genomic locus (200bp bin). The normalized read  
605 counts of all 200bp bins covered by the window were averaged to serve as the summary of the histone  
606 modification signal at the locus. The summarized signals from replicate samples were averaged and log2  
607 transformed after adding pseudocount 1. We then tested different window sizes ( $W=200, 600, 1000,$   
608  $1400, 1800, 2200, 2600$  bp) to find the optimal  $W$  that maximizes the correlation between the  
609 summarized histone modification signal and the bulk DNase-seq signal (i.e., DH level at 200-bp  
610 resolution as described before) across all genomic loci (**Supplementary Fig. 9a**). For H3K27ac with 100  
611 and 600 cells, the optimal  $W$  was 2200. For H3K4me3 with 100 cells, the optimal  $W$  was 2200. For  
612 H3K4me3 with 600 cells, the optimal  $W$  was 1800. The summarized MOWChIP-seq signals using these  
613 optimal  $W$  were then compared with BIRD and ATAC-seq in **Figure 2** and **Supplementary Figure 10**.

614

615 H3K27ac and H3K4me3 ChIP-seq data for bulk GM12878 samples were obtained from ENCODE  
616 (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/>). For each  
617 200bp genomic bin, reads from the bulk ChIP and input control samples were counted. Bin read counts  
618 in each sample were normalized by the sample's total read count and then scaled by multiplying  
619 1,000,000. Signals were then calculated as the difference between the ChIP and input samples. Similar  
620 to MOWChIP-seq, we used the average signal of a  $W$ -bp long window ( $W=200, 600, 1000, 1400, 1800,$   
621 2200, 2600 bp) centered at each genomic locus to represent its summarized histone modification signal.  
622 The summarized signals from replicate samples were then averaged and log2 transformed after adding  
623 pseudocount 1. The optimal  $W$  was determined by maximizing the correlation with the bulk DNase-seq  
624 signal. For bulk H3K27ac and H3K4me3, the optimal  $W$  was 1000 and 1400 respectively (**Supplementary**  
625 **Fig. 9c**). The summarized ChIP-seq signals using these optimal  $W$  were then used for generating  
626 **Supplementary Figure 10**.

627  
628 For TFBS prediction using MOWChIP-seq data, we also first optimized the window size  $W$  for each  
629 MOWChIP-seq dataset. For each histone mark and cell number, we obtained the scaled AUC (scaling is  
630 done by dividing the AUC of using true DNase-seq data to predict TFBSs) for all 34 TFs using different  
631 window sizes. For each TF, different window sizes were then ranked based on the scaled AUC. The  
632 average rank of each window size  $W$  across all 34 TFs was computed (**Supplementary Fig. 9b**).  $W$  with  
633 the best average rank was identified. For H3K27ac MOWChIP-seq with 100 and 600 cells, the optimal  $W$   
634 was 1800. For H3K4me3 MOWChIP-seq with 100 cells, the optimal  $W$  was 2200. For H3K4me3  
635 MOWChIP-seq with 600 cells, the optimal  $W$  was 1800. The summarized MOWChIP-seq signals using  
636 these optimal  $W$  were then compared with BIRD and ATAC-seq in **Figure 3** and **Supplementary Figure 8**.  
637 We note that since TFBSs of different TFs may be associated with different histone modification  
638 signatures (e.g., different types of histone modifications), a better way to predict TFBS might be to use  
639 multiple types of histone modification data and develop a prediction model specific for each TF.  
640 However, that would make the TFBS prediction more difficult to apply in reality because one would  
641 need to collect more MOWChIP-seq data and have knowledge on the histone modification signature for

642 each TF. For this reason, our analyses here were primarily focused on evaluating the performance of  
643 using one data type (i.e., H3K27ac, H3K4me3, ATAC-seq, or RNA-seq) and a common prediction  
644 procedure for all TFs. This makes the comparison among MOWChIP-seq, ATAC-seq (ATAC-b500) and  
645 BIRD (BIRD-b30) relatively fair in the sense that different TFBS prediction methods have similar level of  
646 complexity in terms of data collection and computational analysis.

647

#### 648 **Chromatin accessibility prediction based on single-cell RNA-seq data**

649 We downloaded two datasets from GEO: (1) GM12878 single-cell RNA-seq data (GSE44618, 28 cells in  
650 total), (2) HSMM single-cell RNA-seq data (GSE52529, 69 cells from undifferentiated HSMM were used  
651 for our analysis). For these samples, reads were mapped to human genome hg19 using Tophat (Kim et al.  
652 2013). Gene expression values were then computed using Cufflinks in the same way as how we  
653 processed the Epigenome Roadmap RNA-seq data. For each dataset, we randomly sampled  $k$  cells ( $k = 1$ ,  
654 5, 10, 20, 28 for GM12878;  $k = 1, 5, 10, 20, 30, 40, 50, 69$  for HSMM) and calculated their average gene  
655 expression profile. The average gene expression profile was then used as the input for BIRD to predict  
656 the DH profile. This is the “approach 1” in **Figure 5g-h**. For each  $k$  (except for  $k = 1$  and 28 for GM12878,  
657 and  $k = 1$  and 69 for HSMM), the random sampling was repeated 10 times. The mean and standard  
658 deviation (SD) of the results from the 10 analyses were shown in **Figures 4b and 5g-h**. For  $k=1$ , the  
659 analysis was performed for every single cell.

660

661 We also tried a second approach to predict DH using single-cell RNA-seq (i.e., “approach 2” in **Fig. 5g-h**).  
662 In this approach, we first applied BIRD to predict DH for every single cell using the single-cell RNA-seq  
663 data. Then, we pooled a random group of  $k$  cells (note: the same cells sampled in “approach 1” were  
664 used to keep the comparison consistent) and computed the average of their predicted DH profile. The  
665 random sampling was repeated 10 times as above, and the mean and SD of the prediction performance  
666 from the 10 analyses were shown in **Figure 5g-h**. A comparison between approach 1 and approach 2  
667 shows that using multiple cells’ average expression as predictor had much higher prediction accuracy  
668 than using each cell’s expression to make prediction and then average the predicted DH profile.

669

670 **Chromatin accessibility based on single-cell ATAC-seq data**

671 Two single-cell ATAC-seq datasets for GM12878 were obtained. Dataset 1 (ATAC1) was obtained from  
672 GEO (GSM1647121). This dataset was a mixture of human GM12878 cells and mouse Patski cells. Paired-  
673 end reads were trimmed by Trimmomatic (Bolger et al. 2014) to remove adaptor content and aligned to  
674 human genome hg19 using bowtie2 (Langmead and Salzberg 2012) with parameter -X2000. PCR  
675 duplicates were removed using Picard. The aligned reads were then assigned to individual cells based on  
676 the barcode information and only GM12878 cells were retained for subsequent analyses. For each cell,  
677 bin-level fragment coverage was obtained for each genomic locus (i.e., 200bp bin), and bin fragment  
678 coverage was normalized in the same way as the bulk ATAC-seq data. The single-cell ATAC-seq data are  
679 highly discrete. According to the original report describing this data (Cusanovich et al. 2015), the  
680 sequencing has reached saturation and the median value of total read counts per cell was 2503. We  
681 identified GM12878 cells (n=222) with more than 500 non-zero-coverage loci and used them for the  
682 subsequent analyses. Dataset 2 (i.e. ATAC2) was obtained from GEO (GSE65360). This dataset contains  
683 GM12878 ATAC-seq for 384 single cells. For each single cell, paired-end reads were trimmed by  
684 Trimmomatic to remove adaptor content and aligned to human genome hg19 using bowtie2 with  
685 parameter -X2000. PCR duplicates were removed using Picard. Then, bin-level fragment coverage for  
686 each cell was computed, normalized and transformed in the same way as single-cell ATAC-seq dataset 1.  
687 340 cells with more than 500 non-zero-coverage loci were retained for the subsequent analyses.

688

689 For the single-cell ATAC-seq dataset 1, we randomly sampled a group of  $k$  cells ( $k = 1, 5, 10, 20, 28, 50,$   
690  $100, 222$ ) and calculated their average ATAC-seq profile (i.e., average of the normalized bin fragment  
691 coverage). The average profile was then log2-transformed after adding pseudocount 1. For each  $k$   
692 (except for  $k=1$  and 222), we repeated the random sampling 10 times. The mean and SD of results from  
693 the 10 analyses were shown in **Figure 4b**. For  $k=1$ , the analysis was performed on every single cell. The  
694 same analysis was also performed for the single-cell ATAC-seq dataset 2 with  $k$  cells ( $k = 1, 5, 10, 20, 28,$   
695  $50, 100, 222$  and 340).

696

697 **Hybrid prediction based on combining single-cell RNA-seq and single-cell ATAC-seq**

698 For the hybrid approach, we randomly sampled  $x$  ( $x = 22, 72, 194$  and  $312$ ) cells from the single-cell  
699 ATAC-seq dataset 2. Dataset 2 was used since it performed better than dataset 1 based on analyses in  
700 **Figure 4b**. We obtained the average ATAC-seq profile of the sampled cells using the protocol described  
701 above. We also obtained BIRD-predicted DH from pooled single-cell RNA-seq using 28 cells. The average  
702 of the ATAC-seq profile and BIRD predicted DH profile was then computed. The total number of cells  
703 used by this hybrid approach was  $k = x+28$  (i.e.,  $k= 50, 100, 222$  and  $340$ ). In **Figure 4b**, this hybrid  
704 approach was compared to pooled single-cell ATAC-seq using the same number of cells. For the hybrid  
705 approach, the sampling of cells from scATAC-seq was repeated 10 times. The mean and SD of results  
706 from the 10 analyses were shown in **Figure 4b**.

707

708 **ACKNOWLEDGMENTS**

709 This research is supported by grants from the National Institutes of Health (R01HG006282 and  
710 R01HG006841) and a Seed Fund of the Johns Hopkins Institute for Data Intensive Engineering and  
711 Science.

712

713

714

715

716

717

718

719

720

721

722

723 **FIGURE LEGENDS**

724 **Figure 1.** BIRD predicts DH and TFBSSs using bulk RNA-seq.

725 (a) Overview of the study. Roadmap Epigenomics DNase-seq and RNA-seq data are used to train BIRD  
726 prediction models which are then applied to new RNA-seq samples to predict DH. The predicted DH can  
727 be coupled with DNA motifs to predict TFBSSs.

728 (b) Two examples of true and predicted DH signals across five different samples. Each track is a sample.  
729 Regions highlighted with boxes demonstrate that the predicted DH captures the true DH variation.

730 (c) Statistics used to evaluate prediction performance.

731 (d)-(f) Prediction performance of BIRD and random prediction models (“BIRD-permute”) in leave-one-  
732 cell-type-out cross-validation.

733 (d) Distribution and mean of cross-locus correlation  $r_L$  from all samples.

734 (e) Distribution and mean of cross-sample correlation  $r_C$  from all loci.

735 (f) Squared prediction error ( $\tau$ ).

736 (g) Genomic loci are grouped into four categories by coefficient of variation (CV) of the predicted DH  
737 across samples at each locus. Distribution of  $r_C$  of all loci, stratified using the four CV categories, is shown  
738 for BIRD.

739 (h) Distribution and mean of  $r_C$  in each CV category.

740 (i) Distribution of  $r_C$  for locus-level predictions vs. pathway-level predictions.

741 (j)-(k) Sensitivity-rank curve for predicting YY1 binding sites in GM12878 and JUN binding sites in K562  
742 cells using true DNase-seq (“True”), BIRD, and mean DH profile of training samples (“Mean”). For each  
743 method, the curve shows the percentage of true TFBSSs discovered by top predicted motif sites.  $q$ -values  
744 corresponding to top 5000, 15000, and 25000 BIRD predictions are shown on top of each plot.

745

746 **Figure 2.** Predicting DH using small-cell-number RNA-seq data.

747 (a) Cross-locus correlation between the bulk GM12878 DNase-seq signal and chromatin accessibility  
748 predicted or measured by different methods. “Mean”: mean DH profile of training samples. “BIRD-b10”,  
749 “BIRD-b30”, “BIRD-b100”: BIRD-predicted DH based on small-cell-number RNA-seq samples with 10, 30

750 and 100 cells. “BIRD-bulk”: BIRD-predicted DH based on bulk RNA-seq. “ATAC-b500”, “ATAC-b50k”:  
751 ATAC-seq with 500 and 50,000 cells. “BIRD-b30+ATAC-b500”, “BIRD-b30+ATAC-b50k”: average of BIRD-  
752 predicted DH from 30 cells and ATAC-seq from 500 or 50,000 cells. “Mean+ATAC-b500”, “Mean+ATAC-  
753 b50k”: average of mean DH profile of training samples and ATAC-seq from 500 or 50,000 cells.  
754 “H3K27ac-b100”, “H3K27ac-b600”, “H3K4me3-b100” and “H3K4me3-b600”: MOWChIP-seq for histone  
755 modification H3K27ac or H3K4me3 with 100 or 600 cells.

756 (b) An example that compares chromatin accessibility predicted or measured by different methods. True  
757 bulk DNase-seq signal is shown on the bottom track as a reference. Regions highlighted by boxes  
758 illustrate that BIRD predicted DH better than “Mean” and “ATAC-b500”.

759 (c)-(h) Scatterplots comparing true bulk DNase-seq signal with chromatin accessibility predicted or  
760 measured by ATAC-b50k, ATAC-b500, BIRD-b30, BIRD-b30+ATAC-b500, H3K27ac-b600 and H3K4me3-  
761 b600. Each dot is a genomic locus. The cross-locus correlation is shown on top of each plot.

762

763 **Figure 3.** Predicting TFBSs using small-cell-number RNA-seq data.

764 (a)-(f) Sensitivity-rank curve for predicting E2F4, MAX, SPI1, ELF1, RFX5 and USF2 binding sites in  
765 GM12878 using true DNase-seq (“True”), ATAC-seq from 500 or 50,000 cells (“ATAC-500”, “ATAC-b50k”),  
766 mean DH profile of training samples (“Mean”), BIRD-predicted DH using 30 cells (“BIRD-b30”), the  
767 average of BIRD-predicted DH using 30 cells and ATAC-seq using 500 cells (“BIRD-hybrid”), and  
768 MOWChIP-seq for H3K27ac and H3K4me3 using 600 cells (“H3K27ac-b600”, “H3K4me3-b600”). The  
769 performance for MOWChIP-seq using 100 cells was generally worse than using 600 cells and hence is  
770 shown in **Supplementary Figure 8** but not shown here for clarity of display. The *q*-values for BIRD-b30  
771 predictions are shown on the top of each plot.

772 (g) Scaled area under the curve (AUC) for different methods in TFBS prediction. Each row is a TF, and  
773 each column is a method. For each TF, different methods are ranked based on the AUC value, and the  
774 worst AUC value of all methods is shown on the right using a blue bar. The average rank of each method  
775 across all TFs is shown on the bottom using a red bar. Smaller rank means better performance.

776

777 **Figure 4.** BIRD predicts DH using pooled single-cell RNA-seq data.

778 (a) An example comparing chromatin accessibility reported by different single-cell methods. “ATAC1-  
779 sc10”, “ATAC1-sc28” and “ATAC1-sc222”: pooled single-cell ATAC-seq from 10, 28 or 222 cells using  
780 scATAC-seq dataset 1. “ATAC2-sc10”, “ATAC2-sc28” and “ATAC2-sc222”: pooled single-cell ATAC-seq  
781 from 10, 28 or 222 cells using scATAC-seq dataset 2. “BIRD-sc10”, “BIRD-sc28”: BIRD-predicted DH  
782 based on pooled single-cell RNA-seq data from 10 or 28 cells. “BIRD-hybrid-sc222”: the average of BIRD-  
783 sc28 and single-cell ATAC-seq from 194 cells using scATAC-seq dataset 2. As references, bulk ATAC-seq  
784 from 50,000 cells (“ATAC-b50k”) and DNase-seq are shown on the top and bottom respectively.

785 (b) Cross-locus correlation between the true bulk DNase-seq signal and chromatin accessibility predicted  
786 or measured by different single-cell methods. The correlation is shown as a function of pooled cell  
787 number. Error bars are standard deviation based on 10 independent samplings of cells (**Methods**).  
788 “ATAC1”: scATAC-seq dataset 1. “ATAC2”: scATAC-seq dataset 2. “BIRD”: BIRD-predicted DH using  
789 pooled single-cell RNA-seq. “BIRD-hybrid”: the average of BIRD-predictions based on 28 cells and pooled  
790 ATAC-seq from scATAC-seq dataset 2 (here x-axis is the total number of cells used by scRNA-seq and  
791 scATAC-seq). Prediction performance using the mean DH profile of training samples (“Mean”) is shown  
792 as a dashed line.

793 (c) Scatterplots comparing true bulk DNase-seq signal with chromatin accessibility predicted or  
794 measured by ATAC1, ATAC2 and BIRD (or BIRD-hybrid for 222 cells) using 10, 28 and 222 cells. Each dot  
795 is a genomic locus. The cross-locus correlation is shown on top of each plot.

796

797 **Figure 5.** BIRD predicts TFBSS using pooled single-cell RNA-seq data and a comparison between two  
798 different prediction strategies.

799 (a)-(f) Sensitivity-rank curve for predicting BHLHE40 and SP1 binding sites in GM12878 using true DNase-  
800 seq (“True”), mean DH profile of training samples (“Mean”), and BIRD and scATAC-seq by pooling  
801 different number of cells. (a,d) Pooled scATAC-seq and BIRD using 10 cells. (b,e) Pooled scATAC-seq and  
802 BIRD using 28 cells. (c,f) Pooled scATAC-seq and BIRD-hybrid using a total of 222 cells. “ATAC1” and

803 “ATAC2” correspond to two different scATAC-seq datasets. *q*-values shown in each plot are calculated  
804 based on BIRD-sc10, BIRD-sc28 and BIRD-hybrid-sc222 predictions, respectively.

805 (g) Cross-locus correlation between the true bulk DNase-seq signal and BIRD-predicted DH using two  
806 different prediction strategies in the HSMM dataset. Approach 1 (“5 cells(1)”, ..., “69 cells(1)”: pool  
807 scRNA-seq from multiple cells first and then use the pooled scRNA-seq to make predictions; Approach 2  
808 (“5 cells(2)”, ..., “69 cells(2)”: use scRNA-seq from each single cell to make prediction first and then pool  
809 predictions from different cells by averaging. Error bars are standard deviation based on 10 independent  
810 samplings of cells (**Methods**).

811 (h) Cross-locus correlation between the true bulk DNase-seq signal and BIRD-predicted DH using two  
812 different prediction strategies in GM12878. Approach 1 (“5 cells(1)”, ..., “28 cells(1)”) and Approach 2 (“5  
813 cells(2)”, ..., “28 cells(2)”) are the same as above.

814

815

816

817

818

819

820

821

822

823

824

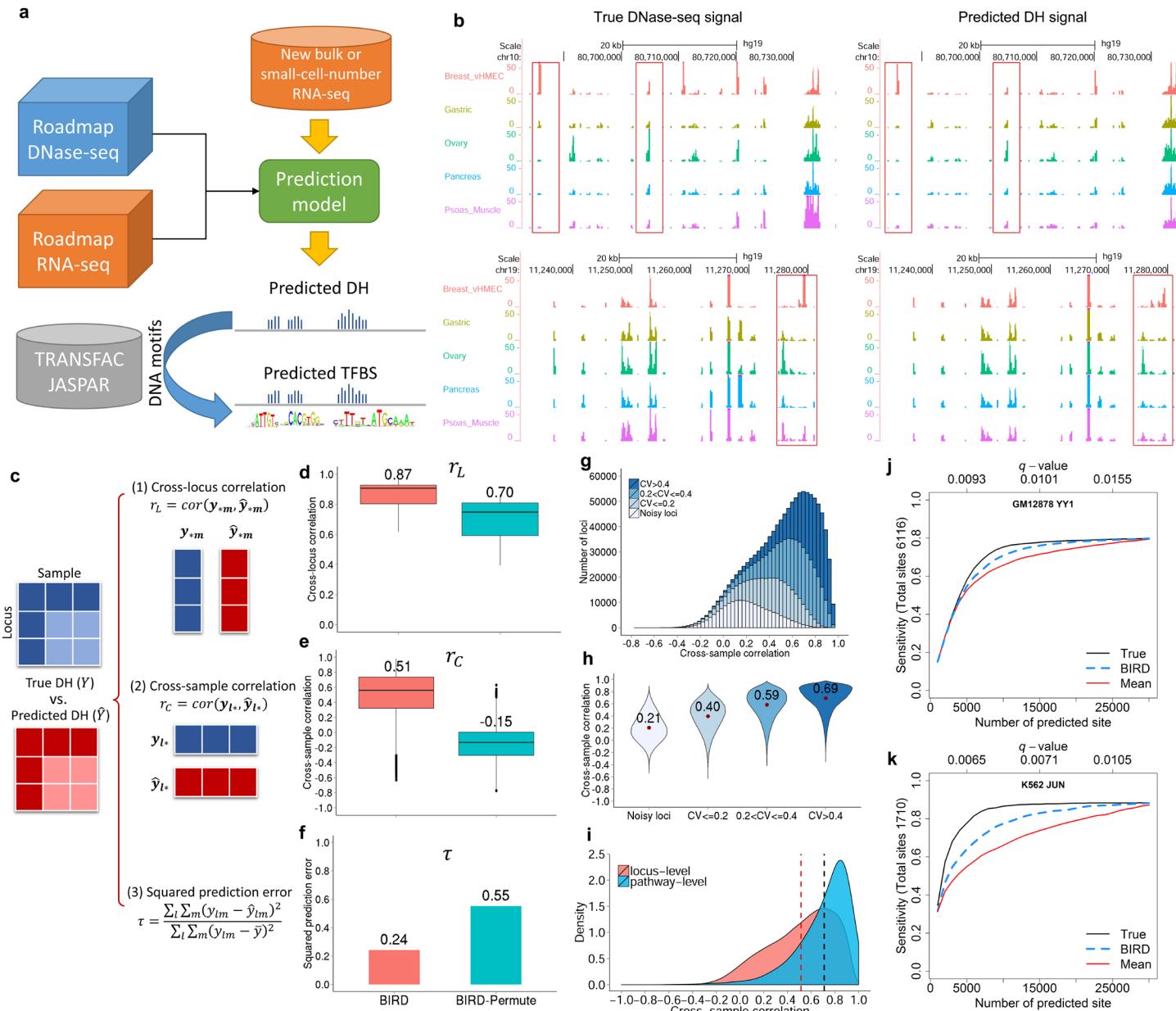
825

826

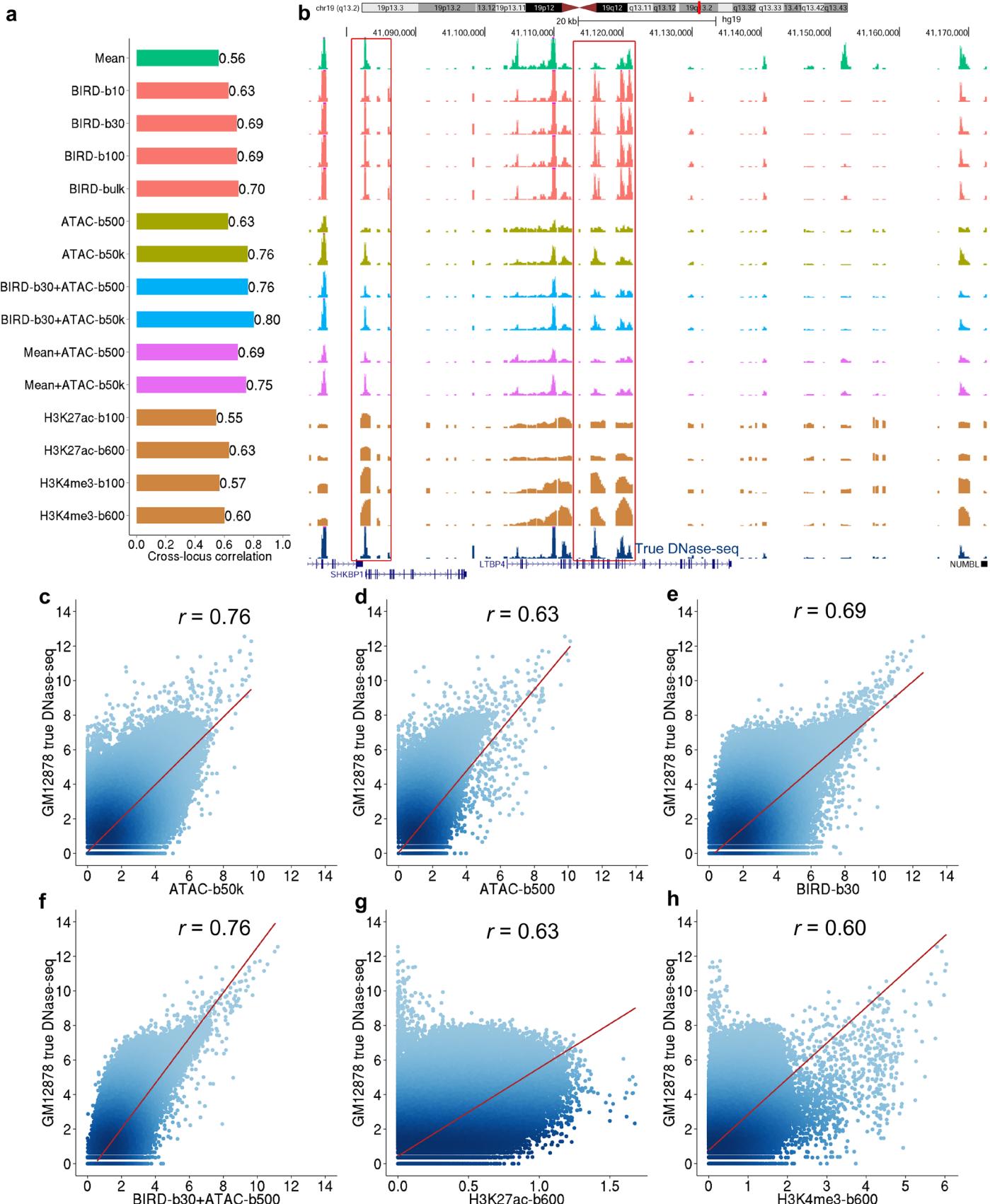
827

828

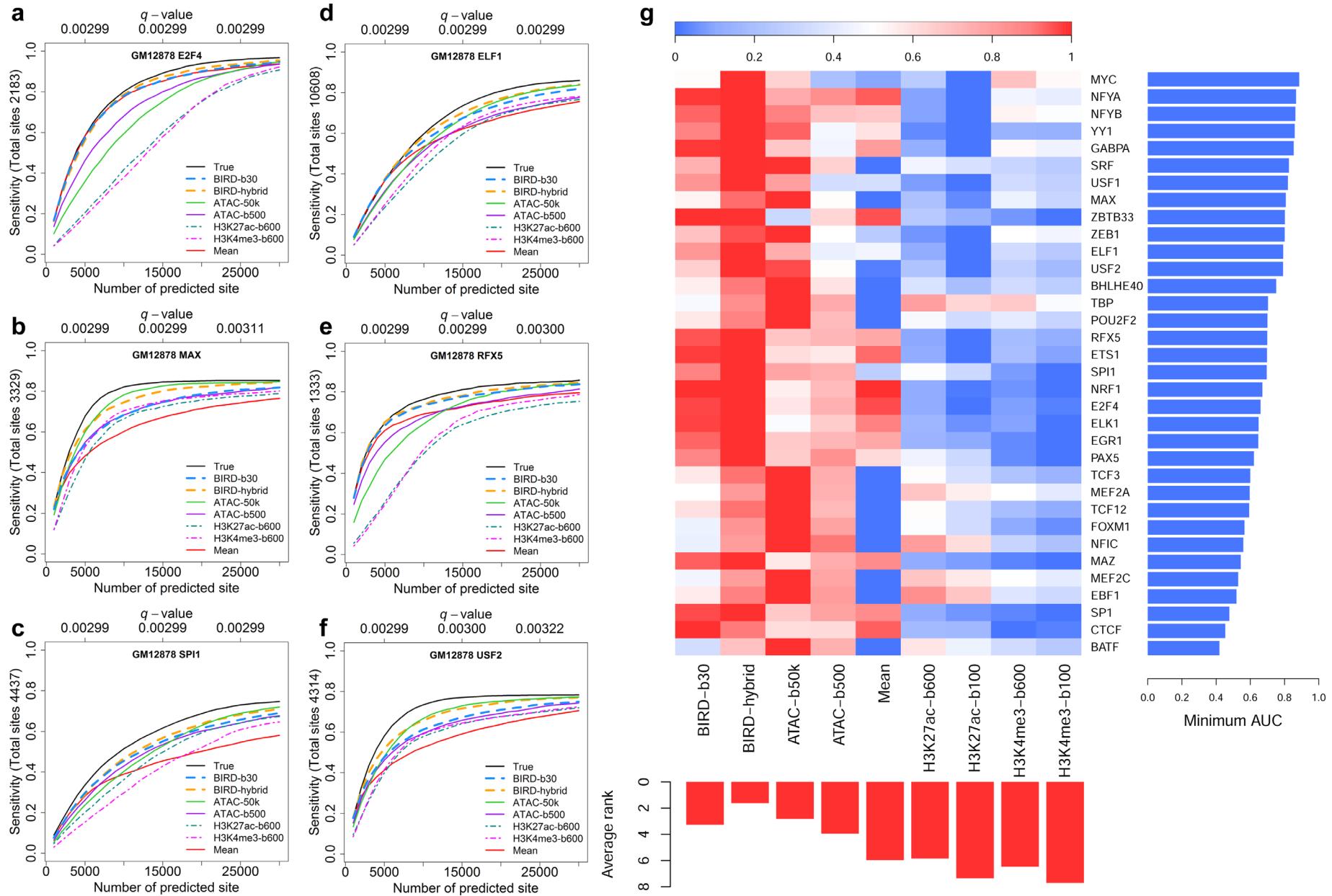
**Figure 1. BIRD predicts DH and TFBSS using bulk RNA-seq data.**



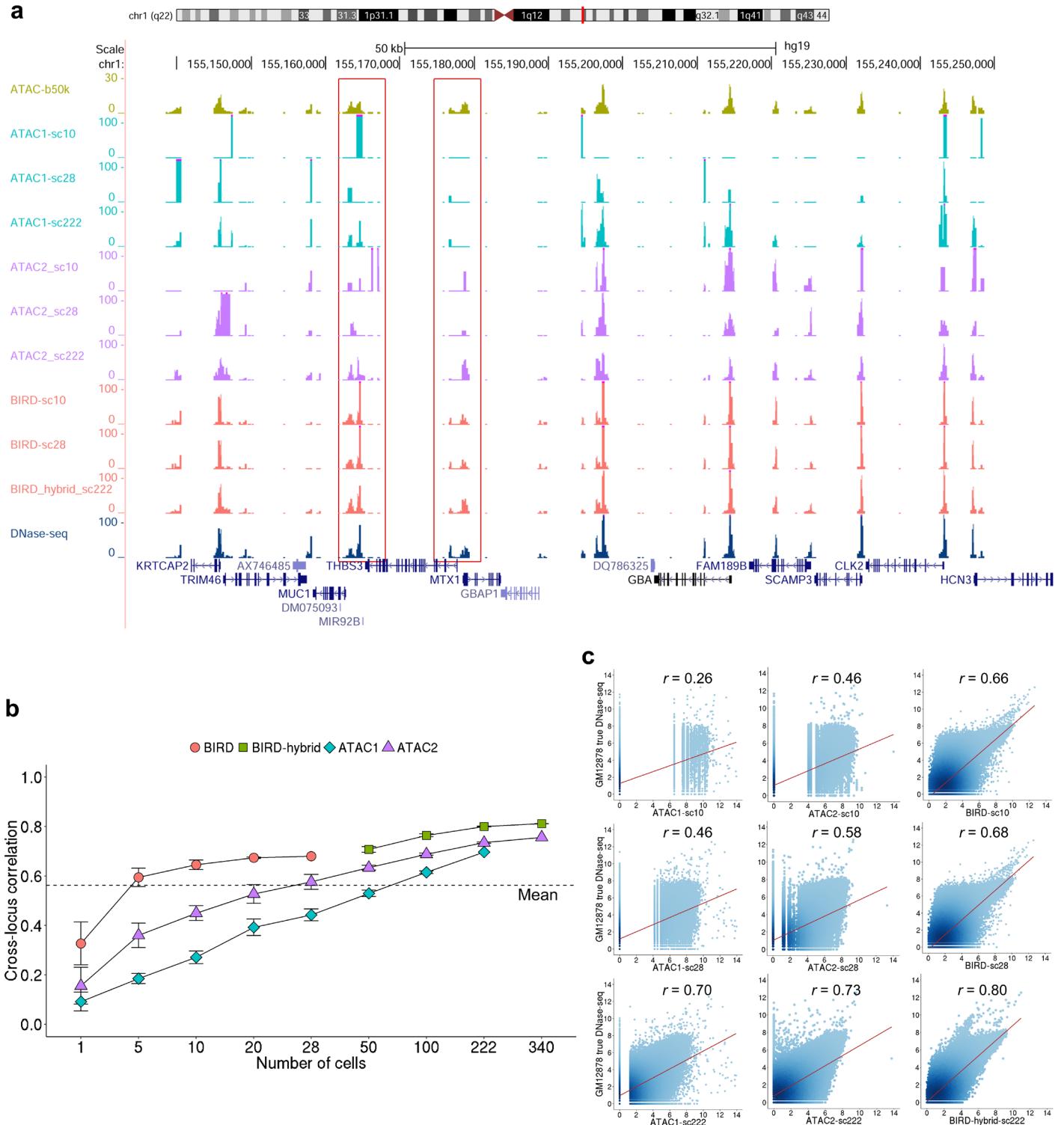
**Figure 2.** Predicting DH using small-cell-number RNA-seq data.



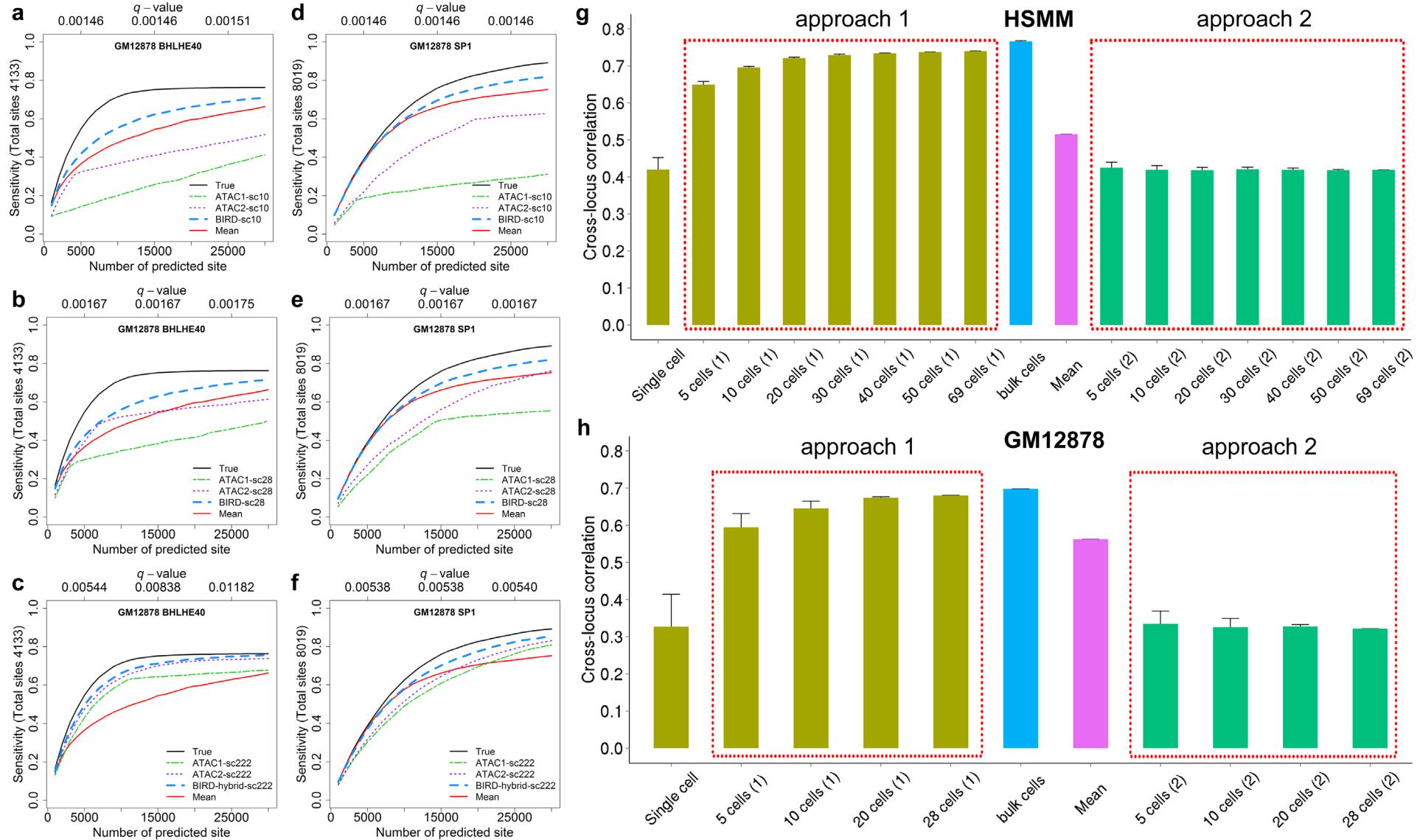
**Figure 3.** Predicting TFBSS using small-cell-number RNA-seq data.



**Figure 4. BIRD predicts DH using pooled single-cell RNA-seq data.**



**Figure 5.** BIRD predicts TFBSS using pooled single-cell RNA-seq data and a comparison between two different prediction strategies.



## REFERENCES

- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.
- 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.
- Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**: 486-490.
- Cao Z, Chen C, He B, Tan K, Lu C. 2015. A microfluidic device for epigenomic profiling using 100 cells. *Nature methods*.
- 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.
- Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**: 910-914.
- Dabney A, Storey JD. . qvalue: Q-value estimation for false discovery rate control. *R package version 1.40.0*.
- ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.
- 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.
- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760-1774.

- He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M. 2010. Nucleosome dynamics define transcriptional enhancers. *Nat Genet* **42**: 343-347.
- 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.
- Jin W, Tang Q, Wan M, Cui K, Zhang Y, Ren G, Ni B, Sklar J, Przytycka TM, Childs R, et al. 2015. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**: 142-146.
- Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497-1502.
- Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nature methods* **11**: 740-742.
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36-2013-14-4-r36.
- 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, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357-359.
- 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.
- Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* **24**: 496-510.
- 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.
- Ramsköld D, Luo S, Wang Y, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**: 777-782.
- Rotem A, Ram O, Shores N, Sperling RA, Goren A, Weitz DA, Bernstein BE. 2015. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol*.
- 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.
- Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA. 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell stem cell* **6**: 468-478.
- Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**: 381-386.
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515.
- Wilcoxon F. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*: 80-83.
- Zhou W, Sherwood B, Ji Z, Du F, Bai J, Ji H. submitted. Genome-wide Prediction of DNase I Hypersensitivity Using Gene Expression.