Supplementary Data

Supplementary Data include Supplementary Methods, Supplementary References, 5 Supplementary Figures, and 2 Supplementary Tables.

Supplementary Methods

Detailed Computing Procedure

BART predicts functional transcription factors (TFs) that bind at genomic cis-regulatory regions to regulate gene expression, given a query gene set or a ChIP-seq dataset. BART leverages 3485 publicly available TF ChIP-seq datasets in human and 3055 in mouse. Detailed procedure is as follows:

1. *Generate cis-regulatory profile.* If the input is a gene set ("geneset" mode), MARGE (MARGE-express v1.0, with default parameters) (S. Wang *et al.*, 2016) is used to generate a cis-regulatory profile, in which a score is given to each UDHS. The higher the score is, the more likely that this DHS is bound by transcription factors that regulate the input gene set. If the input is a ChIP-seq dataset ("profile" mode), a RPKM value is calculated on a 1kb-wide region surrounding each UDHS (500bp upstream and downstream from UDHS center) to generate the cis-regulatory profile. All UDHSs are decreasingly ranked by MARGE scores or RPKM values for the next step.

2. *Find transcription factors whose binding profile best match the cis-regulatory profile*. A TF binding profile is a binary list indicates whether this TF binds on (value 1) each UDHS or not (value 0). A receiver-operating characteristic (ROC) curve is generated to evaluate the performance of predicting the TF binding profile from the ranked cis-regulatory profile, using highly-ranked UDHS in the cis-regulatory profile as positive prediction and bound/unbound TF as true or false. The area under the ROC curve (AUC) value is calculated for each TF binding profile. Among the over 3,000 TF profiles, those with higher AUC values are better associated with the cis-regulatory profile.

3. *Statistical tests for each TF on AUC values*. Most TFs have multiple ChIP-seq datasets in the data compendium of over 3,000 ChIP-seq datasets. Although datasets for the same TF may come from different cell types or experimental conditions with different qualities, their collected binding patterns have better statistical power of prediction than individual datasets. For each TF, the Wilcoxon rank sum test (Wilcoxon, 1945), a non-parametric statistical test, is conducted to compare the AUC values of this TF with AUC values of all TF datasets, and a Wilcoxon statistic score and a p-value are calculated and assigned to this TF.

4. *Standardization of statistic scores using public data compendium*. Analysis of Wilcoxon statistic scores of all TFs on gene sets over a broad range of biological functions shows that different TFs usually have different genome-wide binding distribution patterns. Some TFs tend to have higher AUC scores and some have lower. In order to correct this bias and to standardize the significance assessment, 505 gene sets from Molecular Signature Database (MSigDB) (Liberzon *et al.*, 2011), each of which has at least 200 genes, were collected as human gene set compendium, while

366/267 H3K27ac profiles were collected for human or mouse data profile compendium, respectively. BART analysis was performed on the data compendium and a Wilcoxon statistic score was generated for each TF on each gene set and each H3K27ac profile. For each TF i, we calculate a standard score

$$Z_i = (X_i - \mu_i)/\sigma_i,$$

where X_i is the Wilcoxon statistic score of TF i calculated from the query data, μ_i and σ_i are mean and standard deviation of Wilcoxon statistic scores of TF i across all datasets from the compendium.

5. *Rank summarization of all TFs*. The average rank of standard Z-score (decreasing), Wilcoxon *P*-value (increasing), and maximum AUC value (decreasing) is calculated as the summary rank in the TF prediction output. For each TF, the maximum AUC value is obtained from all AUC values of this TF's ChIP-seq datasets. In the final output result, all TFs are ranked by this average rank score and higher ranked TFs are predicted as higher possible functional TFs.

Union DNaseI Hypersensitive Sites

We use a collection of union DNaseI hypersensitive sites (UDHS) as all candidate cis-regulatory regions in the genome. UDHS were derived from 468 human and 116 mouse DNase-seq datasets, which include 2,723,010 unique non-overlapping DNase-seq peaks in human (hg38) and 1,529,448 in mouse (mm10), each DHS with a minimum length of 50 bp.

Public Transcription Factor Binding Profiles

We collected 3485 public transcription factor ChIP-seq datasets in human and 3055 in mouse from various cell types and tissues from Cistrome Data Browser (Mei *et al.*, 2017). For each dataset, TF binding sites (ChIP-seq peaks) were identified using MACS2 with default parameters and peaks with fold enrichment of at least 5 were retained. For each TF ChIP-seq dataset, a TF binding binary profile is generated on the UDHS. A UDHS is assigned as bound (score 1) if the UDHS region is overlapped with a TF ChIP-seq narrow peak, otherwise assigned as unbound (score 0).

Public Data Compendium for Standardization

Two datasets were collected as compendium for the analysis of human datasets. For gene set input, 505 gene sets with at least 200 genes were selected from chemical and genetic perturbation sets of molecular signature database (MSigDB): http://software.broadinstitute.org/gsea/msigdb. For ChIP-seq dataset input, 366 H3K27ac profiles from various cell lines were collected. For the analysis of mouse datasets, 267 H3K27ac profiles were used for both geneset and ChIP-seq input. Wilcoxon statistic scores of all TFs across compendium were calculated and used for the standardization of statistic score for each TF for each input.

Test Gene Sets

Six gene sets were used as examples to test the prediction performance of BART. These gene sets were derived from differentially expression analysis and can be treated as target genes of known transcription factors. These gene sets include: up-regulated genes upon activation of the estrogen receptor ESR1 in breast cancer cell line MCF7 (Carroll *et al.*, 2006); up-regulated genes upon activation of the androgen receptor AR in prostate cancer cell line LNCaP (Q. Wang *et al.*, 2007); up-regulated genes upon activation of the glucocorticoid receptor NR3C1 in lung cancer cell line A549 (Muzikar *et al.*, 2009); induced genes during adipogenesis regulated by the peroxisome proliferator-activated receptor gamma (PPARG) (Mikkelsen *et al.*, 2010); gamma secretase inhibitor (GSI) sensitive genes as NOTCH1 targets in leukemia cell line CUTLL1 (H. Wang *et al.*, 2014); and down-regulated genes upon shRNA knock-down of POU5F1 in human embryonic stem cell line H1 (Kunarso *et al.*, 2010).

Supplementary References

Carroll,J.S. *et al.* (2006) Genome-wide analysis of estrogen receptor binding sites. *Nat Genet*, **38**, 1289–1297.

Kunarso,G. *et al.* (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*, **42**, 631–634.

Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Mei,S. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research*, **45**, D658–D662.

Mikkelsen,T.S. *et al.* (2010) Comparative epigenomic analysis of murine and human adipogenesis. *Cell*, **143**, 156–169.

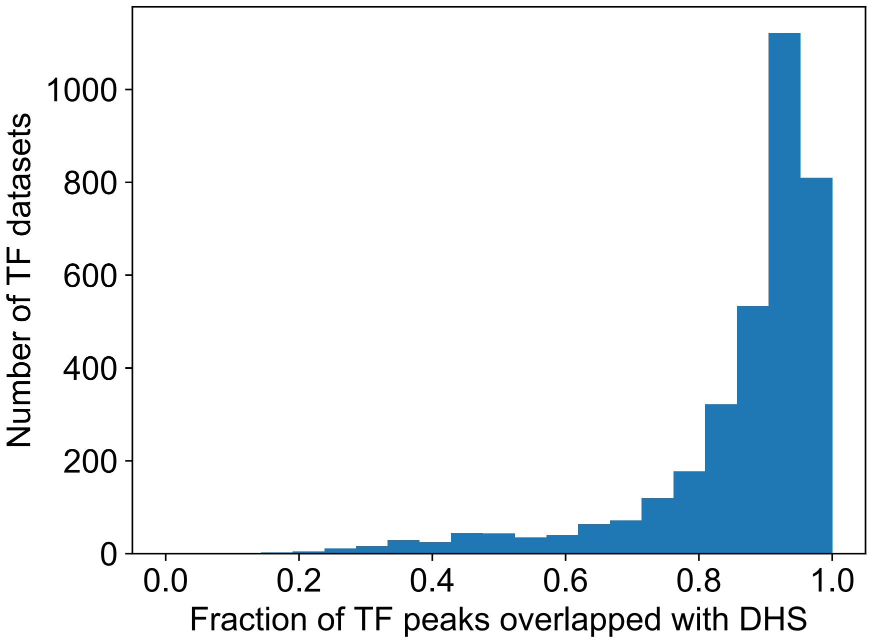
Muzikar,K.A. *et al.* (2009) Repression of DNA-binding dependent glucocorticoid receptormediated gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 16598–16603.

Wang,H. *et al.* (2014) NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 705–710.

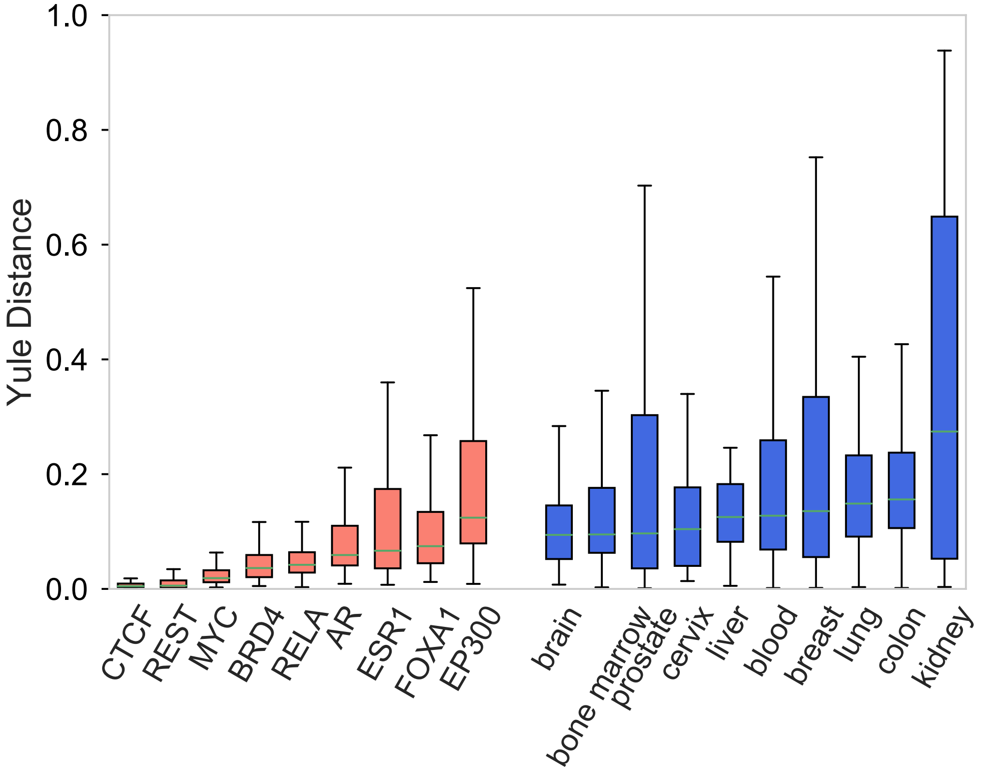
Wang,Q. *et al.* (2007) A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. *Molecular Cell*, **27**, 380–392.

Wang,S. *et al.* (2016) Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. *Genome Research*, **26**, 1417–1429.

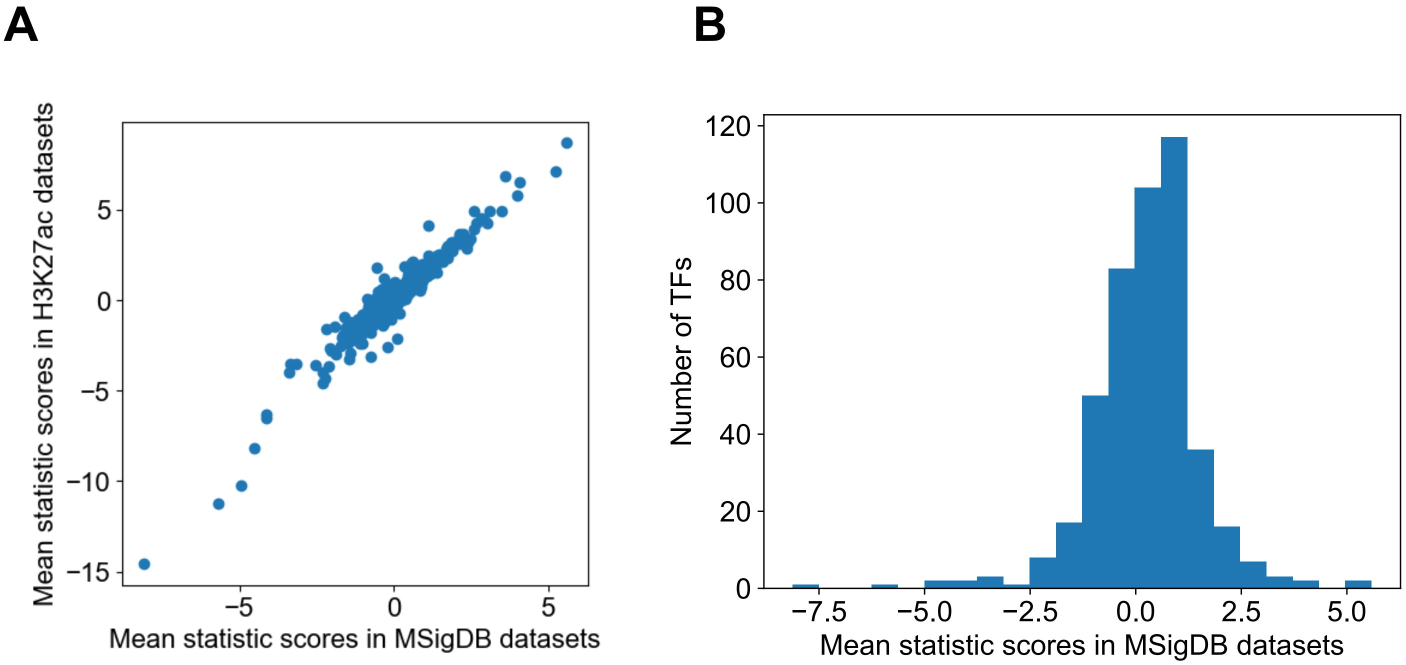
Wilcoxon,F. (1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, **1**, 80–83.



Supplementary Figure S1. Distribution of the fraction of peaks overlapping with UDHS for each human ChIP-seq dataset. For most transcription factors (81%), more than 80% of peaks in their ChIP-seq profiles are overlapped with UDHS.



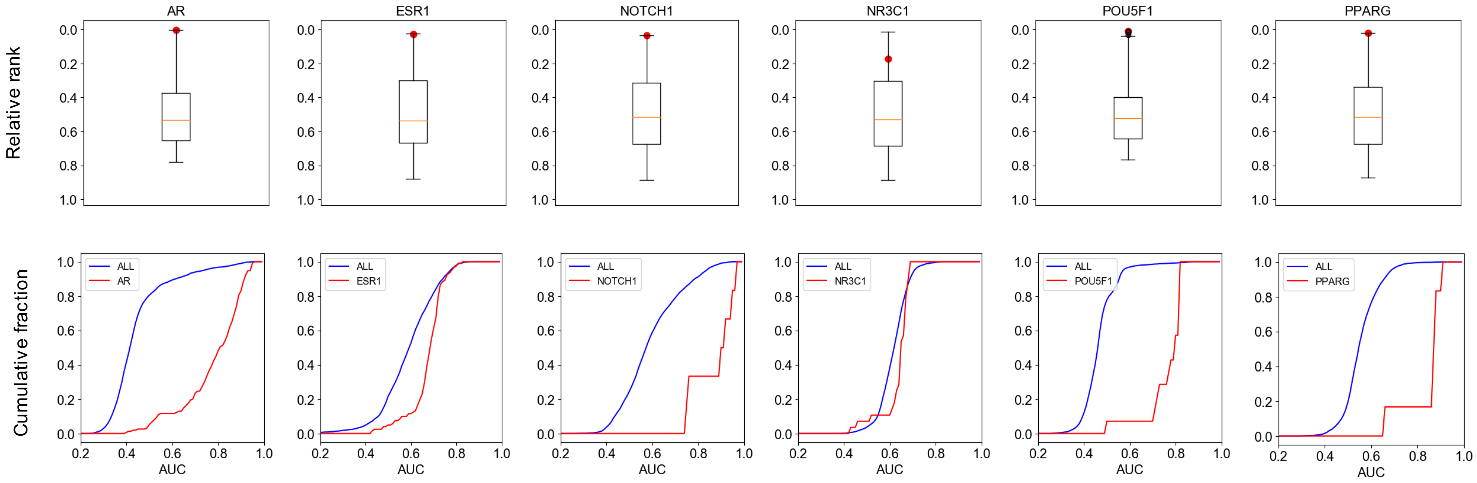
Supplementary Figure S2. Genome-wide transcription factor binding specificity compared with tissue/cell-type specificity. For each transcription factor with more than 50 ChIP-seq datasets among the 3485 collected human datasets, the Yule distance between each pair of datasets from different tissue types was calculated (orange). For each tissue type with more than ChIP-seq 50 datasets, the Yule distance between each pair of datasets of different factors was calculated (blue). Higher Yule distance indicates less similarity. P=0.0004, by one-side Mann Whitney U test.



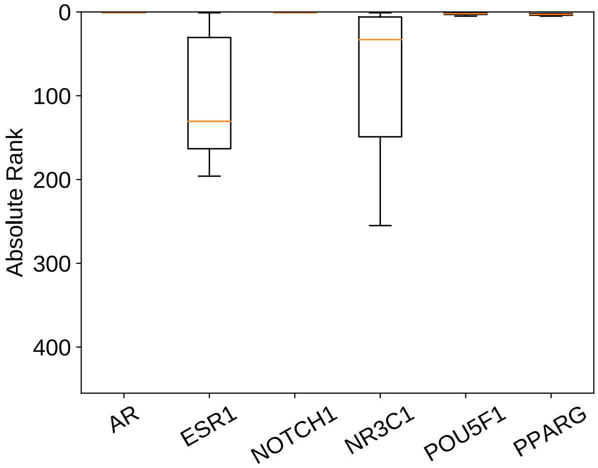
Supplementary Figure S3. Global distribution of transcription factors on UDHS.

(**A**) X-axis represents the mean value of Wilcoxon statistic scores in all 505 MSigDB datasets for each of the 454 TFs in human, and y-axis the mean value of Wilcoxon statistic scores in H3K27ac datasets. Spearman rank correlation of these two sets of data is 0.95.

(**B**) Distribution of mean Wilcoxon statistic score of each TF across all 505 MSigDB gene sets.



Supplementary Figure S4. BART prediction of functional TFs on six differentially expressed gene sets, which were collected from perturbation of AR, ESR1, NOTCH1, NR3C1, PPARG and deletion treatment of POU5F1. The red dot in the box plot represents the relative rank of target TF in the final results, and the cumulative distribution plot compares the distribution of AUC scores from target TF with AUC scores of all datasets. BART accurately predicted most target TF as most functional associated TF in these 6 gene sets.



Supplementary Figure S5. Distribution of the Rank of the true transcription factor in the BART prediction results for the 6 test gene sets, using each of the 30 MARGE outputs. AR, NOTCH1, POU5F1, and PPARG are robustly ranked on top, indicating the BART prediction is robust to the randomness of MARGE results.

Tool name	Search space from gene TSS	Promoter/enhancer	Sequence motif	ChIP-seq data	Reference
ENCODE ChIP-Seq Significance Tool	[- 5kb, TTS + 5kb]	Promoter + proximal enhancer		√	Auerbach et al., 2013
HOMER	[- 2kb, + 2kb]	Promoter	√		Heinz et al., 2010
iRegulon	[- 10kb, + 10kb]	Promoter + proximal enhancer	√	√	Janky et al., 2014
Pscan	[- 1kb, 0]	Promoter	√		Zambelli et al., 2009
BART	[- 100kb, + 100kb], genome-wide	Promoter + enhancer		√	

Supplementary Table 1: Methodology comparison between 4 existing tools and BART.

	AR	ESR1	NR3C1	NOTCH1	POU5F1	PPARG
ENCODE ChIP-Seq Significance Tool	NA	NA	1/61	NA	NA	NA
HOMER	4/364	221/364	2/364	NA	14/364	28/364
iRegulon	43/62	NA	1/45	NA	1/65	NA
Pscan	97/579	490/579	153/579	NA	244/579	29/579
BART	1/454	2/454	47/454	1/454	1/454	1/454

Supplementary Table 2: Performance comparison between 4 existing tools and BART. For each case study, the same gene set is used as query. For each prediction result, the rank of the true transcription factor is shown as the numerator, and the total number of transcription factors included in each prediction result is shown as the denominator. NA, the tool is not able to identify the true factor.