Supplementary Materials for:

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

Weiqiang Zhou[1], Zhicheng Ji[1], Hongkai Ji[1,*]

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

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

Contents

SUPPLEMENTARY METHODS 3

BIRD Model **3**

SUPPLEMENTARY REFERENCES 6

SUPPLEMENTARY TABLES 7

Supplementary Table 1 ..... **7**

Supplementary Table 2 ..... **7**

Supplementary Table 3 ..... **7**

Supplementary Table 4 ..... **7**

SUPPLEMENTARY FIGURES 8

Supplementary Figure 1 ..... **8**

Supplementary Figure 2 ..... **9**

Supplementary Figure 3 ..... **10**

Supplementary Figure 4 ..... **11**

Supplementary Figure 5 ..... **12**

Supplementary Figure 6 ..... **14**

Supplementary Figure 7 ..... **16**

Supplementary Figure 8 ..... **17**

Supplementary Figure 9 ..... **19**

Supplementary Figure 10 ..... **20**

Supplementary Figure 11 ..... **21**

Supplementary Figure 12 ..... **22**

Supplementary Figure 13 ..... **24**

Supplementary Figure 14 ..... **26**

Supplementary Figure 15 ..... **28**

SUPPLEMENTARY METHODS

BIRD Model

Problem formulation:

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

The function $f_l(\boldsymbol{X})$ is unknown. BIRD trains it using $\boldsymbol{X}$ and $\boldsymbol{Y}$ observed from a number of different samples. The training data are organized into two matrices: a gene expression matrix $\mathbb{X} = (x_{gc})_{G\times C}$ and a DH matrix $\mathbb{Y} = (y_{lc})_{L\times C}$. Rows in these matrices are genes and genomic loci respectively. Columns in these matrices are samples. C is the number of training samples. Each column of $\mathbb{X}$ and $\mathbb{Y}$ is a realization of the random vector $\boldsymbol{X}$ and $\boldsymbol{Y}$ in a specific sample.

Let a_l^y and s_l^y be the mean and standard deviation (SD) of the DH signals at locus l (i.e., row l of $\mathbb{Y}$), and let a_g^x and s_g^x be the mean and SD of the gene expression for gene g (i.e., row g of $\mathbb{X}$). Before training, each row of $\mathbb{X}$ in the training data is standardized to have zero mean and unit SD using a_g^x and s_g^x. Similarly, each row of $\mathbb{Y}$ in the training data is standardized to have zero mean and unit SD using a_l^y and s_l^y. The standardized matrices $\widetilde{\mathbb{X}}$ and $\widetilde{\mathbb{Y}}$ are then used to construct prediction models.

Once the prediction models are trained, they can be applied to new RNA-seq samples to make predictions. Let $\boldsymbol{X}$ be the expression profile of the new sample after being normalized to the stored quantiles of the training data. The expression value of each gene X_g in the new sample is first standardized using $\tilde{X}_g = (X_g - a_g^x)/s_g^x$. Recall that a_g^x and s_g^x are the mean and SD of the gene expression for gene g in the training data, and they have been computed and stored when BIRD was trained. Next, the trained model is applied to the standardized gene expression profile to make predictions. The predicted DH value for each locus, $\tilde{Y}_l$, 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 mean and SD of the DH signals for locus l in the training data. The unstandardized $\hat{Y}_l$ gives the prediction for Y_l, the DH level of genomic locus l in the new sample.

Training the prediction models:

BIRD is constructed in three steps.

- Step 1 – Construct elementary model $\mathrm{BIRD}_{\overline{\mathrm{X}},\mathrm{Y}}$: BIRD first groups correlated predictors into clusters. This is done by clustering rows of the standardized training data matrix $\widetilde{\mathbb{X}}$ into K clusters using k-means clustering (Hartigan and Wong 1979) (Euclidean distance used as similarity measure). Based on the clustering result, the gene expression profile $\widetilde{\boldsymbol{X}}$ of each sample is converted into a lower dimensional vector $\overline{\boldsymbol{X}} = (\bar{X}_{\boldsymbol{1}}, \dots, \bar{X}_{\boldsymbol{K}})$, where $\bar{X}_{\boldsymbol{k}}$ is the mean expression level of genes in cluster k. BIRD will use gene clusters' mean expression $\overline{\boldsymbol{X}}$ instead of the expression of individual genes $\widetilde{\boldsymbol{X}}$ as predictors to build prediction models. Clustering reduces the dimension and co-linearity of the predictors, and it also makes predictors less sensitive to measurement noise of individual genes. After clustering, the $G \times C$ matrix $\widetilde{\mathbb{X}}$ is converted into a $K \times C$ matrix $\overline{\mathbb{X}}$ ($G \approx 10^4$, $K \approx 10^2 \sim 10^3$)**.** The predictor dimension is reduced, but it is still high compared to sample size. We further reduce the predictor dimension by using a fast variable screening procedure: for each DHS locus l, the Pearson's correlation between its DH signal (i.e., row l of $\widetilde{\mathbb{Y}}$) and the expression of each gene cluster k (i.e., row k of $\overline{\mathbb{X}}$) across the training samples is computed, and the top N ($\approx 10^1$) clusters with the largest correlation coefficients are selected. Using the selected clusters ($\bar{X}_{\boldsymbol{l_1}}, \dots, \bar{X}_{\boldsymbol{l_N}}$) as predictors, a multiple linear regression $\tilde{Y}_{\boldsymbol{l}} = \beta_{l0} + \beta_{l1}\bar{X}_{\boldsymbol{l_1}} + \cdots + \beta_{lN}\bar{X}_{\boldsymbol{l_N}} + \epsilon_l$ is then fitted. Based on the fitted model, the standardized DH level of locus l in a new sample can be predicted by $\tilde{Y}_{\boldsymbol{l}} = f_l(\widetilde{\boldsymbol{X}}) = \beta_{l0} + \beta_{l1}\bar{X}_{\boldsymbol{l_1}} + \cdots + \beta_{lN}\bar{X}_{\boldsymbol{l_N}}$. The procedure above has two tuning parameters *K* (the number of gene clusters) and *N* (the number of predictors). We set *K*=3000 and *N*=4 when using all 70 samples as training dataset. These values were selected based on a pilot study using a random subset (1%) of the 1,136,465 studied loci. In the pilot study, different values of *K* and *N* were tested (*K*= 100, 200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500; *N*=1, 2, 3, 4, 5, 6, 7, 8). A five-fold cross-validation was performed by dividing the 70 samples into 5 groups (4 groups for training and 1 group for testing in each fold). The combination *K*=3000 and *N*=4 resulted in the highest mean cross-sample correlation r_C of all loci in the pilot set (note: cross-sample prediction performance is used as the primary criterion for parameter optimization since predicting cross-sample variation is more

challenging than predicting cross-locus variation as shown in **Fig. 1d-e**). This parameter combination was used in all subsequent models.

- Step 2 – Construct cluster-level model $\mathrm{BIRD}_{\overline{\mathrm{X}},\overline{\mathrm{Y}}}$: We further group correlated responses (i.e., genomic loci with similar DH patterns) into clusters. This is done by clustering rows of the standardized matrix $\widetilde{\mathbb{Y}}$ into H clusters using k-means clustering (Euclidean distance used as similarity measure). Based on the clustering result, the DH profile $\widetilde{\boldsymbol{Y}}$ of each sample can be converted into a lower dimensional vector $\overline{\boldsymbol{Y}} = (\overline{Y}_{\boldsymbol{1}}, \dots, \overline{Y}_{\boldsymbol{H}})$, where $\overline{Y}_{\boldsymbol{h}}$ is the mean DH level of DHSs in cluster h. Instead of predicting the DH level $\widetilde{\boldsymbol{Y}}$ of individual loci, $\mathrm{BIRD}_{\overline{\mathrm{X}},\overline{\mathrm{Y}}}$ uses the cluster-level gene expression $\overline{\boldsymbol{X}}$ to predict cluster-level DH $\overline{\boldsymbol{Y}}$. The prediction models are constructed using linear regression in a way similar to how the regression models are constructed in $\mathrm{BIRD}_{\overline{\mathrm{X}},\mathrm{Y}}$. For $\mathrm{BIRD}_{\overline{\mathrm{X}},\overline{\mathrm{Y}}}$, H was set to 1000, 2000 and 5000 respectively. For each H, we repeated the model training procedure.

- Step 3 – Construct the final prediction model BIRD: $\mathrm{BIRD}_{\overline{\mathrm{X}},\mathrm{Y}}$ is a special case of $\mathrm{BIRD}_{\overline{\mathrm{X}},\overline{\mathrm{Y}}}$ when DHSs are not clustered (i.e., $H = L$). In $\mathrm{BIRD}_{\overline{\mathrm{X}},\overline{\mathrm{Y}}}$, one may use the predicted cluster mean as the predicted DH level of each individual locus within the cluster. This will also generate a prediction for each locus. This locus-level prediction may be biased, but it is usually associated with smaller variance because cluster mean reduces variability. In the final BIRD model, multiple $\mathrm{BIRD}_{\overline{\mathrm{X}},\overline{\mathrm{Y}}}$ models with different H values are combined through model averaging to achieve a better bias-variance tradeoff and improve the prediction accuracy. Consider making predictions for a sample. Let $\mathcal{H}$ be the set of H values used by $\mathrm{BIRD}_{\overline{\mathrm{X}},\overline{\mathrm{Y}}}$. $\mathcal{H} = \{1000, 2000, 5000, L\}$. For each DHS locus l, let $\hat{Y}_l^{(H)}$ denote the locus-level DH predicted by $\mathrm{BIRD}_{\overline{\mathrm{X}},\overline{\mathrm{Y}}}$ using cluster number H. $\hat{Y}_l^{(L)}$ is the locus-level DH predicted by $\mathrm{BIRD}_{\overline{\mathrm{X}},\mathrm{Y}}$. The final BIRD prediction for locus l is a weighted average: $\frac{\sum_{\boldsymbol{H}\in\mathcal{H}} d_l^H \hat{Y}_l^{(H)}}{\sum_{\boldsymbol{H}\in\mathcal{H}} d_l^H}$**.** Here d_l^H is the weight. For a given cluster number *H*, the weight d_l^H is determined using training data as follows. Let $\widetilde{\boldsymbol{y}}_l = (\tilde{y}_{l1}, \dots, \tilde{y}_{lM})$ be the standardized locus-level DH for locus l observed in *M* training samples. Each locus l is associated with a cluster. Let $\widetilde{\boldsymbol{y}}_l^{(H)} = \left(\tilde{y}_{l1}^{(H)}, \dots, \tilde{y}_{lM}^{(H)}\right)$ represent the average of the standardized DH level of all loci within the cluster. d_l^H is the Pearson's correlation between the two vectors $\widetilde{\boldsymbol{y}}_l^{(H)}$ and $\widetilde{\boldsymbol{y}}_l$. Note that when $H = L$, $\mathrm{BIRD}_{\overline{\mathrm{X}},\overline{\mathrm{Y}}}$ reduces to $\mathrm{BIRD}_{\overline{\mathrm{X}},\mathrm{Y}}$, and therefore $\widetilde{\boldsymbol{y}}_l^{(L)} = \widetilde{\boldsymbol{y}}_l$ and $d_l^L = 1$. Consequently, the weight for $\mathrm{BIRD}_{\overline{\mathrm{X}},\mathrm{Y}}$ is 1.

SUPPLEMENTARY REFERENCES

Hartigan JA, Wong MA. 1979. Algorithm AS 136: A k-means clustering algorithm. *Applied statistics*: 100-108.

SUPPLEMENTARY TABLES

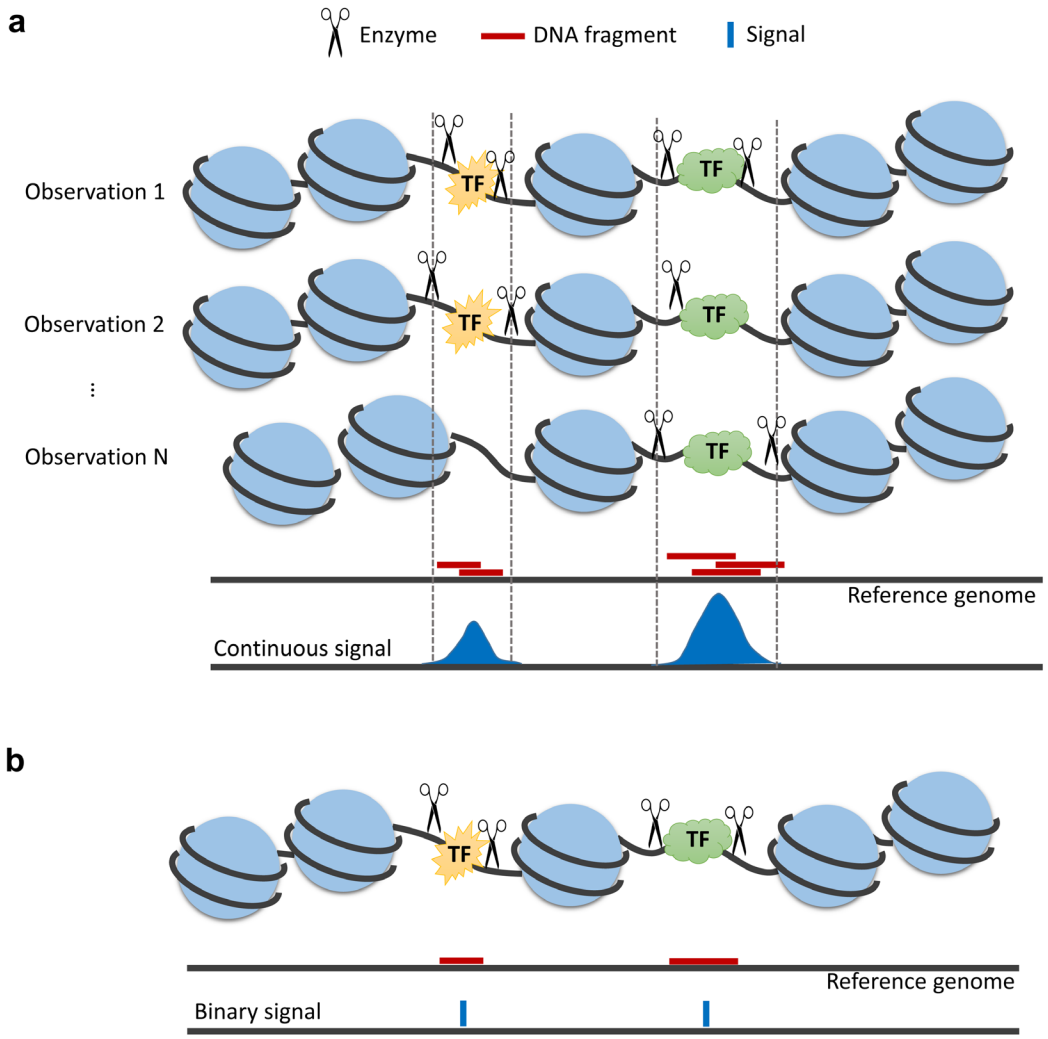
Supplementary Table 1. List of DNase-seq and RNA-seq data used for training and testing. The data contains 70 samples representing 30 different cell types. The table shows the cell types, GEO accession numbers (GSM number) of the DNase-seq and RNA-seq samples, and the donor identifiers. One cell type can contain multiple samples from different donors (see Supplementary_table_1.xlsx).

Supplementary Table 2. Transcription factor binding motifs and their TRANSFAC or JASPAR accession numbers used in this study (see Supplementary_table_2.xlsx).

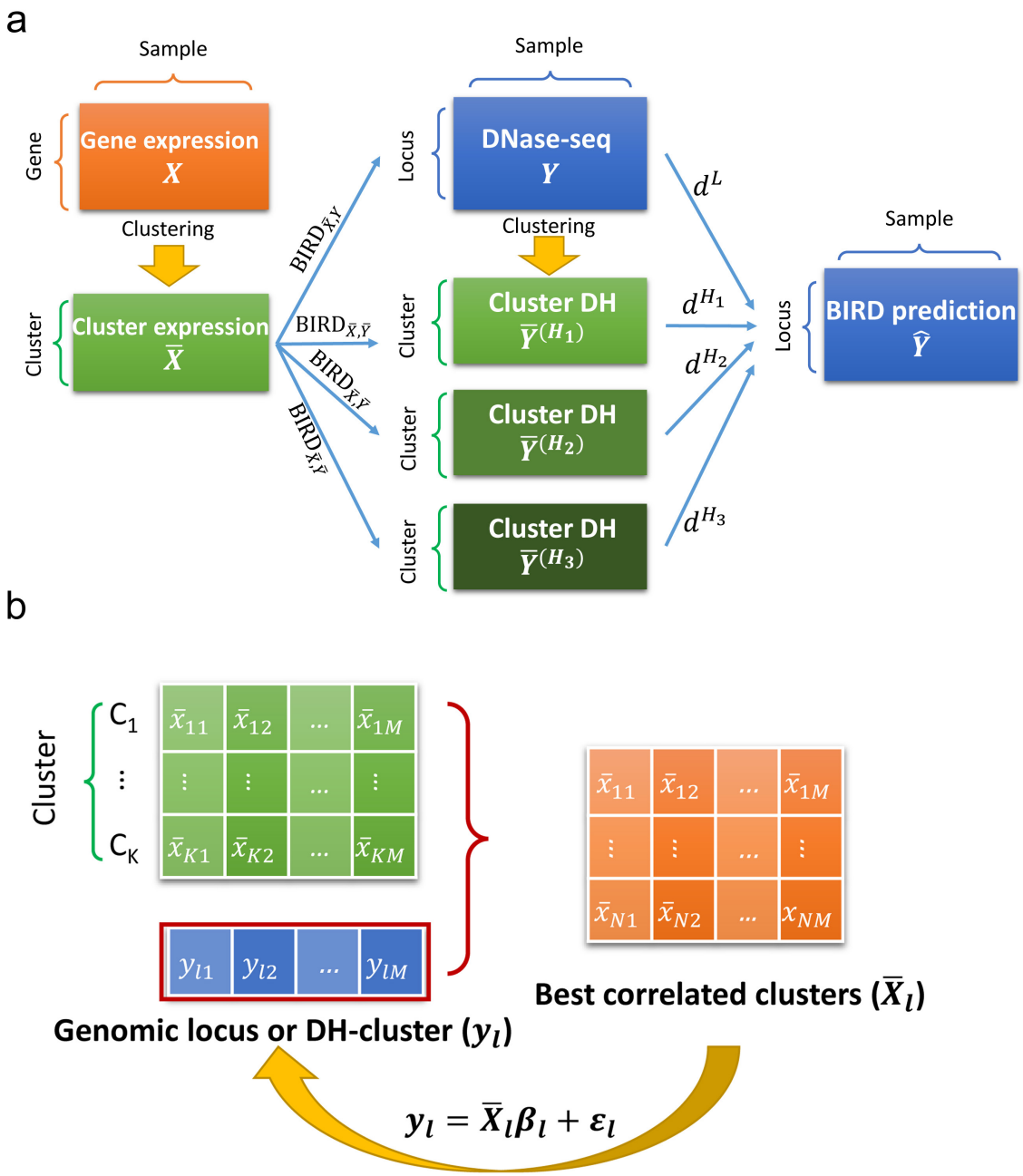
Supplementary Table 3. Scaled area under the curve (AUC) obtained from different TFBS prediction methods for 34 tested TFs in GM12878 using small number of cells (see Supplementary_table_3.xlsx).

Supplementary Table 4. Scaled area under the curve (AUC) obtained from different TFBS prediction methods for 34 tested TFs in GM12878 using pooled single cells (see Supplementary_table_4.xlsx).

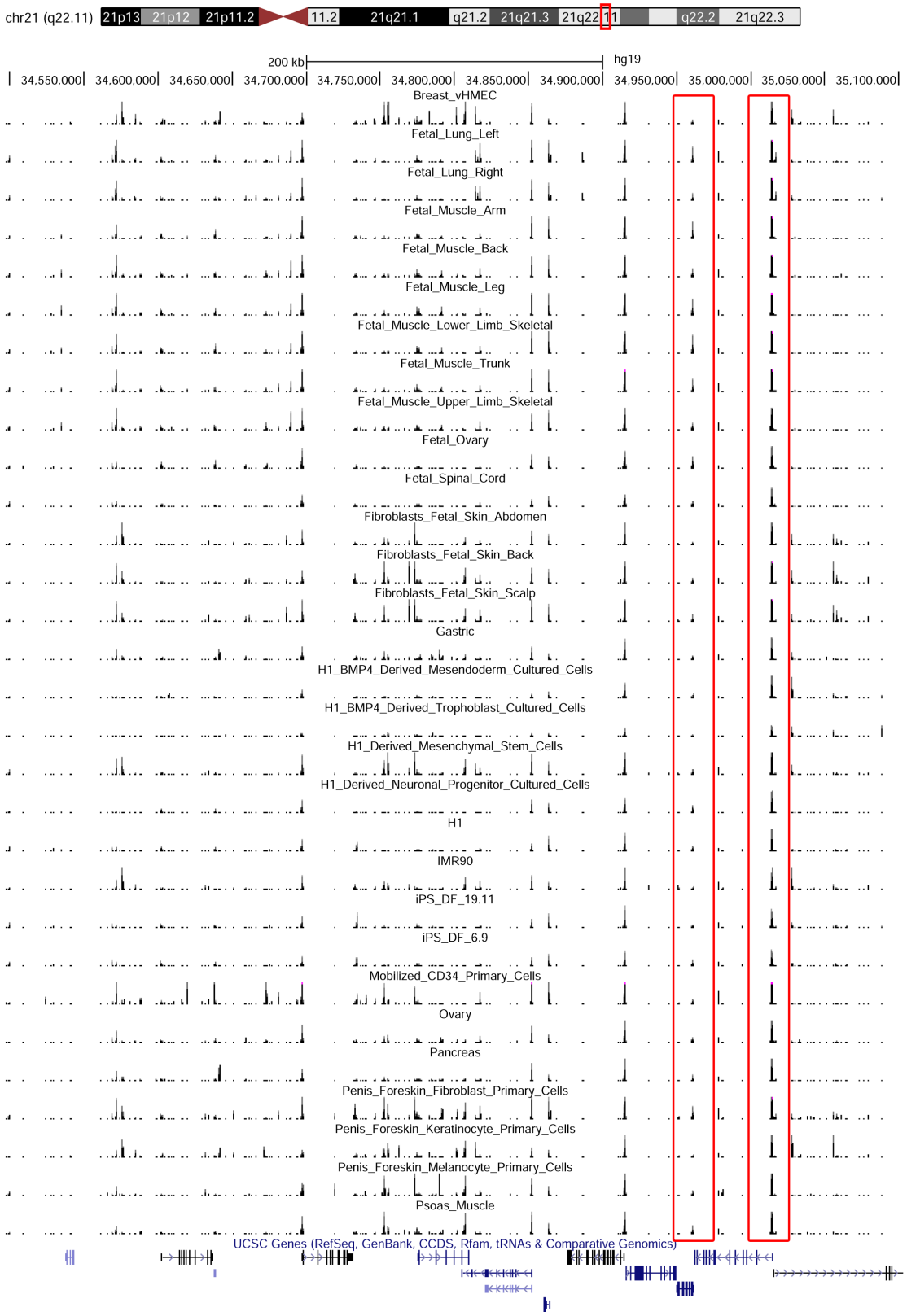
SUPPLEMENTARY FIGURES



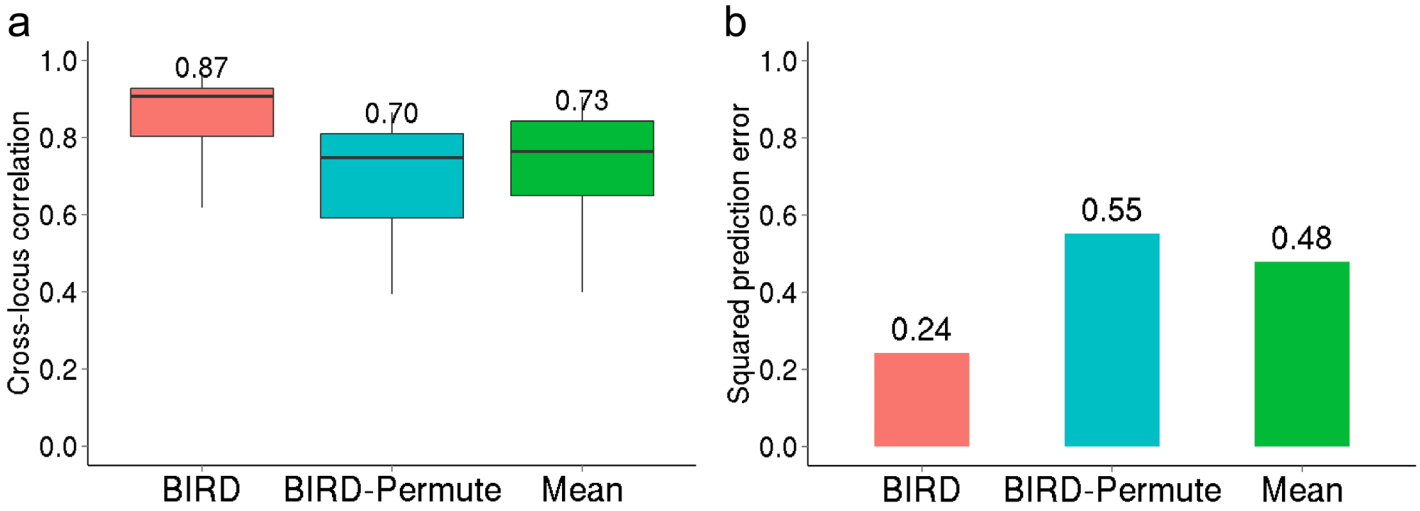
Supplementary Figure 1. Limitation of current high-throughput experimental technologies for mapping single-cell regulome. (**a**) Molecular events such as TF–DNA binding and dissociation, DNase I cleavage (DNase-seq), Tn5 insertion (ATAC-seq) are stochastic. For instance, a TF may bind to DNA at one time point and then dissociate from DNA at the next time point. If one observes a cell at a random time point, there is certain probability to observe TF binding, and there is also certain probability to observe that there is no TF binding. Similarly, if hypothetically one could apply Tn5 treatment to the same cell multiple times, Tn5 insertion may occur sometimes and may not occur at some other times. Chromatin accessibility reflects the probability that such molecular events occur at a genomic locus. Regions with higher accessibility tend to have higher probabilities to host such molecular events. Since probability is a continuous measure, chromatin accessibility is arguably a continuous signal. (**b**) Current single-cell regulome mapping technologies cannot repeatedly measure a cell multiple times. Instead, they can only measure a cell once since the cell is destroyed during the experiment. Moreover, each genomic locus only has two alleles that can be assayed. As a result, the measurements (i.e., event counts) are highly discrete. Such discrete signals obtained at a single time point are not adequate for describing the continuum of chromatin accessibility (i.e., the probability that molecular events occur at each locus). For instance, observing TF binding at a particular time point is not sufficient for describing the average steady-state TF binding behavior of a cell over time since TF binding may not occur at other times.



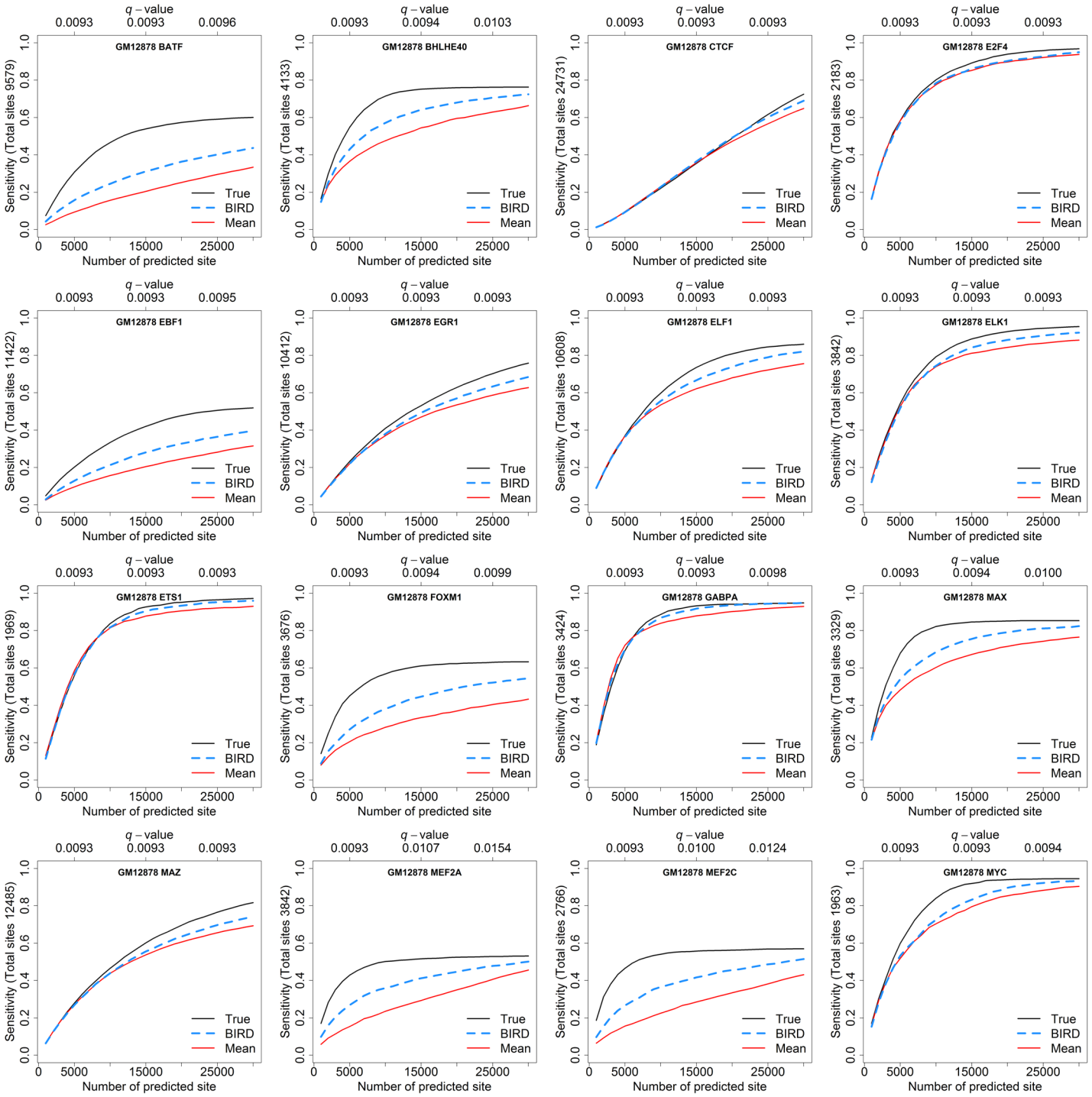
Supplementary Figure 2. Overview of the BIRD prediction model. (**a**) BIRD groups co-expressed genes into gene-clusters and uses the clusters' mean expression levels as predictors. BIRD also groups genomic loci (i.e., DHSs) with similar cross-sample DH variation patterns into DHS-clusters. Then, BIRD builds regression models both for predicting the DH level at each genomic locus (i.e. $\mathrm{BIRD}_{\bar{X},Y}$) and for predicting the mean DH level of each DHS-cluster (i.e. $\mathrm{BIRD}_{\bar{X},\bar{Y}}$). The locus-level predictions and the cluster-level predictions are combined through model averaging to obtain the final predicted DH value for each locus. (**b**) For both $\mathrm{BIRD}_{\bar{X},Y}$ and $\mathrm{BIRD}_{\bar{X},\bar{Y}}$, predictors (i.e. gene-clusters) that best correlate with the response (i.e., DH level) are selected. The selected predictors are then used to build a linear regression model.



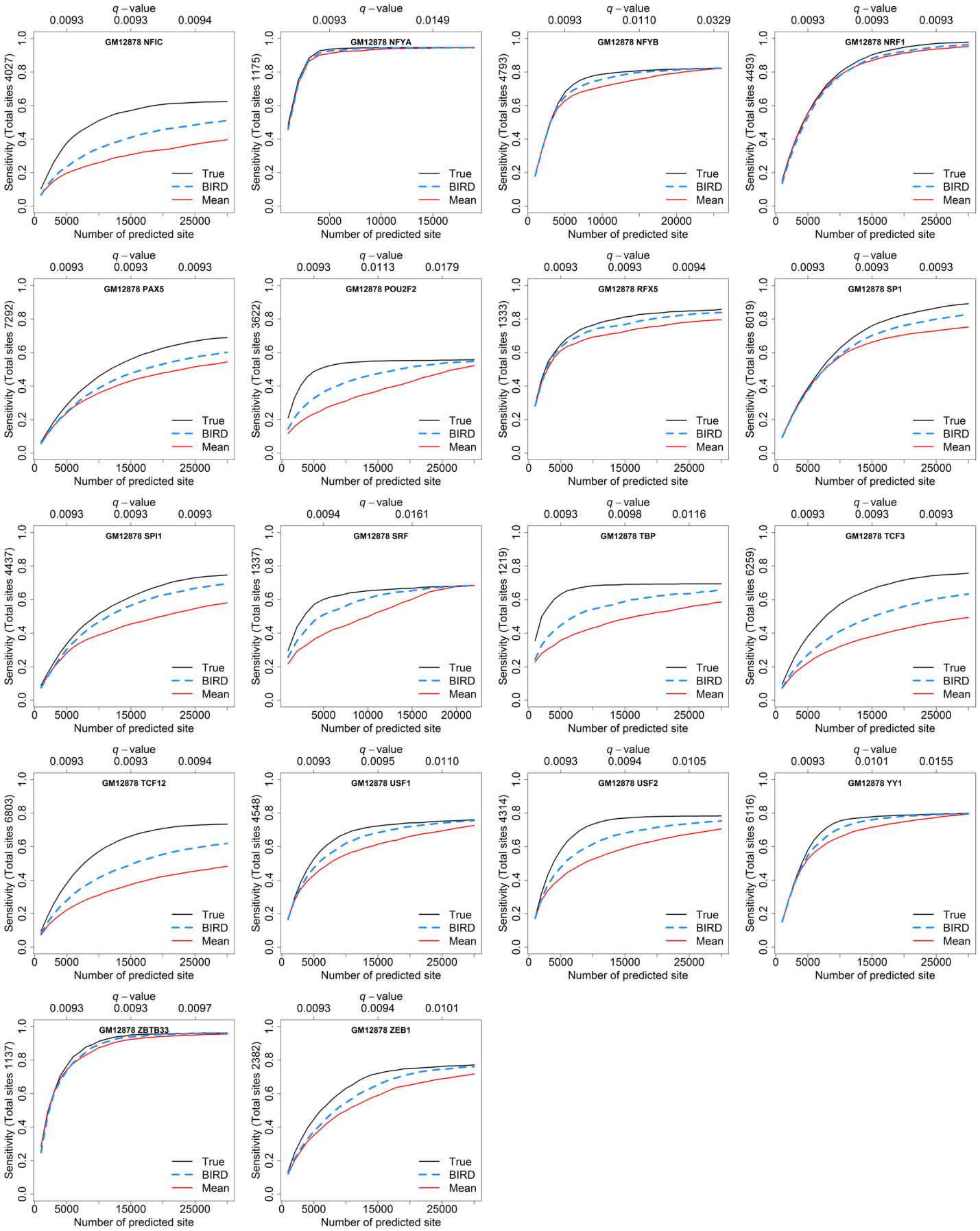
Supplementary Figure 3. Locus effects in DNase-seq data. DNase-seq signals from a genomic region in 30 different Epigenome Roadmap cell types are shown (only one sample is shown for each cell type). The highlighted regions show that DH levels at some genomic loci tend to be consistently higher than the DH levels at other loci in almost all cell types. As a result, the mean DH level across all cell types can predict cross-locus DH variation in a new sample with good accuracy. We note that permuting sample labels will not change the mean DH level at each locus and therefore will not perturb the locus effects.



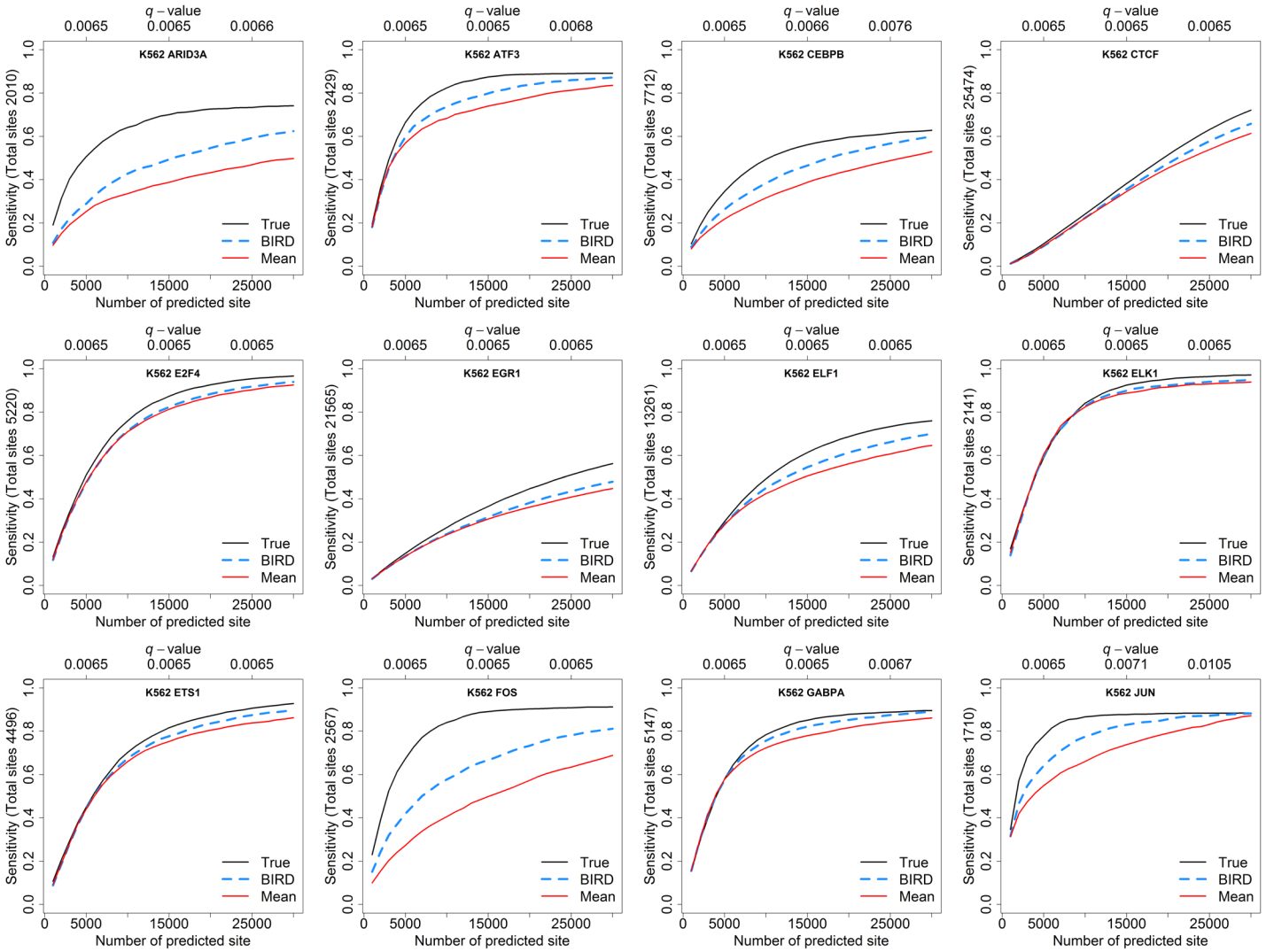
Supplementary Figure 4. Prediction performance for the mean DH profile of training samples ("Mean") in leave-one-out cross-validation. Performance for BIRD and random prediction models ("BIRD-Permute") are also shown as references. (**a**) Cross-locus correlation (r_L) between the true DNase-seq signal and the predicted DH for different methods. For each method, the boxplot shows the distribution of r_L from all samples. The number on top of the box shows the mean of r_L. (**b**) Squared prediction error (τ) for different methods. Cross-sample correlation (r_C) is not shown since it is zero for the mean DH profile (as its prediction remains the same for all new samples).



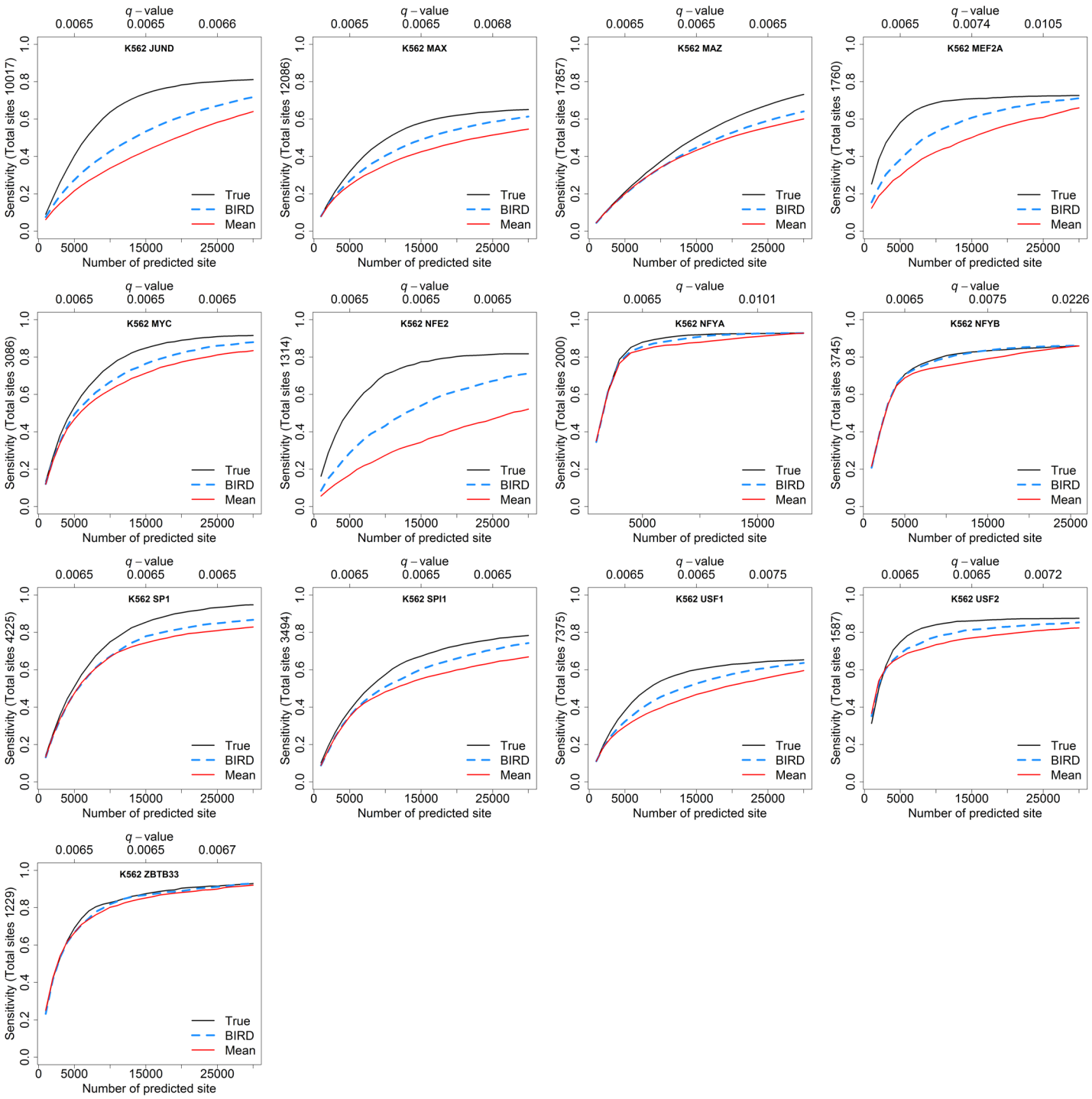
Supplementary Figure 5. Sensitivity-rank curves for predicting transcription factor binding sites in GM12878 using bulk RNA-seq. Predictions were made using the true DNase-seq data ("True"), BIRD-predicted DH based on bulk RNA-seq ("BIRD"), and mean DH profile of training samples ("Mean"). For each method, the curve shows the percentage of true TFBSs discovered by top predicted motif sites. *q*-values corresponding to top 5000, 15000, and 25000 BIRD predictions are shown on top of each plot.



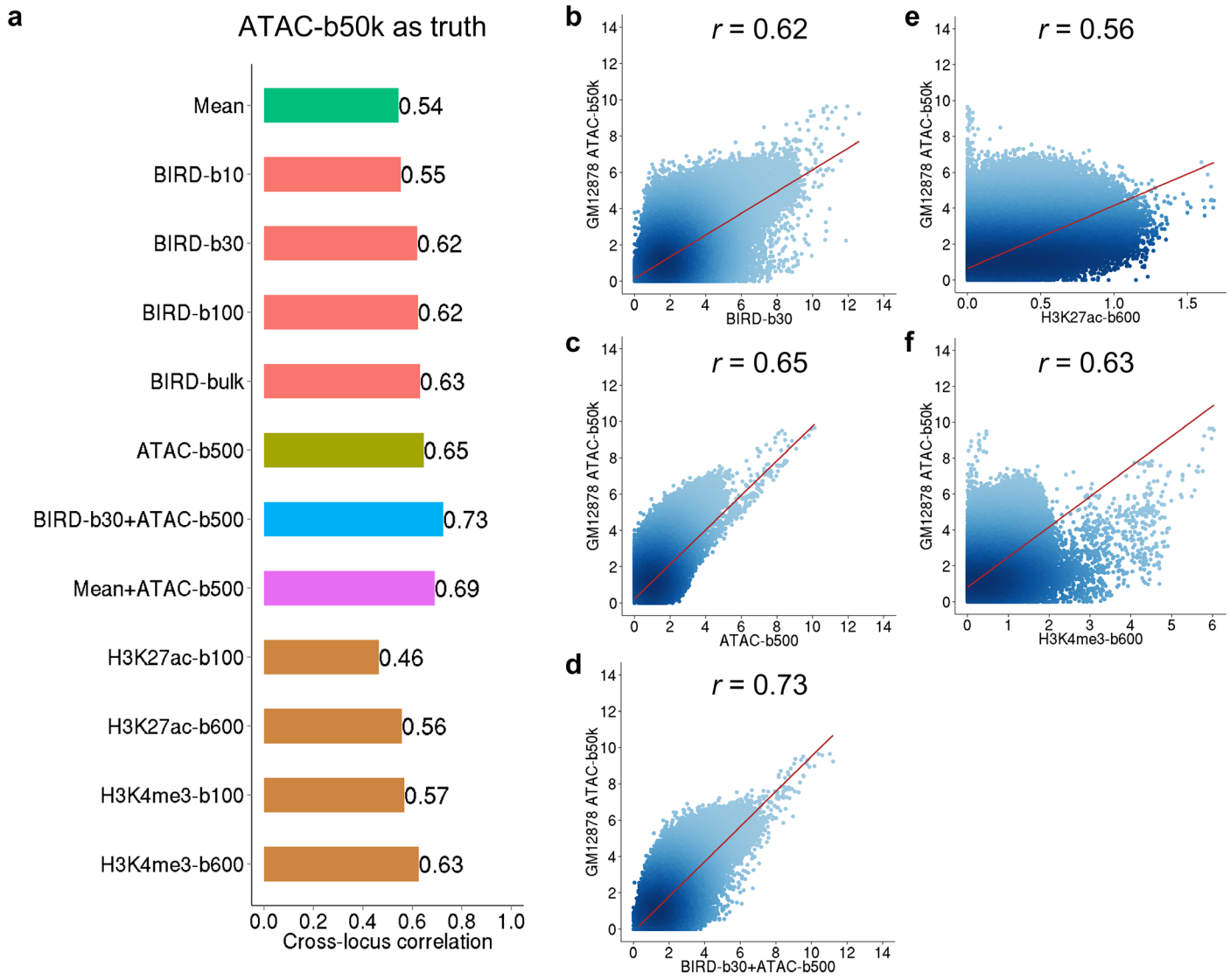
Supplementary Figure 5 (cont'd).



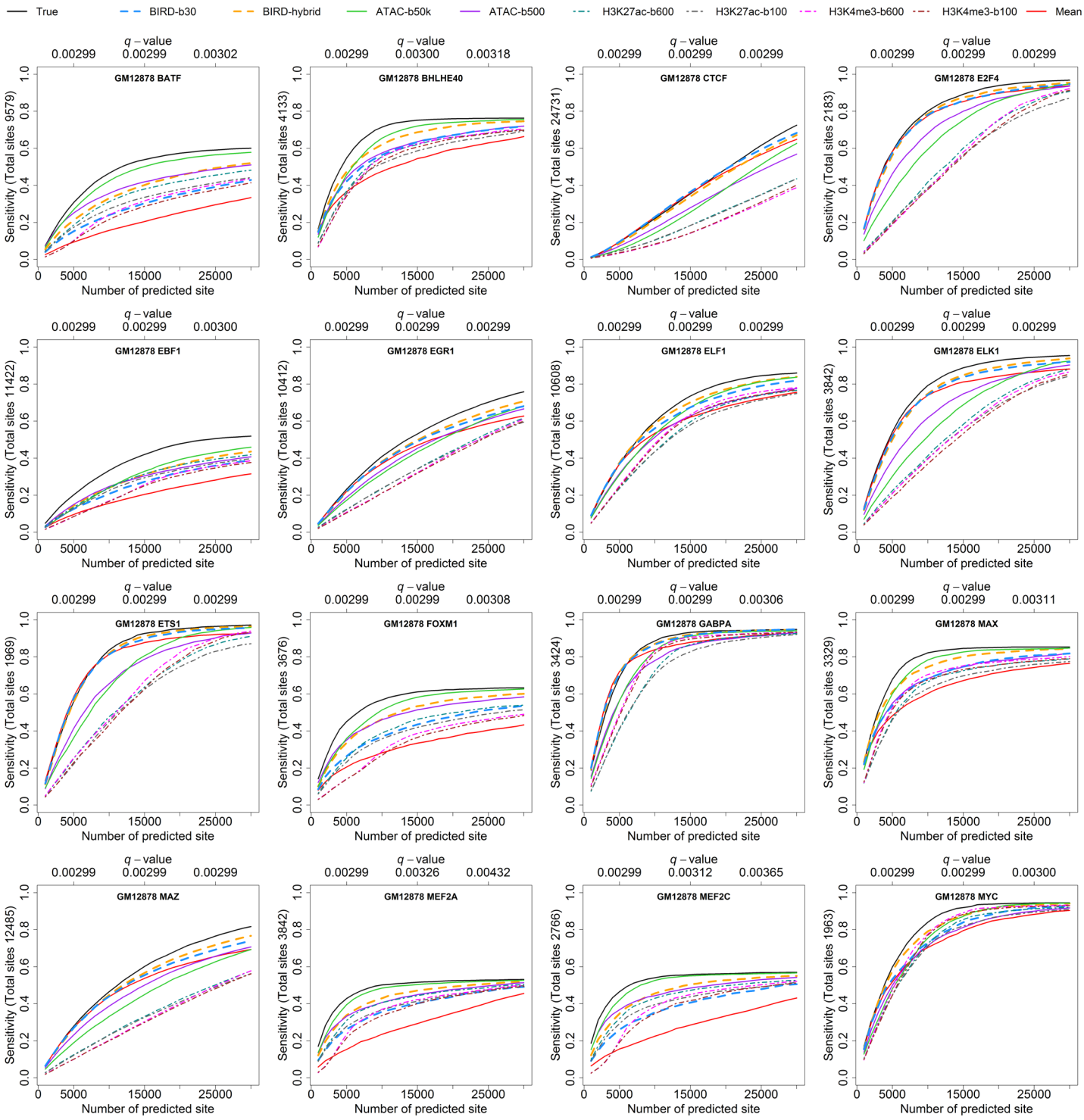
Supplementary Figure 6. Sensitivity-rank curves for predicting transcription factor binding sites in K562 using bulk RNA-seq. Predictions were made using the true DNase-seq data ("True"), BIRD-predicted DH based on bulk RNA-seq ("BIRD"), and mean DH profile of training samples ("Mean"). For each method, the curve shows the percentage of true TFBSs discovered by top predicted motif sites. *q*-values corresponding to top 5000, 15000, and 25000 BIRD predictions are shown on top of each plot.



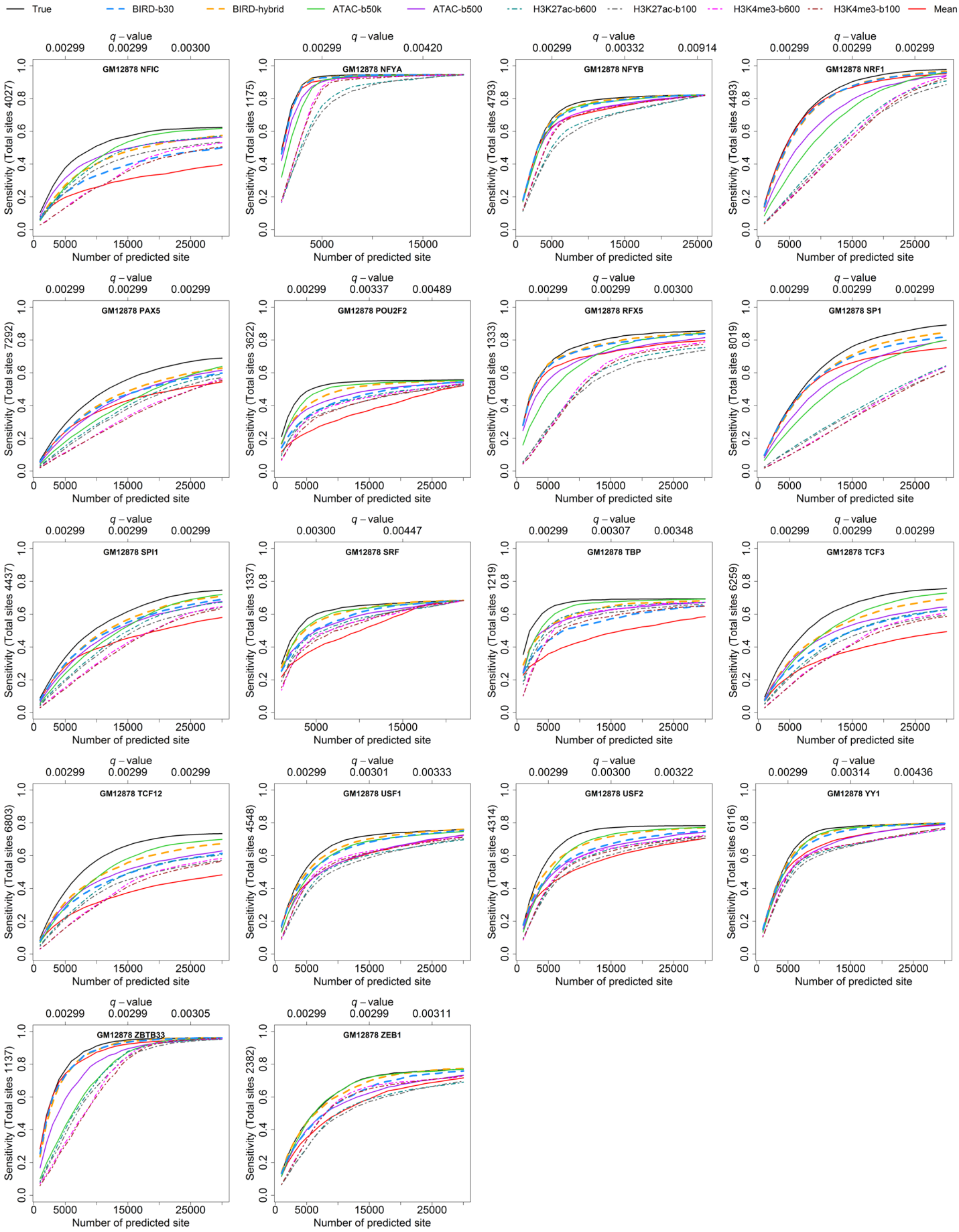
Supplementary Figure 6 (cont'd).



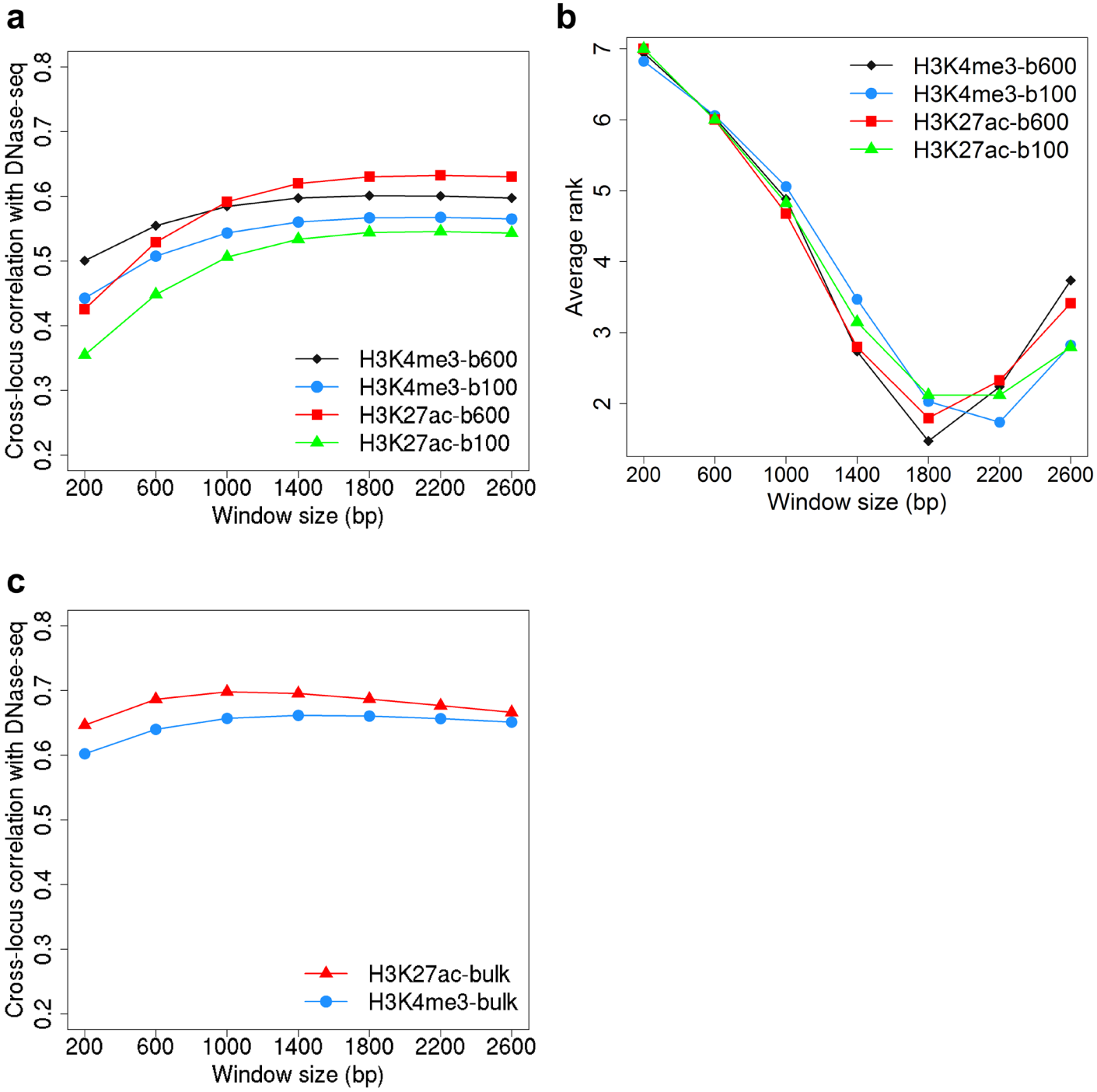
Supplementary Figure 7. Performance of predicting DH in GM12878 using small-cell-number RNA-seq. The performance is evaluated using ATAC-seq with 50,000 cells as gold standard. (**a**) Cross-locus correlation between ATAC-seq with 50,000 cells ("ATAC-b50k") and chromatin accessibility predicted or measured by different methods. "Mean": mean DH profile of training samples. "BIRD-b10", "BIRD-b30", "BIRD-b100": BIRD-predicted DH based on small-cell-number RNA-seq samples with 10, 30 and 100 cells. "BIRD-bulk": BIRD-predicted DH based on bulk RNA-seq. "ATAC-b500": ATAC-seq with 500 cells. "BIRD-b30+ATAC-b500: average of BIRD-predicted DH from 30 cells and ATAC-seq from 500 cells. "Mean+ATAC-b500": average of mean DH profile of training samples and ATAC-seq from 500 cells. "H3K27ac-b100", "H3K27ac-b600", "H3K4me3-b100" and "H3K4me3-b600": MOWChIP-seq for histone modification H3K27ac or H3K4me3 with 100 or 600 cells. (**b**)-(**f**) Scatterplots comparing ATAC-b50k with chromatin accessibility predicted or measured by BIRD-b30, ATAC-b500, BIRD-b30+ATAC-b500, H3K27ac-b600 and H3K4me3-b600. Each dot is a genomic locus. The cross-locus correlation is shown on top of each plot.



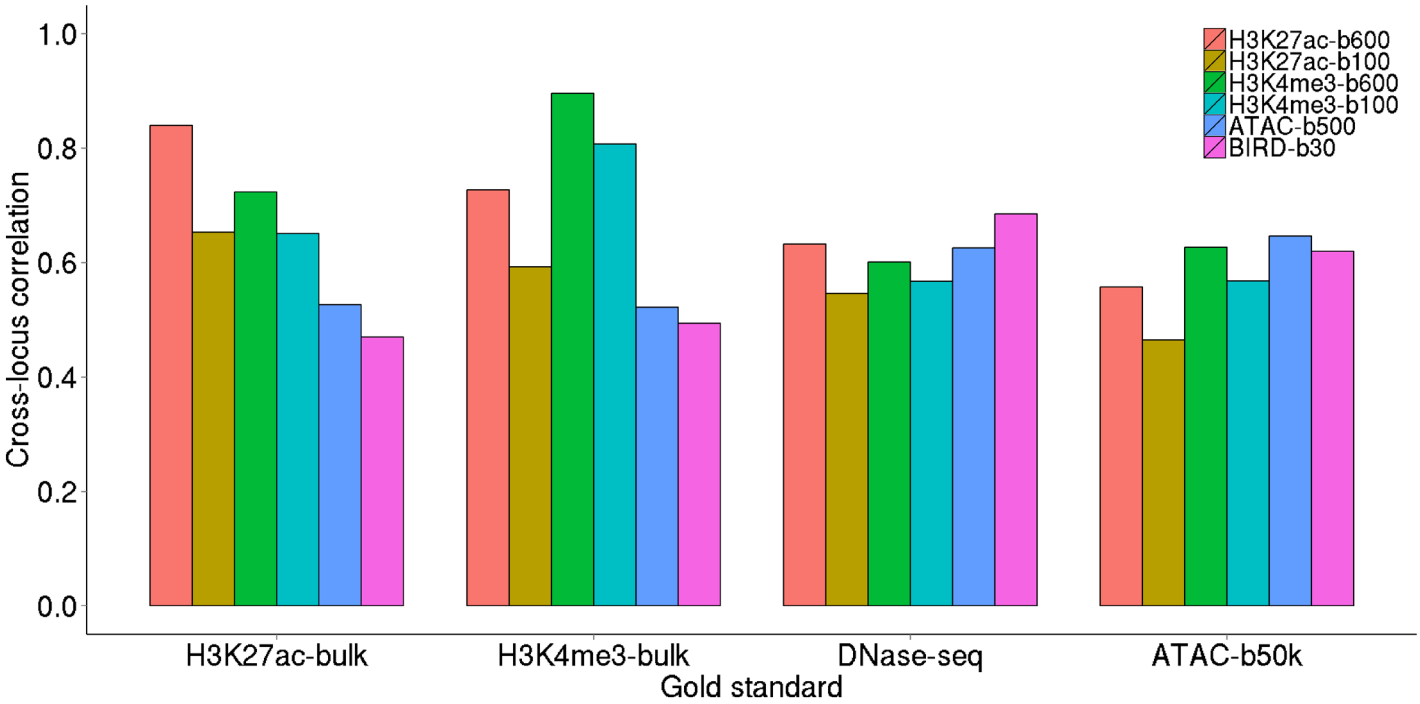
Supplementary Figure 8. Sensitivity-rank curves for predicting transcription factor binding sites in GM12878 using small-cell-number RNA-seq. Predictions were made using true DNase-seq data ("True"), mean DH profile of training samples ("Mean"), ATAC-seq from 500 or 50,000 cells ("ATAC-500", "ATAC-b50k"), BIRD-predicted DH using 30 cells ("BIRD-b30"), the average of BIRD-predicted DH using 30 cells and ATAC-seq using 500 cells ("BIRD-hybrid"), and MOWChIP-seq for H3K27ac and H3K4me3 using 600 cells ("H3K27ac-b600", "H3K4me3-b600") or 100 cells ("H3K27ac-b100", "H3K4me3-b100"). For each method, the curve shows the percentage of true TFBSs discovered by top predicted motif sites. *q*-values corresponding to top 5000, 15000, and 25000 BIRD-b30 predictions are shown on top of each plot.



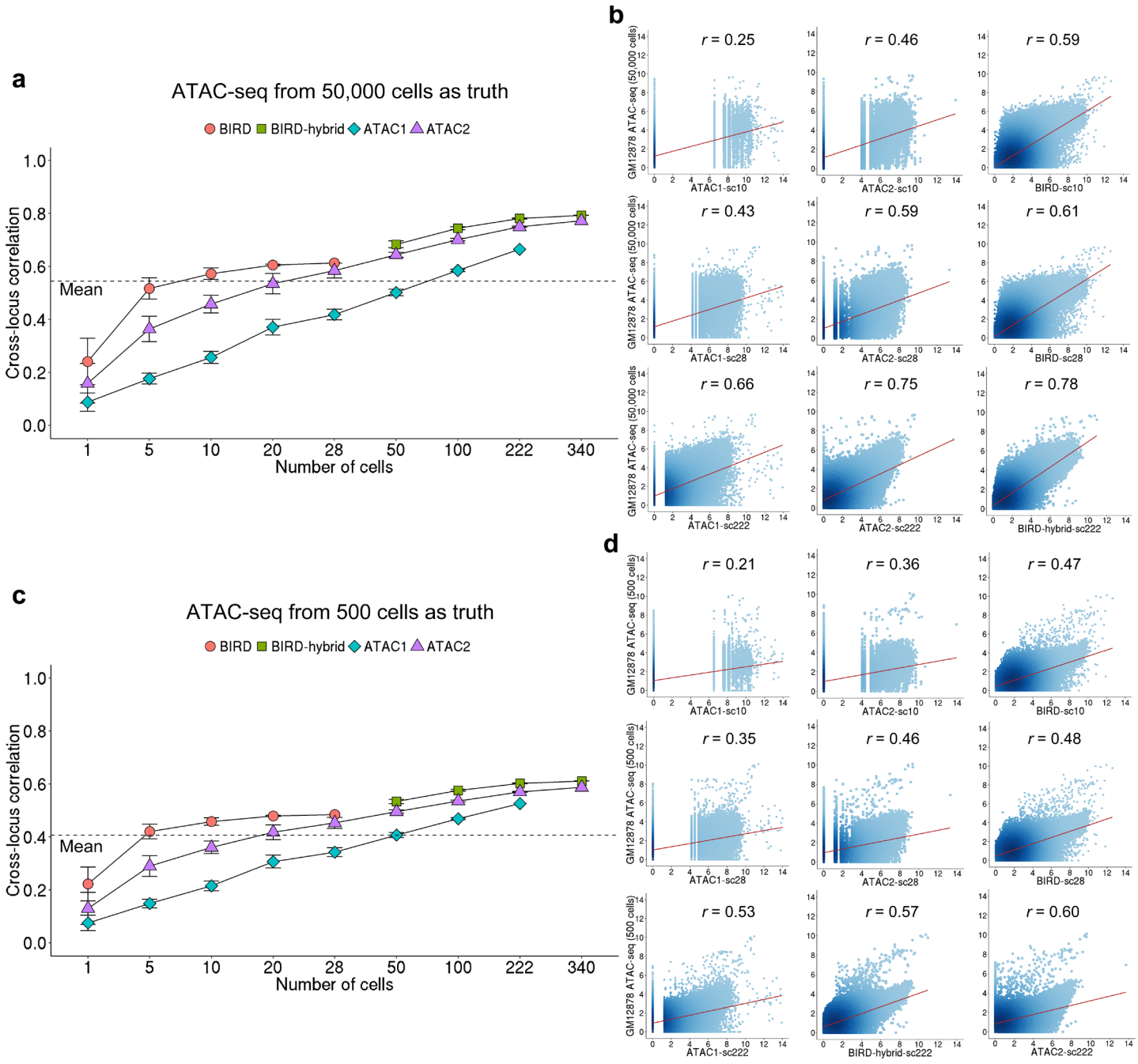
Supplementary Figure 8 (cont'd).



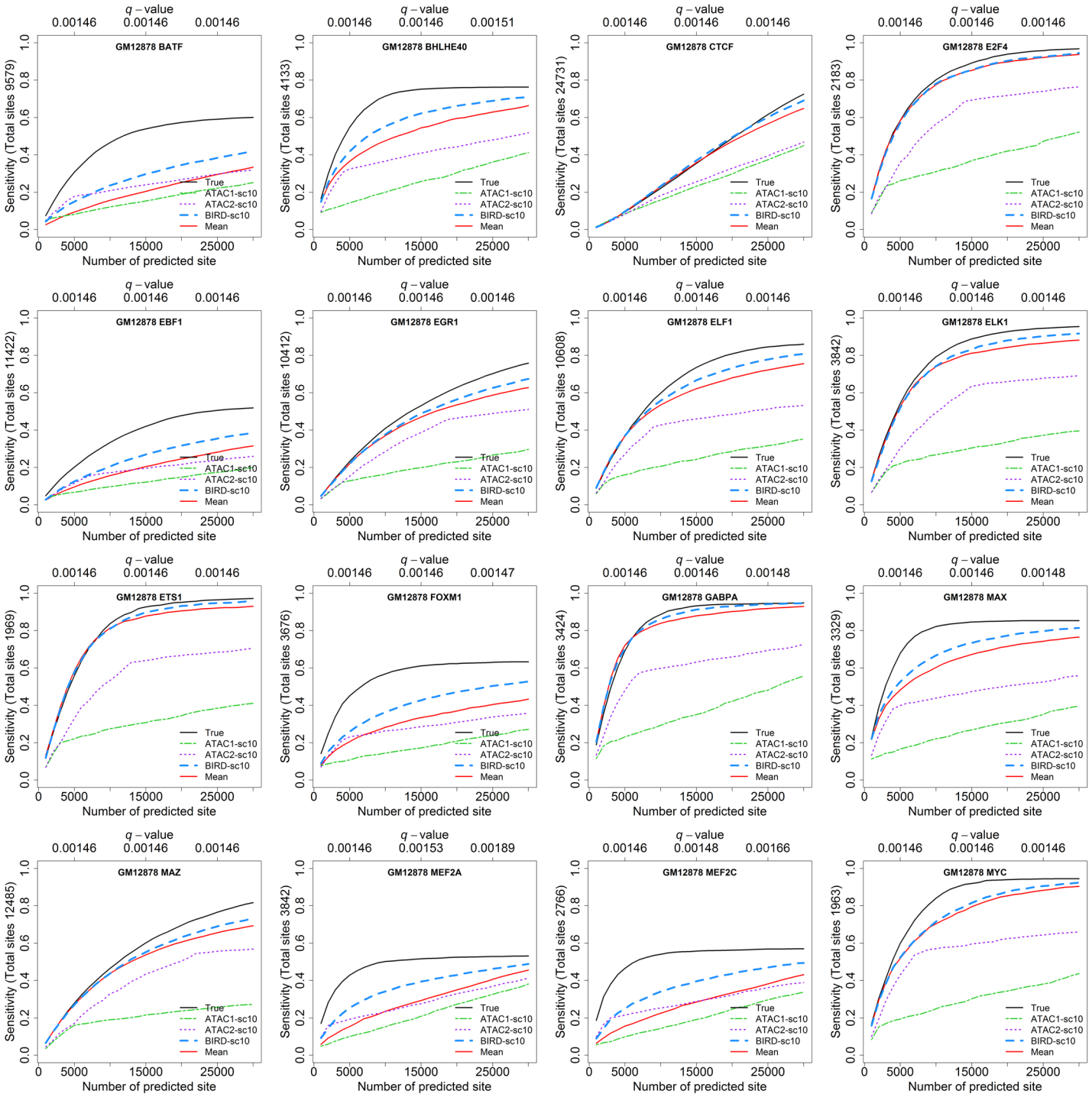
Supplementary Figure 9. Window size selection for analyzing histone modification MOWChIP-seq and ChIP-seq data. (**a**) Cross-locus correlation between MOWChIP-seq and bulk DNase-seq signal for different window sizes (*W*). The optimal *W* that maximizes the correlation was 2200 for H3K27ac with 100 and 600 cells ("H3K27ac-b100", "H3K27ac-b600"), 2200 for H3K4me3 with 100 cells ("H3K4me3-b100") and 1800 for H3K4me3 with 600 cells ("H3K4me3-b600"). (**b**) TFBS prediction performance for different window sizes using MOWChIP-seq (**Methods**). The optimal *W* that maximizes the TFBS prediction performance (in terms of the smallest average rank) was 1800 for H3K27ac-b100 and H3K27ac-b600, 2200 for H3K4me3-b100 and 1800 for H3K4me3-b600. (**c**) Cross-locus correlation between bulk ChIP-seq and bulk DNase-seq signal for different window sizes. The optimal *W* that maximizes the correlation was 1000 for bulk H3K27ac ("H3K27ac-bulk") and 1400 for bulk H3K4me3 ("H3K4me3-bulk").



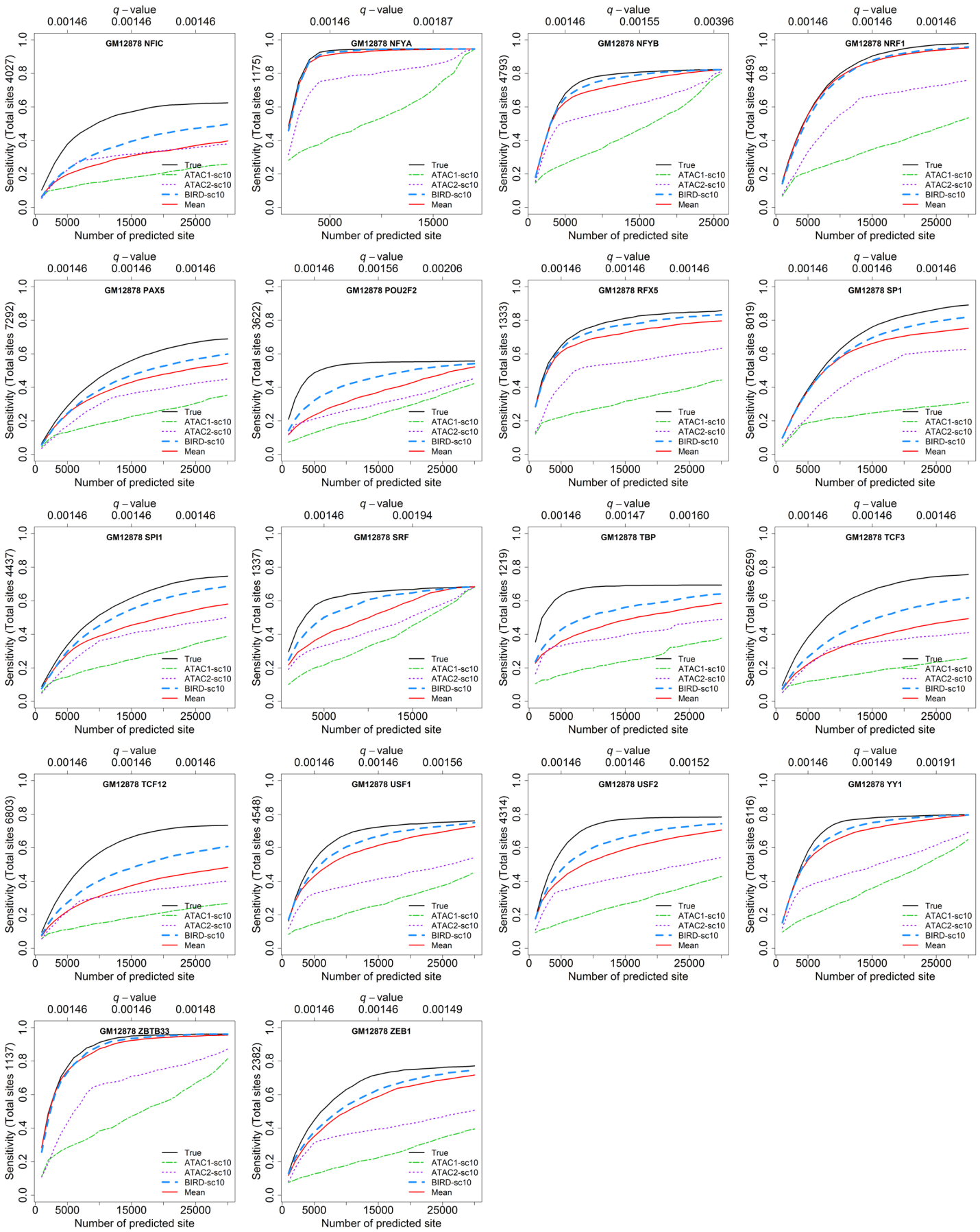
Supplementary Figure 10. Cross-locus correlation between gold standard signal in GM12878 and signals from small-cell-number MOWChIP-seq, ATAC-seq and BIRD. Different types of gold standard are used. Each group of bars corresponds to one type of gold standard. The gold standard signals include bulk H3K27ac ChIP-seq data ("H3K27ac-bulk"), bulk H3K4me3 ChIP-seq data ("H3K4me3-bulk"), bulk DNase-seq data ("DNase-seq") and ATAC-seq data from 50,000 cells ("ATAC-b50k"). Compared signals include MOWChIP-seq data for H3K27ac and H3K4me3 obtained from 100 and 600 cells ("H3K27ac-b100", "H3K27ac-b600", "H3K4me3-b100" and "H3K3me3-b600"), ATAC-seq from 500 cells ("ATAC-b500"), and BIRD predictions based on RNA-seq from 30 cells ("BIRD-b30").



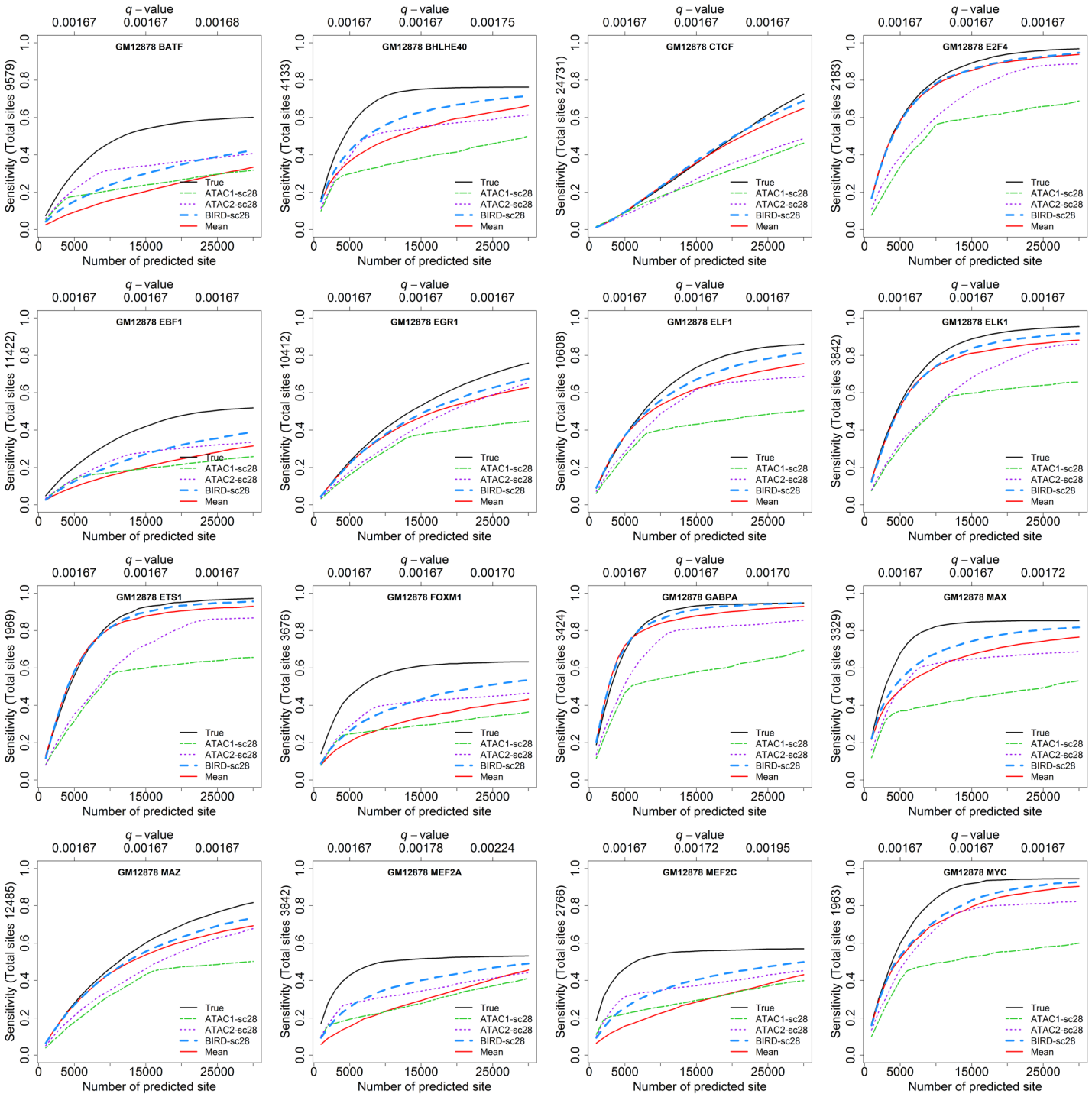
Supplementary Figure 11. Performance for predicting DH using pooled single-cell RNA-seq in GM12878 evaluated using ATAC-seq with 50,000 cells ("ATAC-b50k") or 500 cells ("ATAC-b500") as gold standard. (**a**) Cross-locus correlation between ATAC-seq with 50,000 cells and chromatin accessibility predicted or measured by different single-cell methods. The correlation is shown as a function of pooled cell number. Error bars are standard deviation based on 10 independent samplings of cells (**Methods**). "ATAC1": scATAC-seq dataset 1. "ATAC2": scATAC-seq dataset 2. "BIRD": BIRD-predicted DH using pooled single-cell RNA-seq. "BIRD-hybrid": the average of BIRD-predictions based on 28 cells and pooled ATAC-seq from scATAC-seq dataset 2 (here x-axis is the total number of cells used by scRNA-seq and scATAC-seq). Prediction performance using the mean DH profile of training samples ("Mean") is shown as a dashed line. (**b**) Scatterplots comparing ATAC-seq with 50,000 cells with signals from ATAC1 and ATAC2 using 10 cells ("ATAC1-sc10", "ATAC2-sc10"), 28 cells ("ATAC1-sc28", "ATAC2-sc28") and 222 cells ("ATAC1-sc222", "ATAC2-sc222"), BIRD prediction using 10 cells ("BIRD-sc10"), 28 cells ("BIRD-sc28") and BIRD-hybrid using 222 cells ("BIRD-hybrid-sc222"). (**c**)-(**d**) The same as (**a**)-(**b**) but using ATAC-seq with 500 cells as gold standard.



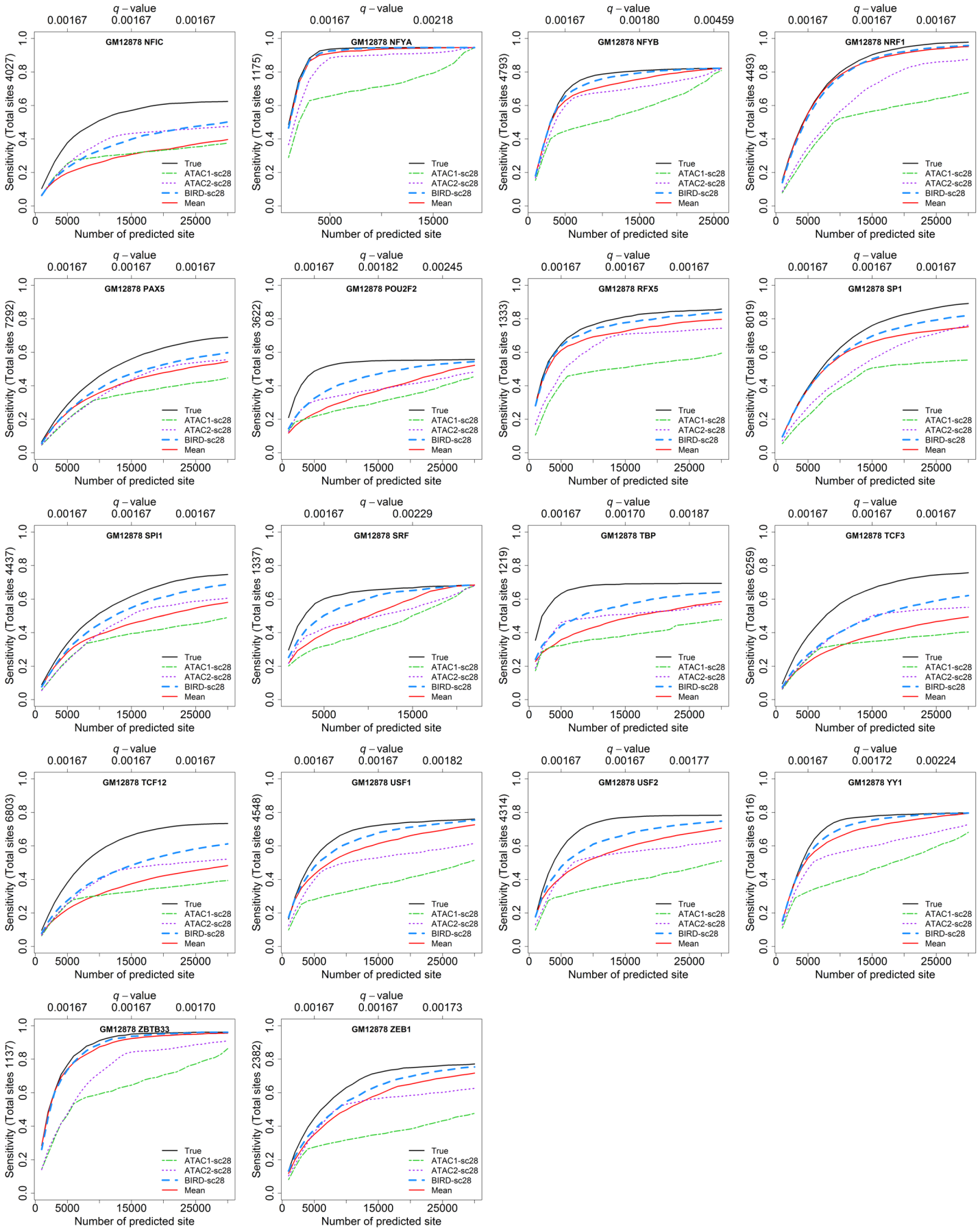
Supplementary Figure 12. Sensitivity-rank curves for predicting transcription factor binding sites in GM12878 by pooling single-cell RNA-seq from 10 cells. Predictions were made using true DNase-seq data ("True"), pooled single-cell ATAC-seq dataset1 (ATAC1) and dataset2 (ATAC2) with 10 cells ("ATAC1-sc10", "ATAC2-sc10"), BIRD-predicted DH using pooled single-cell RNA-seq with 10 cells ("BIRD-sc10"), and mean DH profile ("Mean"). For each method, the curve shows the percentage of true TFBSs discovered by top predicted motif sites. *q*-values corresponding to top 5000, 15000, and 25000 BIRD-sc10 predictions are shown on top of each plot.



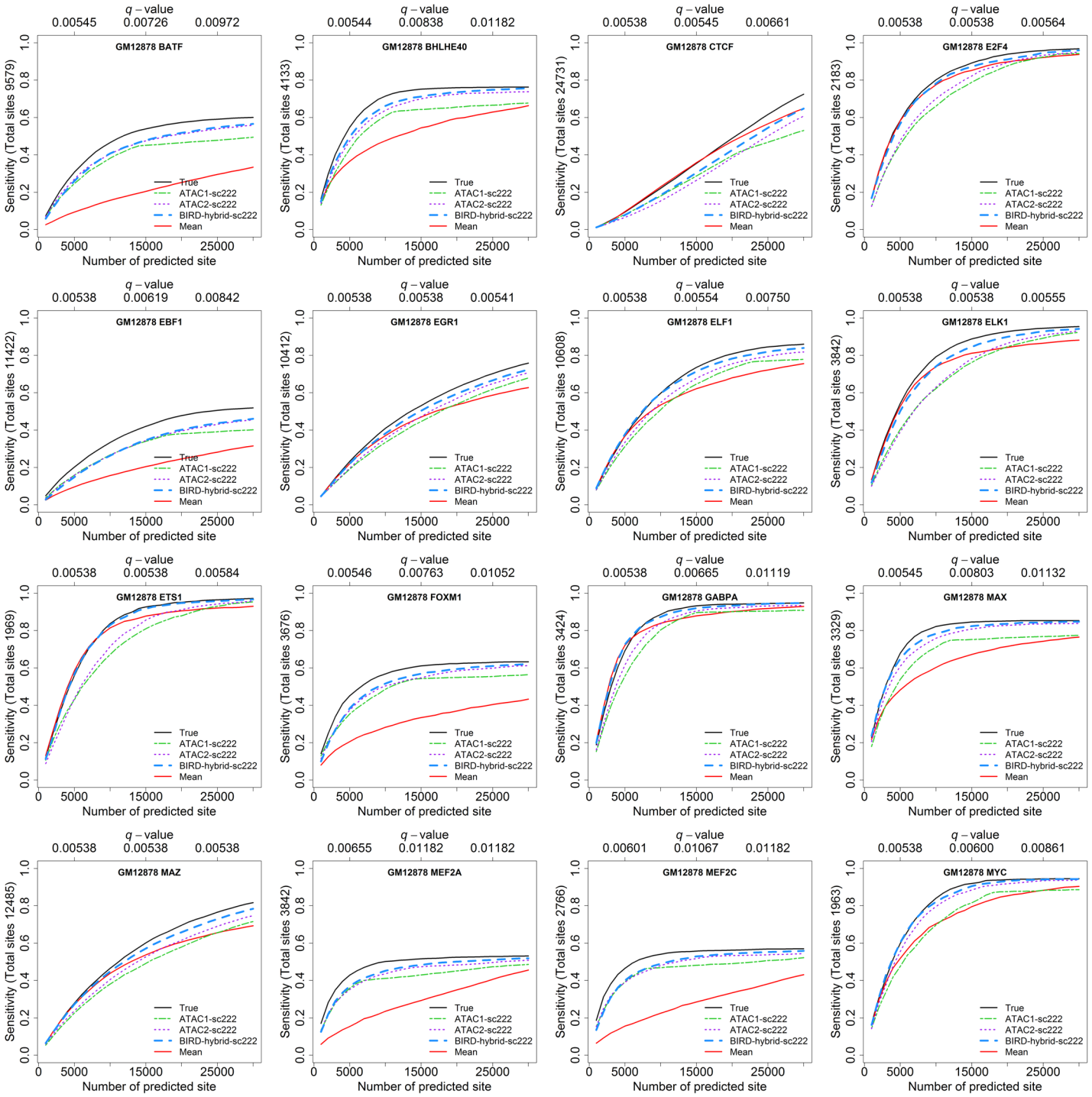
Supplementary Figure 12 (cont'd).



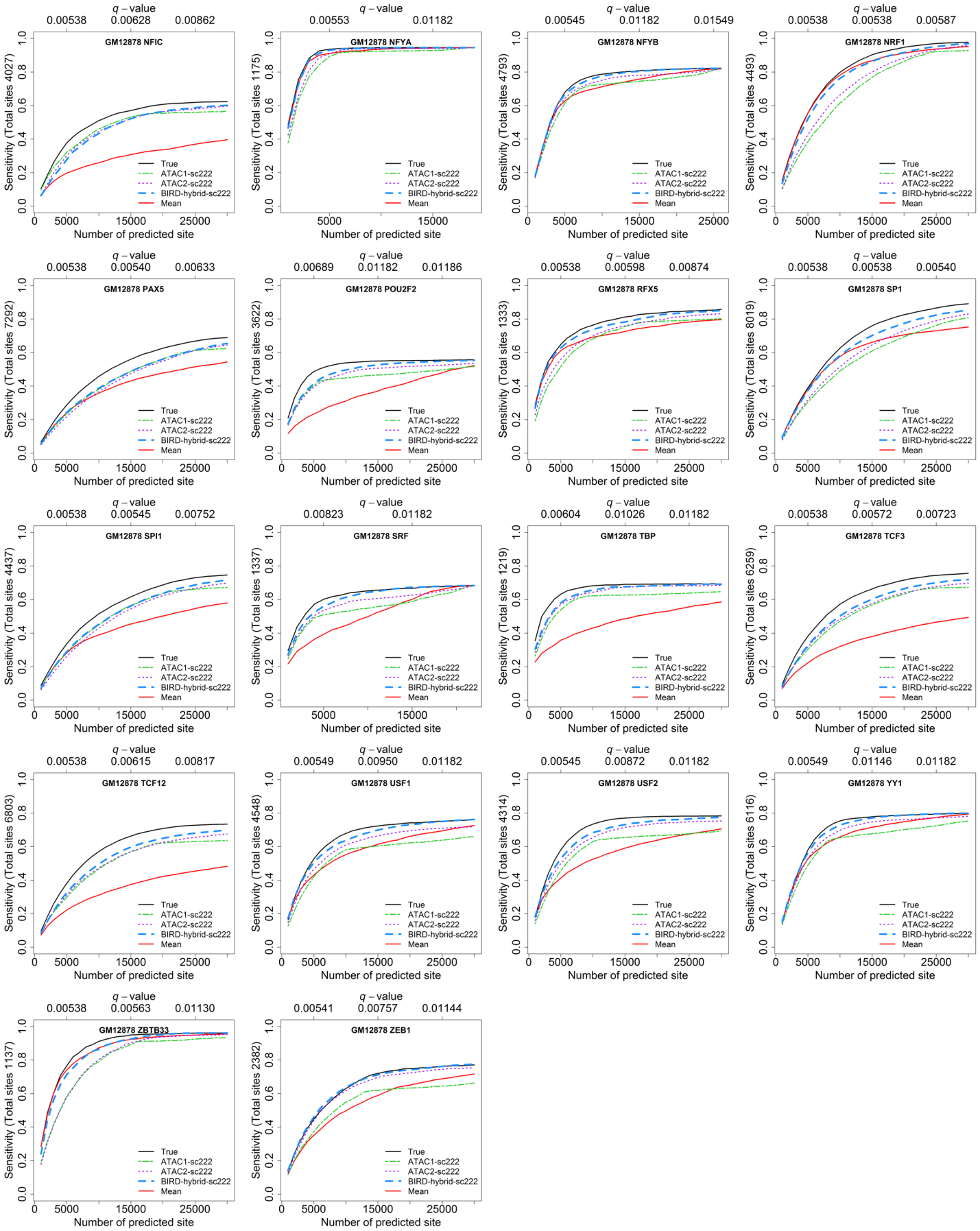
Supplementary Figure 13. Sensitivity-rank curves for predicting transcription factor binding sites in GM12878 by pooling single-cell RNA-seq from 28 cells. Predictions were made using true DNase-seq data ("True"), pooled single-cell ATAC-seq dataset1 (ATAC1) and dataset2 (ATAC2) with 28 cells ("ATAC1-sc28", "ATAC2-sc28"), BIRD-predicted DH using pooled single-cell RNA-seq with 28 cells ("BIRD-sc28"), and mean DH profile ("Mean"). For each method, the curve shows the percentage of true TFBSs discovered by top predicted motif sites. *q*-values corresponding to top 5000, 15000, and 25000 BIRD-sc28 predictions are shown on top of each plot.



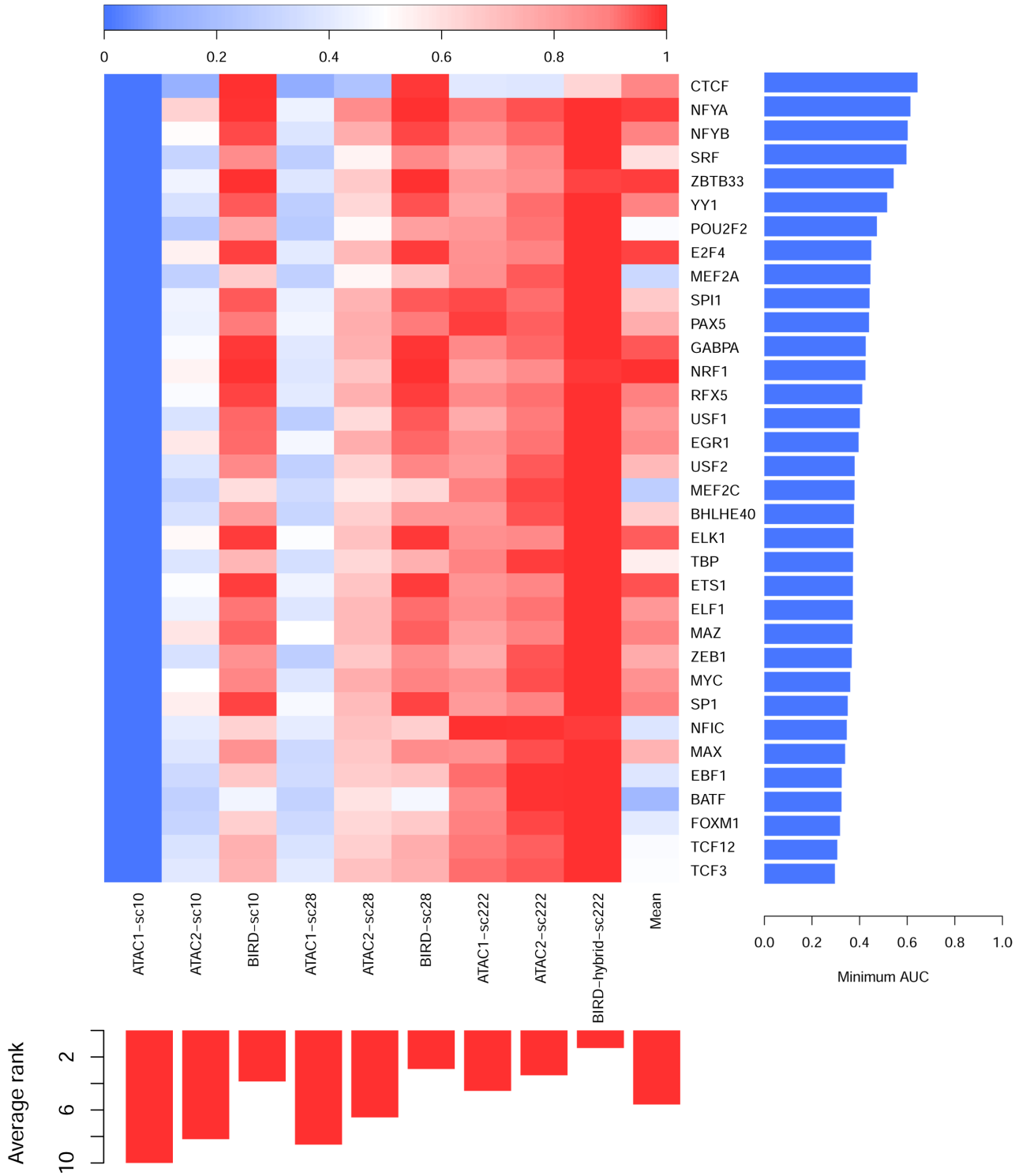
Supplementary Figure 13 (cont'd).



Supplementary Figure 14. Sensitivity-rank curves for predicting transcription factor binding sites in GM12878 by combining BIRD-predicted DH with scATAC-seq using a total of 222 cells. Predictions were made using true DNase-seq data ("True"), pooled single-cell ATAC-seq dataset1 (ATAC1) and dataset2 (ATAC2) with 222 cells ("ATAC1-sc222", "ATAC2-sc222"), the average of BIRD-predicted DH using pooled single-cell RNA-seq from 28 cells and pooled scATAC-seq from 194 cells ("BIRD-hybrid-sc222"), and mean DH profile ("Mean"). For each method, the curve shows the percentage of true TFBSs discovered by top predicted motif sites. *q*-values corresponding to top 5000, 15000, and 25000 BIRD-hybrid-sc222 predictions are shown on top of each plot.



Supplementary Figure 14 (cont'd).



Supplementary Figure 15. Scaled area under the curve (AUC) for different methods that predict TFBSs using pooled single-cell data. Each row is a TF, and each column is a method. For each TF, different methods are ranked based on the AUC value, and the worst AUC value of all methods is shown on the right using a blue bar. The average rank of each method across all TFs is shown on the bottom using a red bar. Smaller rank means better performance. “ATAC1-sc10”, “ATAC2-sc10”, “ATAC1-sc28”, “ATAC2-sc28”, “ATAC1-sc222”, “ATAC2-sc222”: pooled single-cell ATAC-seq from 10, 28 or 222 cells using scATAC-seq dataset 1 (ATAC1) or dataset 2 (ATAC2). “BIRD-sc10”, “BIRD-sc28”: BIRD-predicted DH using pooled single-cell RNA-seq from 10 or 28 cells. “BIRD-hybrid-sc222”: the average of BIRD-predicted DH using pooled single-cell RNA-seq from 28 cells and pooled scATAC-seq from 194 cells. “Mean”: mean DH profile of training samples.