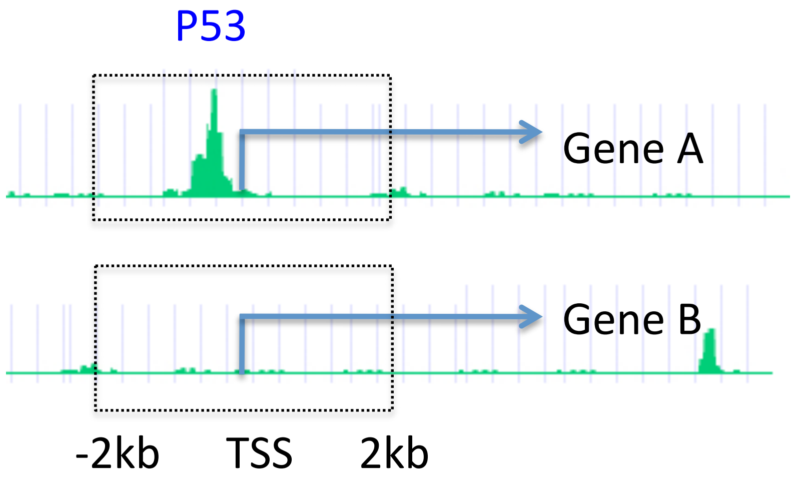
TIP: Transcription Factor Target Identification with a Probabilistic Model

Chao Cheng

The Gerstein Lab

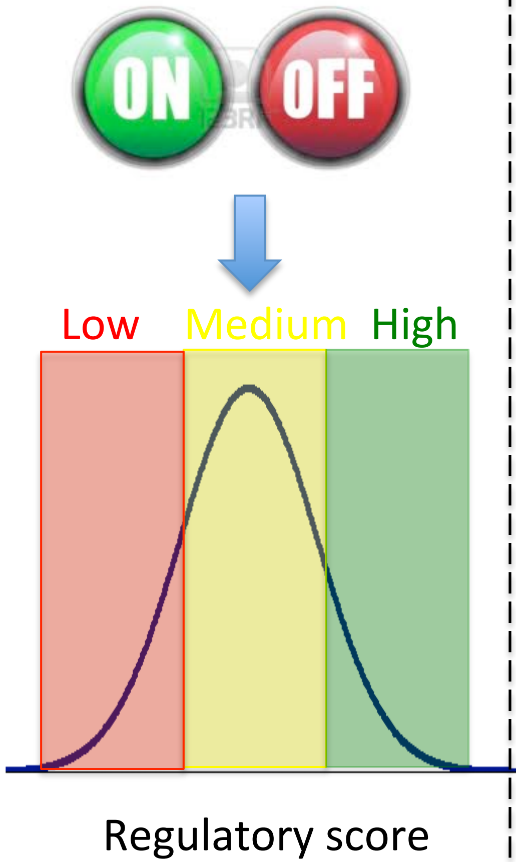
Conventional way to identify TF target genes



P53 → A

P53 X B

TF binding and TF→gene regulation is NOT binary but quantitative



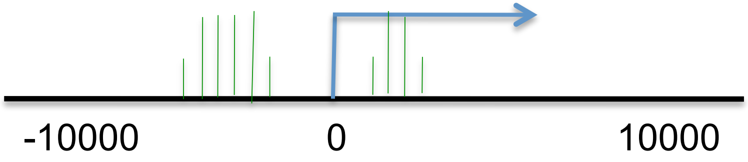
Limitations of peak-based method:

-- sensitive to # binding peaks

-- sensitive to cut-off value (1, 2 or 5 kb of TSS?)

-- no significance estimation

Target Identification with a Probabilistic model (TIP)

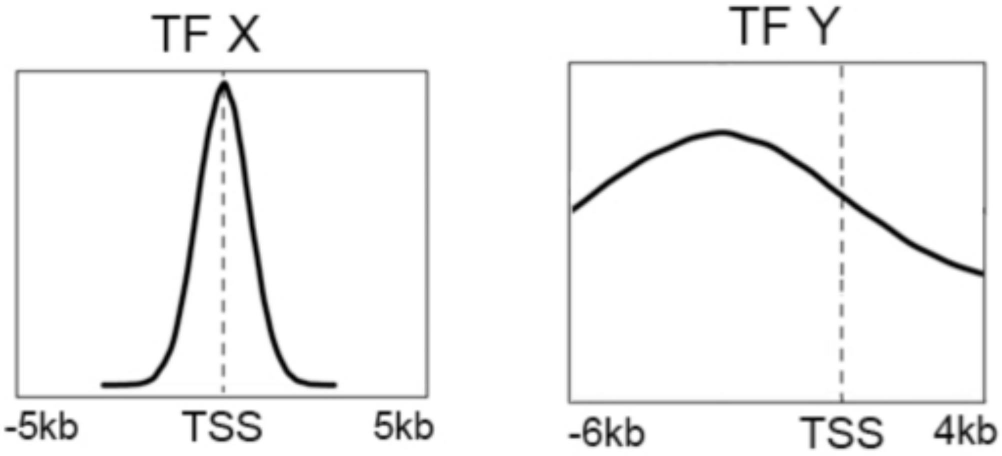


$S_i(g)$: the signal of a TF at nucleotide i of gene g

-- TF binding signal at different position contribute differently

(1) Distance from TSS – binding signal closer to TSS contribute more

(2) Binding preference of a TF – TF specific



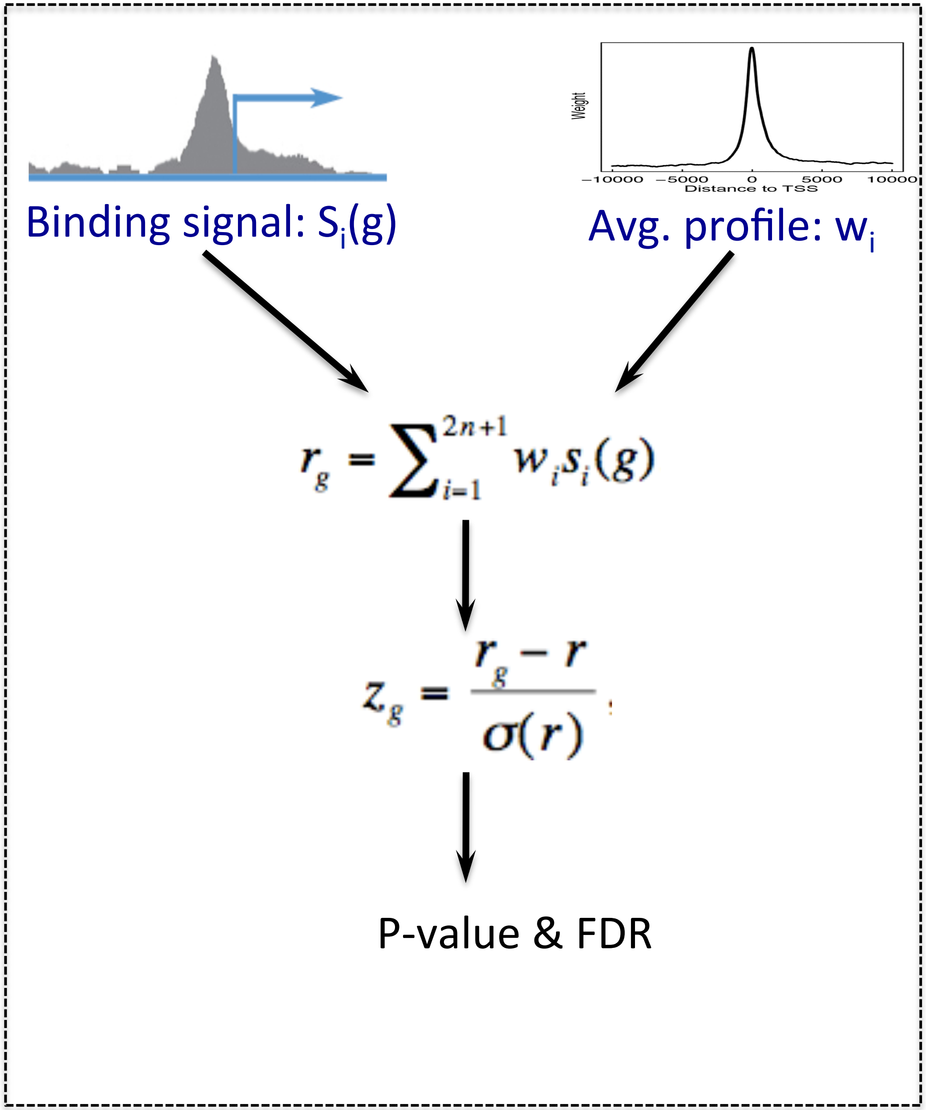
Avg. binding signal across all genes
-- characteristic binding profile of a TF

Three Steps of TIP

Step 1: Given a ChIP-seq wiggle file, estimate the characteristic binding profile of the TF – the aggregation signal in the 20kb DNA region centering at TSS

Step 2: For each gene, calculate the weighted sum of binding signal of the TF in its 20kb TSS region. – denote it as regulatory score.

Step 3: Normalize the regulatory scores of genes into z-scores, can estimate the P-value and FDR. – result in a ranked target gene list.



Binding signal: $S_i(g)$

Avg. profile: w_i

$$r_g = \sum_{i=1}^{2n+1} w_i s_i(g)$$

$$z_g = \frac{r_g - r}{\sigma(r)}$$

P-value & FDR

Cheng et al. 2011, Bioinformatics

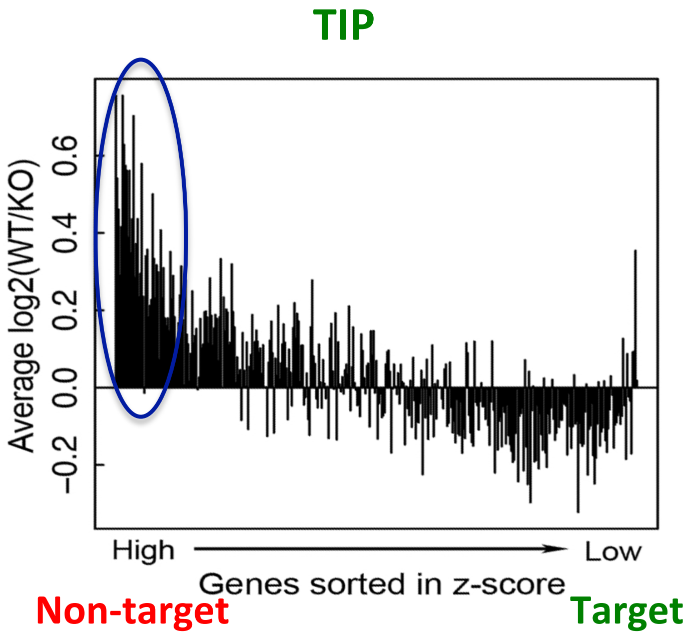
Example 1: STAT4 targets identified by TIP are more down-regulated in STAT4 KO

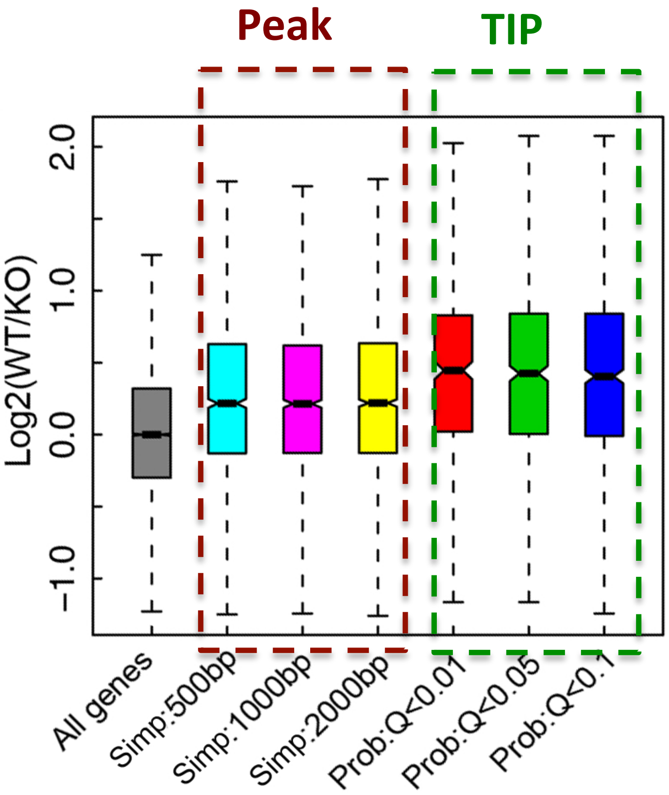
Data

ChIP-seq data:
Stat4 binding in Th1 cell

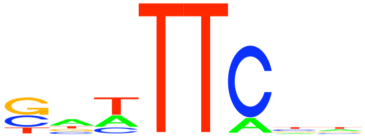
Gene expression:
STAT4KO + WT mice

Wei et al., 2010

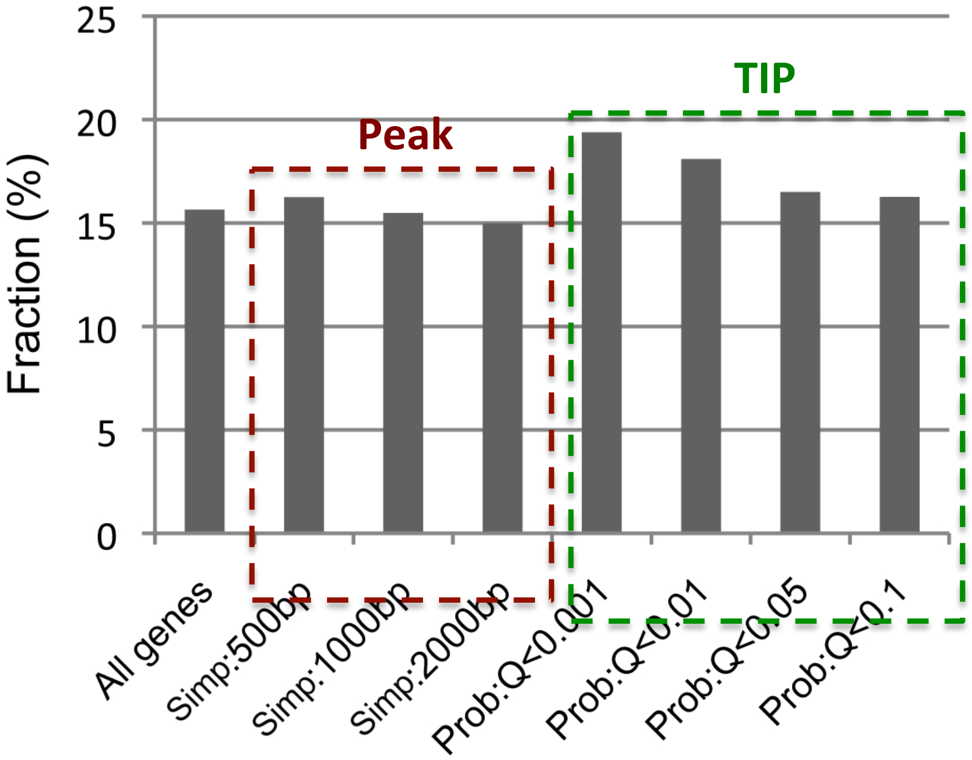




Example 1: STAT4 binding motif are more enriched in targets identified by TIP



STAT4 motif



Example 2: ERα targets identified by TIP are more responsive to estrogen treatment

Data

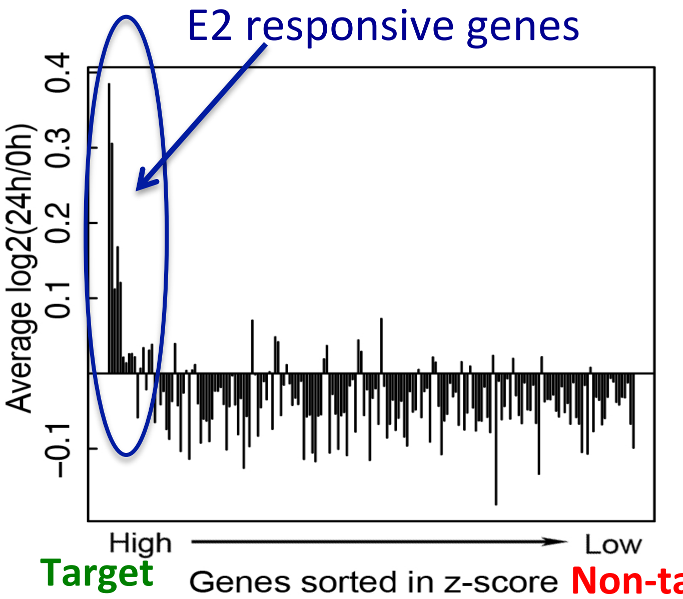
ChIP-seq data:

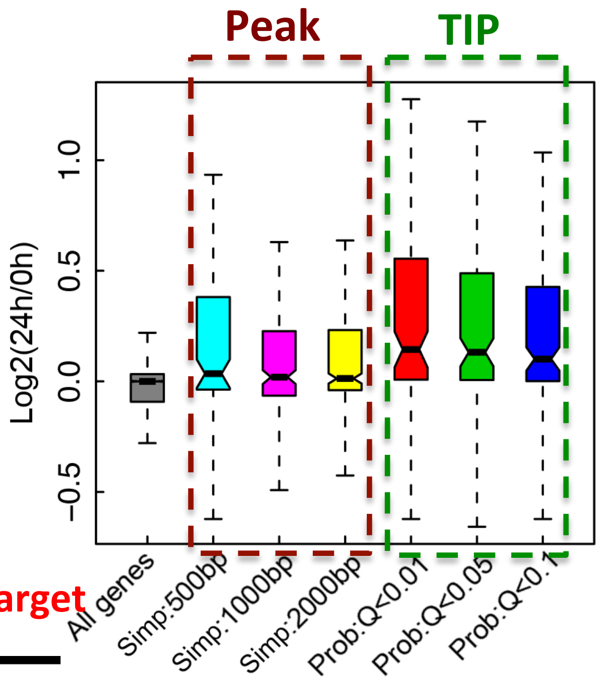
ERα binding in MDA-MB-231 cell line

Gene expression:

before and after E2 treatment

Stender et al., 2010



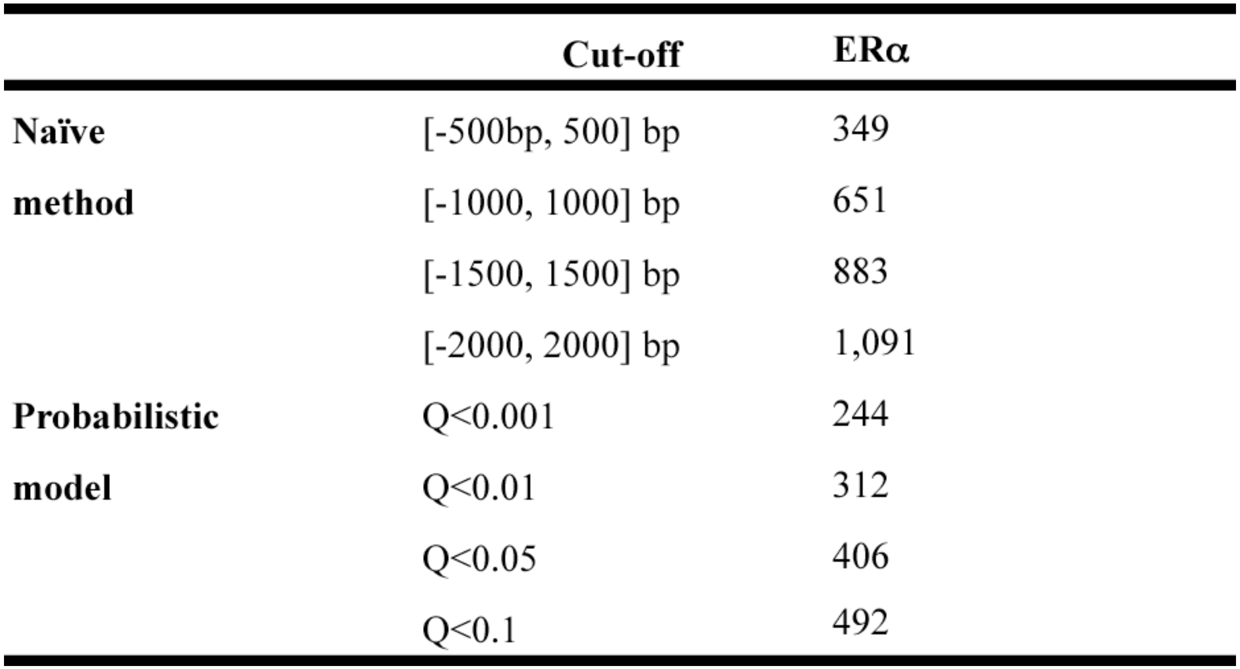


# target genes

	Cut-off	ERα
Peak-based	[-500bp, 500] bp	349
	[-1000, 1000] bp	651
	[-1500, 1500] bp	883
	[-2000, 2000] bp	1,091
TIP	Q<0.001	244
	Q<0.01	312
	Q<0.05	406
	Q<0.1	492

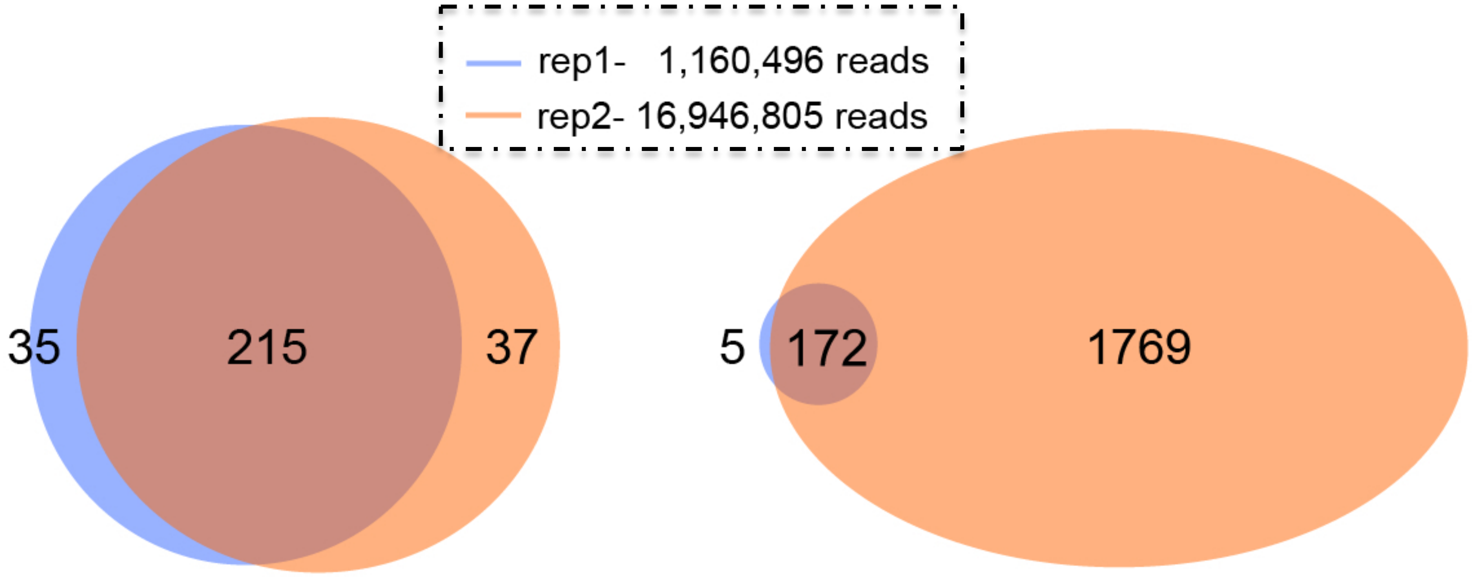
# TIP target

# target genes



	Cut-off	ERα
Naïve method	[-500bp, 500] bp	349
	[-1000, 1000] bp	651
	[-1500, 1500] bp	883
	[-2000, 2000] bp	1,091
Probabilistic model	Q<0.001	244
	Q<0.01	312
	Q<0.05	406
	Q<0.1	492

TIP is insensitive to sequencing depth



TIP (P<0.001)

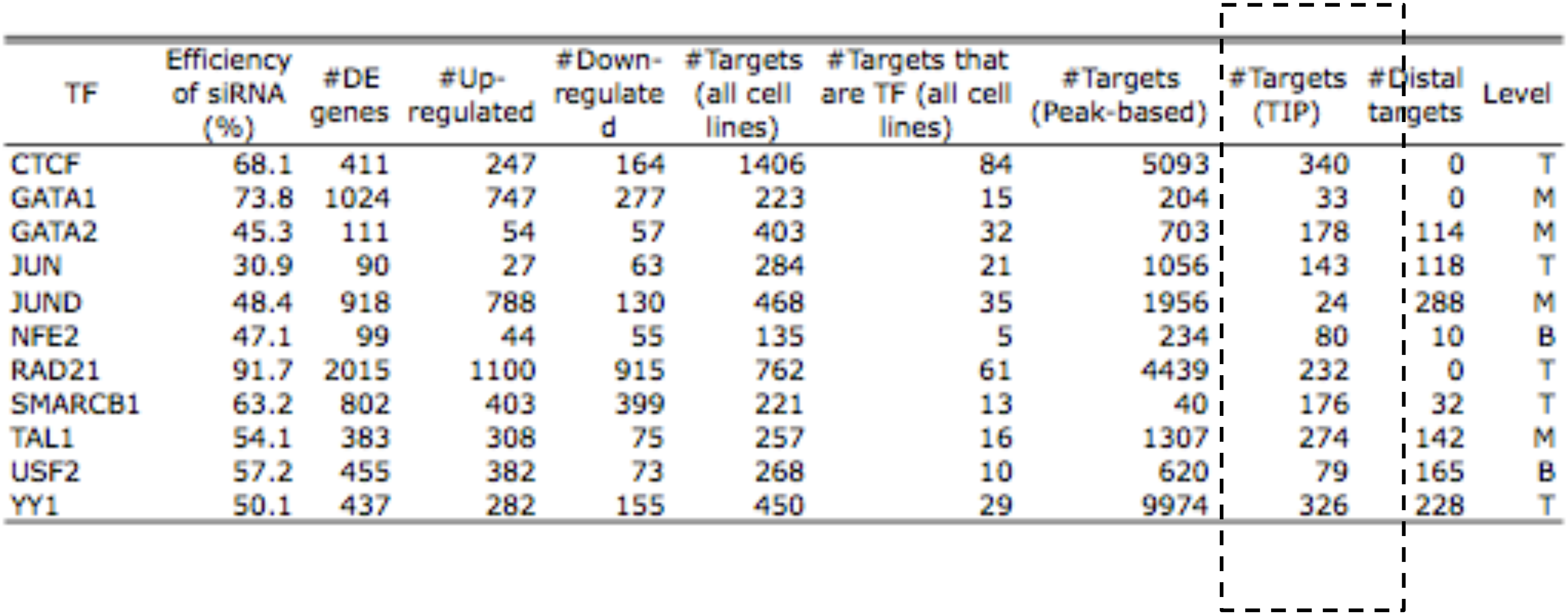
Peak based method

TCF4: *Mokry et al., 2010*

TIP is cost-effective-- do not require high read-depth for target gene identification!

Apply TIP to ENCODE data

Table: TFs with both ChIP-seq and knock-down expression data



TF	Efficiency of siRNA (%)	#DE genes	#Up-regulated	#Down-regulated	#Targets (all cell lines)	#Targets that are TF (all cell lines)	#Targets (Peak-based)	#Targets (TIP)	#Distal targets	Level
CTCF	68.1	411	247	164	1406	84	5093	340	0	T
GATA1	73.8	1024	747	277	223	15	204	33	0	M
GATA2	45.3	111	54	57	403	32	703	178	114	M
JUN	30.9	90	27	63	284	21	1056	143	118	T
JUND	48.4	918	788	130	468	35	1956	24	288	M
NFE2	47.1	99	44	55	135	5	234	80	10	B
RAD21	91.7	2015	1100	915	762	61	4439	232	0	T
SMARCB1	63.2	802	403	399	221	13	40	176	32	T
TAL1	54.1	383	308	75	257	16	1307	274	142	M
USF2	57.2	455	382	73	268	10	620	79	165	B
YY1	50.1	437	282	155	450	29	9974	326	228	T

Target expression in TF knockdown cells -- ENCODE data

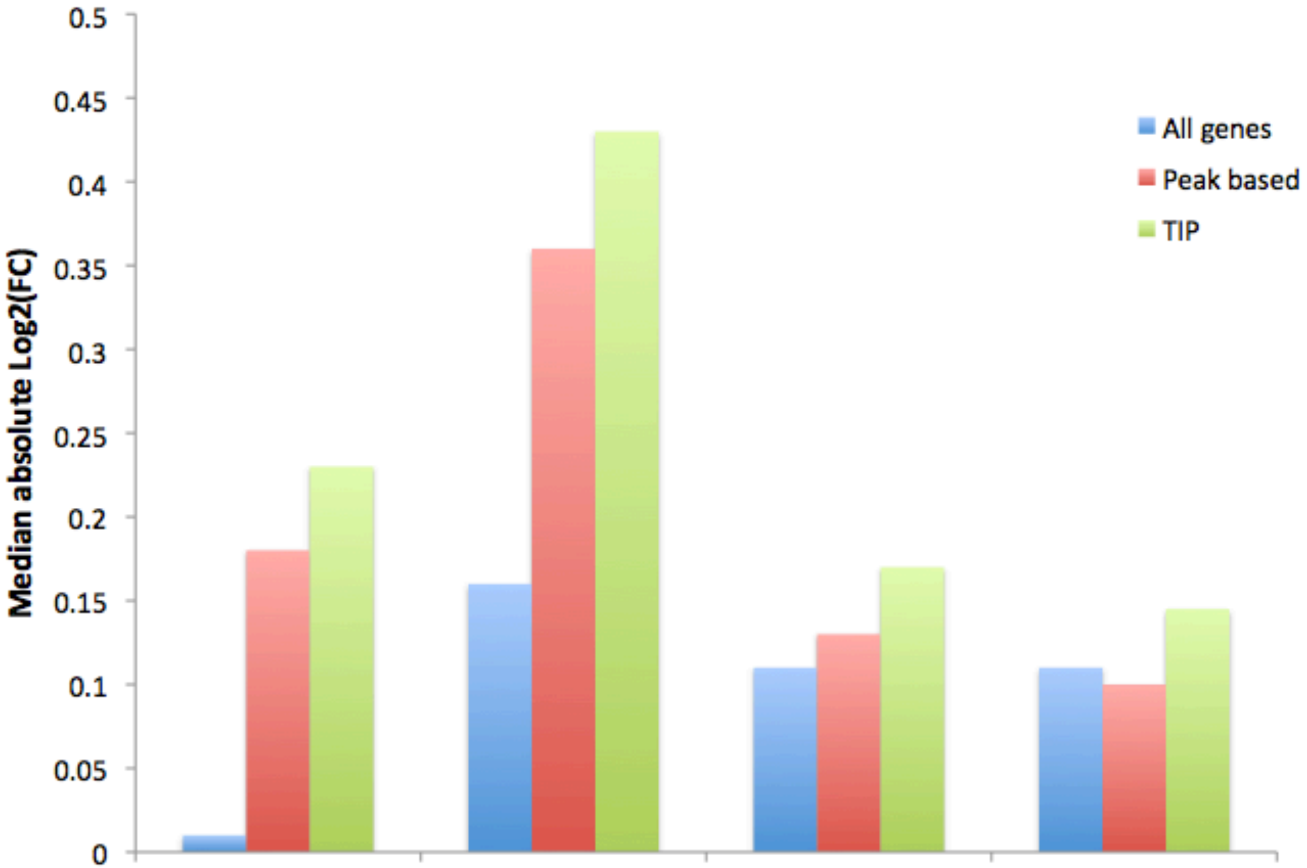
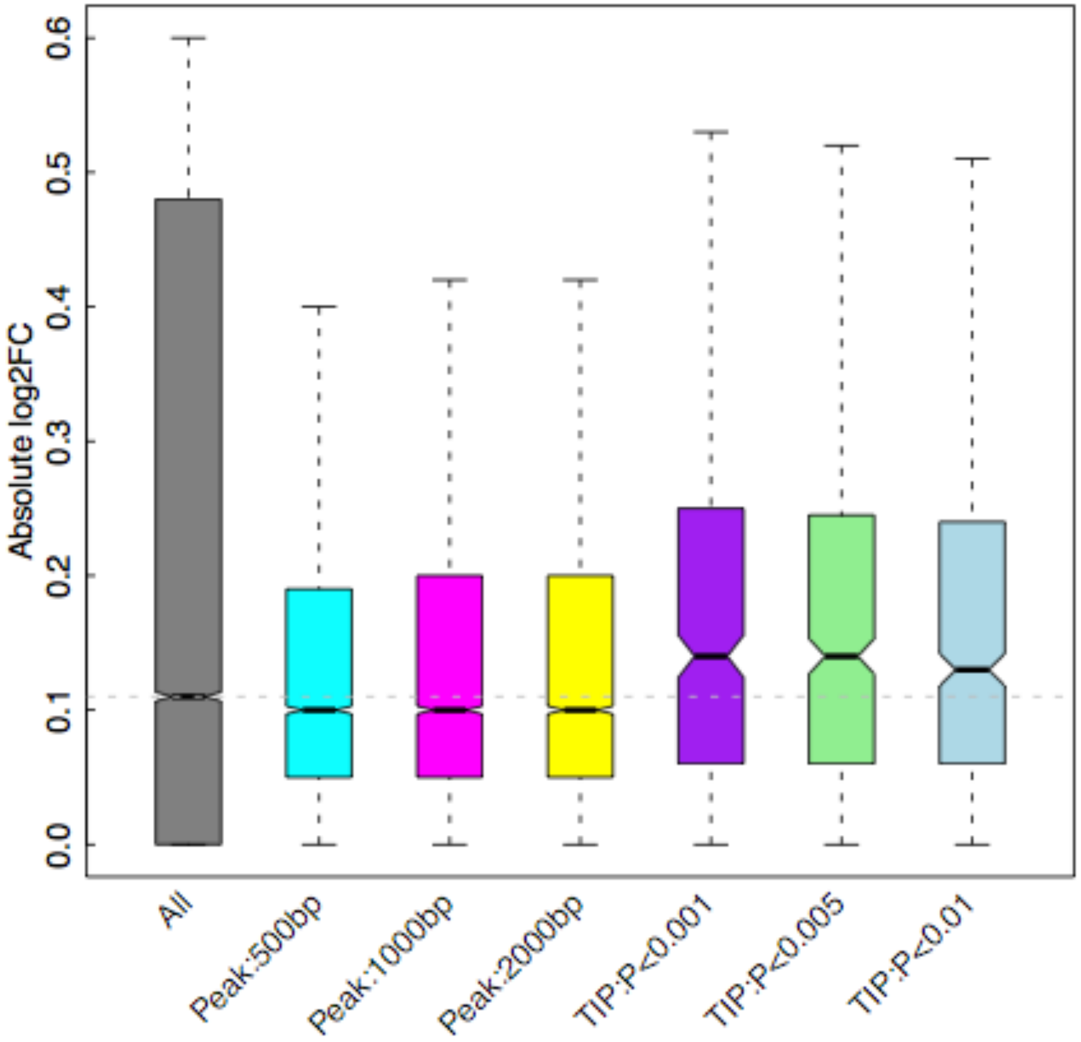


Figure S1D

Expression changes of GATA1 targets in its knockdown K562 cells



Extended version of TIP

(1) Combine TIP with peak based method

- **for each target gene, list all peaks nearby its TSS**
- **calculate the relative contribution of each peak to the regulatory score of a target gene**

(2) Improve P-value calculation.

- **current version is too conservative in estimate significance.**
- **using 2-component mixture model (non-target and target genes)**

(3) Build a webserver to implement TIP on line.