Integrating binding and expression data to predict transcription factors combined function

Mahmoud Ahmed

Gyeongsang National University

August 24, 2020
https://bit.ly/3fyxwZ9

Outline

Predicting the interaction of two factors using independent datasets of binding and expression

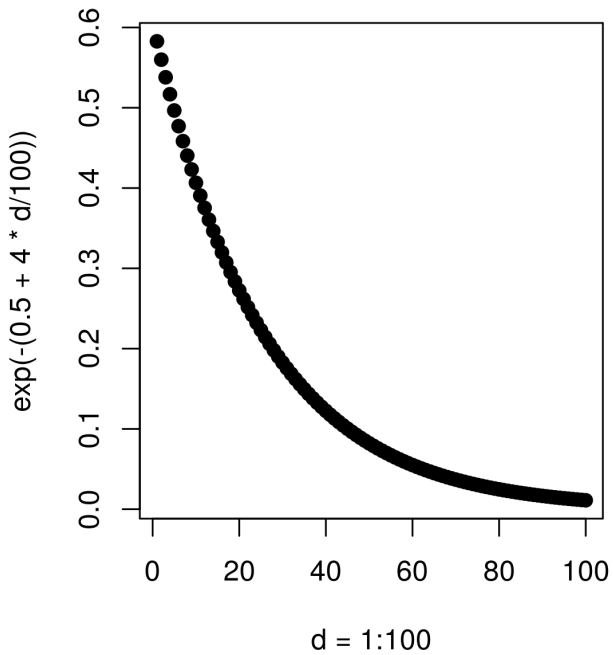
- The binding of a transcription factor to a genomic region (e.g. gene promoter) induces or represses its expression [Rougemont and Naef, 2012].
- Transcription factors share their binding sites with other factor, co-factors and/or DNA-binding proteins. The DNA-binding proteins may form complexes which bind to the DNA as one units.
- The integration of the overlapping binding sites and the effect of the gene expression of perturbed factors can be used to infer their combined function; cooperative or competitive.

Modeling the binding sites as the discounted distances of the ChIP peaks

Peak Score (S_p): is the distance (Δ) from transcription start site (TSS) relative to a 100 kb [Wang et al., 2013].

$$S_p = e^{-(0.5+4\Delta)}$$

- The shape of the function approximate empirical observations [Tang et al., 2011].

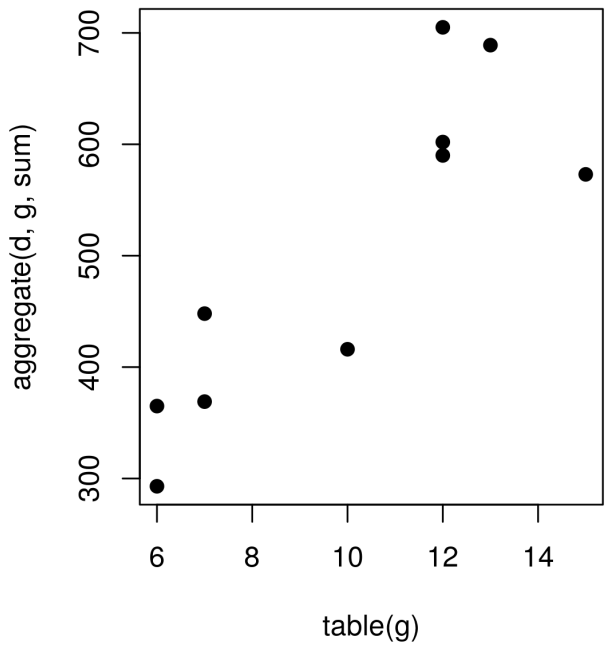


Modeling the regulatory potential as the sum of the weighted peaks

Gene Score (S_g): is the sum of the scores (S_p) of the k nearby peaks from the TSS.

$$S_g = \sum_{i=1}^{k} S_{pi}$$

- Regulatory potential increases with the number of binding sites [Tang et al., 2011].

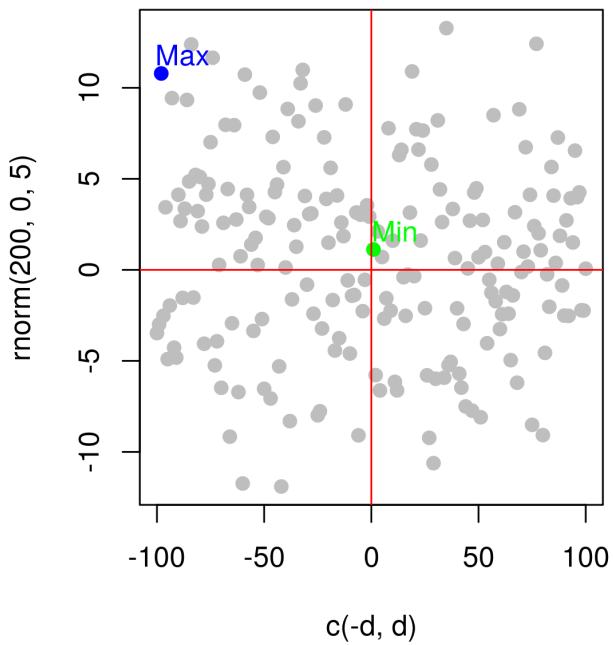


Integrating the factor binding and the expression information

Rank Product (RP_g): the gene score (R_{gb}) is multiplied by the gene statistics (R_{ge}) from differential expression [Breitling et al., 2004].

$$RP_g = \frac{R_{gb} \times R_{ge}}{n^2}$$

- Integrate the binding events and the functional effect [Tang et al., 2011].



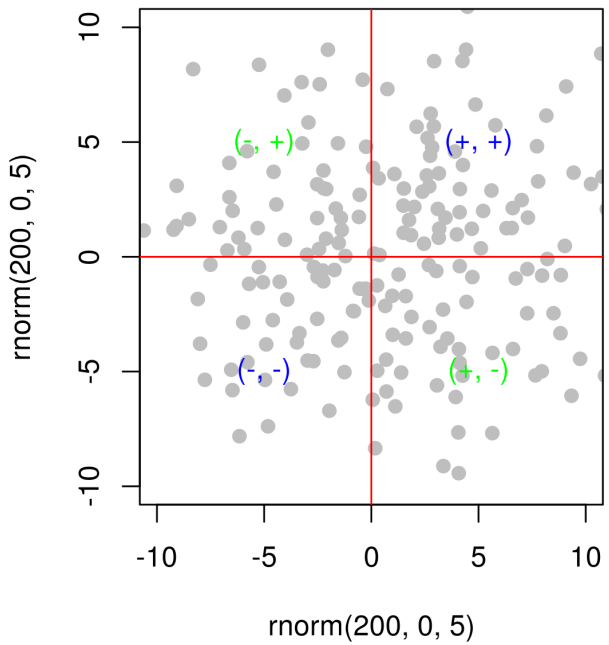
Modeling the interaction of two factors using independent perturbations

Regulatory Interaction (RI): is the product of the gene statistics from differential expression of the perturbation of the two factors (X and Y) separately.

$$RI_g = x_{ge} \times y_{ge}$$

and,

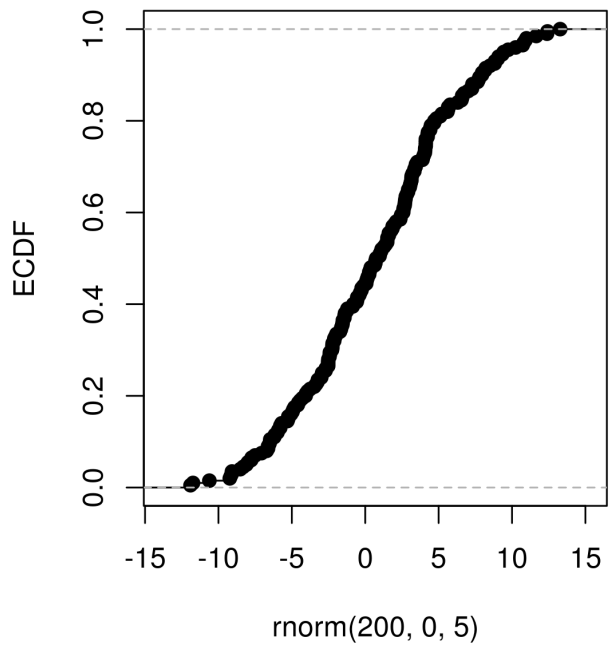
$$RP_g = \frac{R_{gb} \times RI_{ge}}{n^2}$$



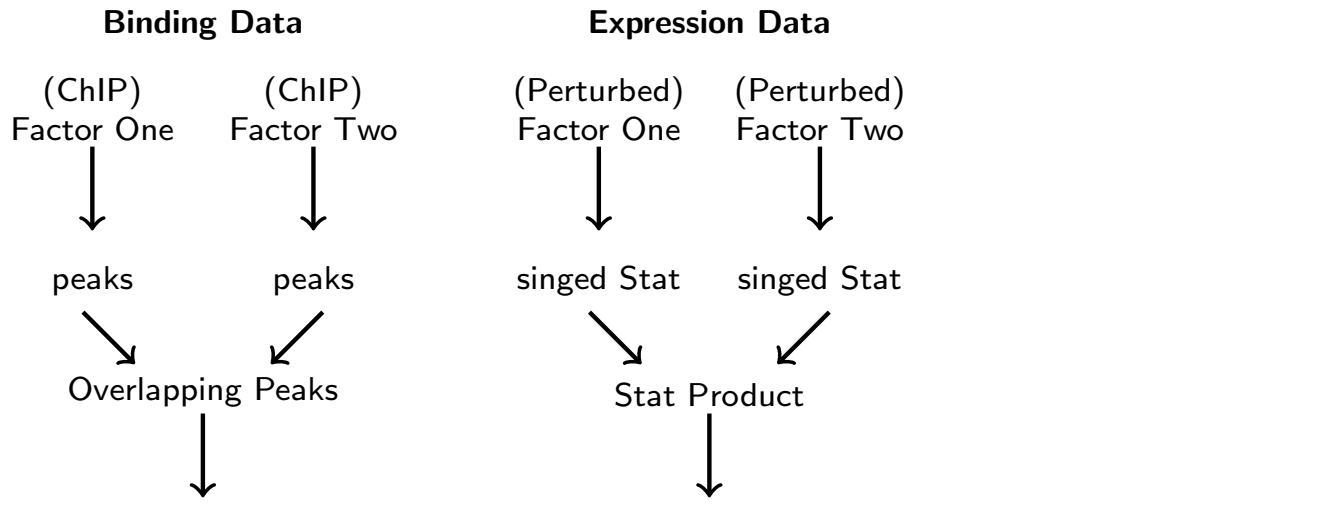
Aggregating the effect of the binding events

Empirical Cumulative Distribution Function (ECDF): the proportion of genes in a category (up- or down-regulated genes) that are ranked at or better than the x-axis (regulatory potential value) value [Tang et al., 2011].

- Aggregate the effect of the factor perturbation in relation to its regulatory potential.



Workflow for integrating binding and expression data



The region regulatory potential as the sum of the transformed distances of the peaks in this region.

The product of the singed statistics of perturbed factors determine their combined function.

The ranking of the region by significance where low values mean higher significance.

Functions in the target R package.

Function	Description	Input	Output
merge_ranges	Merge overlapping peaks & regions.	peaks & regions	Merged ranges
find_distance	Calculate the distance between the centers of peaks & regions.	peaks & regions	Distances
score_peaks	Calculate regulatory scores for peaks in relation to regions.	Distances	Peak scores
score_regions	Calculate regulatory scores for regions.	Peak scores & region IDs	Regions scores
rank_product	Rank regions based on the regulatory potential & expression statistics.	Regions scores, expression statistics & region IDs	Regions rank products
associated_peaks	Select overlapping peaks & regions & calculate a score for each peak in relation to a region.	peaks & regions	Assigned peaks
direct_target	Select & rank regions with overlapping peaks.	peaks & regions	Assigned targets
plot_predictions	Plot the ECDF of the regions' ranks by group.	Ranks & group factor	ECDF plot
test_predictions	Test the ECDF of the ranks in the regions in each group are from different distribution.	Ranks & group factor	t-statistics & p-values

Comparison with existing R packages

- **rTRM** attempts to identify the transcriptional regulatory modules (TRMs) which are complexes of transcription factors and co-factors by integrating ChIP, gene expression and protein-protein interactions [Diez et al., 2013]
- **TFEA.ChIP** takes the approach of curating large quantities of data from different sources and using this data to build a model or database where queries of transcription factor targets can be constructed [Puente-Santamaria et al., 2019].
- **transcriptR** integrates ChIP- and RNA-Seq data for an entirely different purpose [Karapetyan AR, 2019]. It uses the ChIP data to *denovo* identify transcripts which are then used to quantify the expression in the RNA-Seq data.

Limitations of target

- Comparable sets of data for the two factors are required; binding data using ChIP and gene expression data under factor perturbation (overexpression or knockdown).
- Assume that the interaction between two DNA-binding proteins is linear which may not be the case always.
- Cannot detect assisted binding.

Availability

- target is available as an open source R/Bioconductor package (to be submitted)
- An interactive application can be invoked locally through R or accessed directly on the web (`https://mahshaaban.shinyapps.io/target-app/`)
- The source code for the package and the interactive application are available at (`https://github.com/MahShaaban/target`).

Case Study: of two evolutionary and functionally related transcription factors YY1 & YY2

Yin And Yang 1 (YY1)

- Belong to transcription factor GLI- Kruppel class of zinc finger proteins
- Involved in repressing and activating a diverse number of promoters
- Direct histone deacetylases and histone acetyl-transferases to a promoter

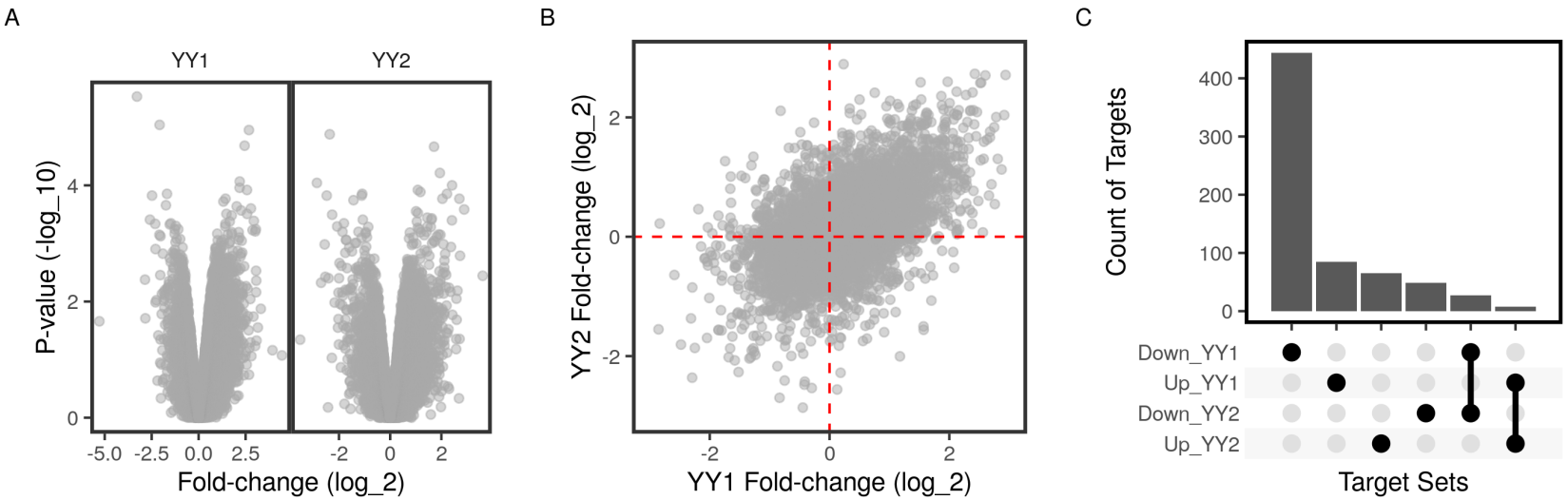
Yin And Yang 2 (YY2)

- Arisen by retrotransposition of the related YY1 gene on chromosome 14
- Exhibit positive and negative control on a large number of genes
- Antagonize YY1 and function in development and differentiation.

Datasets of YY1 and YY2 ChIP-Seq and microarray knockdown in HeLa cells

Dataset	Factor	Type	Source	Ref.
GSE31417	YY1	ChIP-Seq	ChIP-Atlas	[Michaud et al., 2013]
GSE96878	YY2	ChIP-Seq	ChIP-Atlas	[Wu et al., 2017]
GSE14964	YY1-kd	Microarray	KnockTF	[Chen et al., 2010]
GSE14964	YY2-kd	Microarray	KnockTF	[Chen et al., 2010]

Differential expression of YY1 and YY2 in knockdown vs control HeLa cells

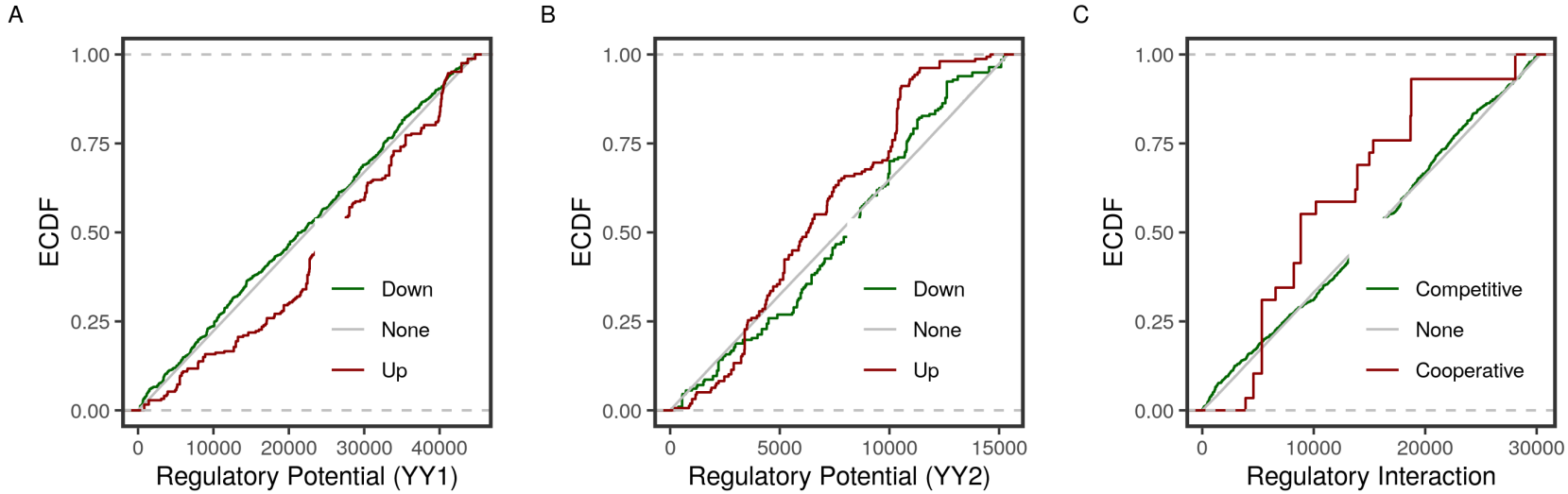


Both factors induce and repress large number of genes.

The effect of the factors knockdown on gene expression is correlated.

The effect of the factors knockdown is correspondent on a small number of shared targets

Predicted functions of YY1 and YY2 on specific and shared targets in HeLa cells



The knockdown of YY1 & YY2 has opposing effects on their specific targets

YY1 & YY2 cooperate on most shared targets, except for a few strong targets where they compete

Testing YY1 and YY2 combined functions

Factor	Test	Statistic	P Value
YY1	Down vs Up	0.79	0e+00
YY2	Up vs Down	0.41	5e-13
Two Factors	Cooperate vs Compete	0.97	0e+00

Summary

- We provided a fast and flexible implementation of the BETA algorithm for predicting direct targets of transcription factors from binding and expression data.
- We extended the method to determine the combined function of two factors on the same region.
- The algorithm is available as an R package and an interactive web application.
- We applied the method to ChIP-Seq and knockdown microarrays of YY1 & 2 in HeLa cells. We found that the two factors cooperate on most shared targets

Find out more

Mahmoud Ahmed, Do Sik Min and **Deok Ryong Kim**. (2020). Integrating binding and expression data to predict transcription factors combined function. *BMC Genomics*. https://10.1186/s12864-020-06977-1

Poster #B2. Until 4:00 PM.

References I

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004).

Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.

FEBS Letters.

Chen, L., Shioda, T., Coser, K. R., Lynch, M. C., Yang, C., and Schmidt, E. V. (2010).

Genome-wide analysis of YY2 versus YY1 target genes.

Nucleic Acids Research, 38(12):4011–4026.

Diez, D., Hutchins, A. P., and Miranda-Saavedra, D. (2013).

Systematic identification of transcriptional regulatory modules from protein-protein interaction networks.

Nucleic Acids Research.

Karapetyan AR (2019).

An Integrative Tool for ChIP- And RNA-Seq Based Primary Transcripts Detection and Quantification.

Michaud, J., Praz, V., James Faresse, N., Jnbaptiste, C. K., Tyagi, S., Schütz, F., and Herr, W. (2013).

HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy.

Genome research, 23(6):907–16.

Puente-Santamaria, L., Wasserman, W. W., and del Peso, L. (2019).

TFEA.ChIP: a tool kit for transcription factor binding site enrichment analysis capitalizing on ChIP-seq datasets.

Bioinformatics.

Rougemont, J. and Naef, F. (2012).

Computational Analysis of Protein–DNA Interactions from ChIP-seq Data.

In *Gene Regulatory Networks*, pages 263–273. Springer.

References II

Tang, Q., Chen, Y., Meyer, C., Geistlinger, T., Lupien, M., Wang, Q., Liu, T., Zhang, Y., Brown, M., and Liu, X. S. (2011).
A comprehensive view of nuclear receptor cancer cistromes.
Cancer research, 71(22):6940–7.

Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., Tang, Q., Meyer, C. A., Zhang, Y., and Liu, X. S. (2013).
Target analysis by integration of transcriptome and ChIP-seq data with BETA.
Nature Protocols.

Wu, X.-N., Shi, T.-T., He, Y.-H., Wang, F.-F., Sang, R., Ding, J.-C., Zhang, W.-J., Shu, X.-Y., Shen, H.-F., Yi, J., Gao, X., and Liu, W. (2017).
Methylation of transcription factor YY2 regulates its transcriptional activity and cell proliferation.
Cell discovery, 3:17035.