Integrating binding and expression data to predict transcription factors combined function

Mahmoud Ahmed and Deok Ryong Kim

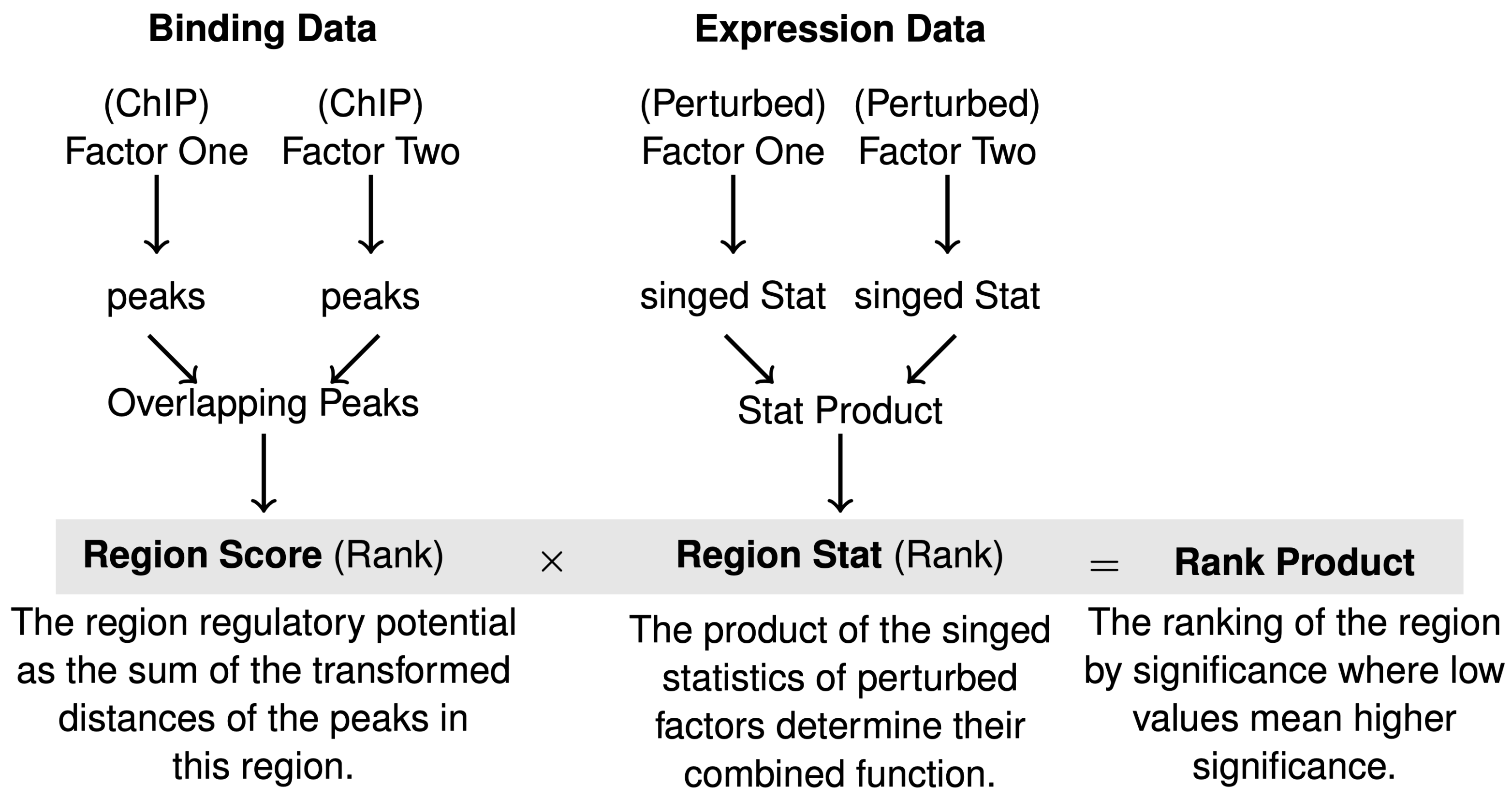
Department of Biochemistry and Convergence Medical Sciences and Institute of Health Sciences, Gyeongsang National University School of Medicine, Jinju, Korea

Summary

Background Transcription factor binding to the regulatory region of a gene induces or represses its gene expression. Transcription factors share their binding sites with other factors, co-factors and/or DNA-binding proteins. These proteins form complexes which bind to the DNA as one-units. The binding of two factors to a shared site does not always lead to a functional interaction. **Results** We propose a method to predict the combined functions of two factors using comparable binding and expression data (target) (Figure 1). We based this method on binding and expression target analysis (BETA), which we re-implemented in R and extended for this purpose (Table 1). target ranks the factor's targets by importance and predicts the dominant type of interaction between two transcription factors. We applied the method to simulated and real datasets of transcription factor-binding sites and gene expression under perturbation of factors. Yin Yang 1 transcription factor (YY1) and YY2 are evolutionary and functionally related. The knockdown of either factors produced wide changes in the gene expression of HeLa cells (Figure 2). We found that YY1 and YY2 have antagonistic and independent regulatory targets in HeLa cells, but they may cooperate on a few shared targets (Figure 3 & Table 2). **Conclusion** We developed an R package and a web application to integrate binding (ChIP-seq) and expression (microarrays or RNA-seq) data to determine the cooperative or competitive combined function of two transcription factors.

Background

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. Two factors work cooperatively when they share a binding site and where they both induce or repress the gene [2]. By contrast, two factors may compete on a specific sites where the binding of either has an effect on the gene expression opposite to the other [3]. In this study, we provide an implementation of an algorithm to integrate the binding and expression data to predict transcription factors direct target and extend the method to predict the combined functions of two factors using comparable binding and expression data.



Binding Data

(ChIP) Factor One

(ChIP) Factor Two

peaks

peaks

Overlapping Peaks

Region Score (Rank)

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

Expression Data

(Perturbed) Factor One

(Perturbed) Factor Two

singed Stat

singed Stat

Stat Product

Region Stat (Rank)

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

×

=

Rank Product

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

Figure 1: Integrating binding and expression data to predict the combined function of transcription factors. The binding data from ChIP experiments of two factors are used to find the peaks in the genomic regions of interest. The distances between the peaks and the regions are used to calculate peak scores. The sum of the scores of all peaks in assigned to a region is its regulatory potential. The product of signed statistics from gene expression experiments of the factors perturbation is used to determine the magnitude and the direction of their regulatory interactions. The rank product of the region score and statistics is the region significance.

Implementation

Binding and expression target analysis (BETA)

The BETA algorithm in its simplest form, *minus* [6], is composed of three steps:

1. Select the peaks (p) within a certain range from the regions of interest (g).
2. Calculate the distance (Δ) between the center of the peak and each of the regions expressed relative to a distance of 100 kb.
3. Calculate the peak scores (S_p) as the transformed exponential of the distance, Δ, as follows:

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

4. Calculate the region/gene regulatory potential (S_g) as the sum of the scores, S_p [5], as follows:

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

where p is $\{1, ..., k\}$ peaks near the region of interest. In BETA *basic*, another step is added to predict real region/gene targets.

5. Rank all regions based on their regulatory potential, S_g, to give their binding potential (R_{gb}) and based on their differential expression (R_{ge}). The product of the two ranks predicts real region/gene targets.

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

where n is the number of regions g.

Regulatory interaction (RI) term for predicting combined functions

To determine the relation of two factors x and y on a common peak near a region of interest, we define a new term; the regulatory interaction (RI) as the product of two signed statistics from comparable perturbation experiments. The rank of this term is used to calculate a rank product (PR_g) for each region of interest as described above [1].

$$RI_g = x_{ge} \times y_{ge} \quad \text{and} \quad RP_g = \frac{R_{gb} \times RI_{ge}}{n^2}$$

This term would represent the interaction magnitude assuming a linear relation between the two factor. The sign of the term would define the direction of the relation were positive means cooperative and negative means competitive. The regions can be divided into meaningful groups and tested for significance. The original BETA paper suggested generating distribution functions for the groups and apply the one-tailed Kolmogorov-Smirnov test to test whether the groups are drawn from the same distribution [4].

Table 1: Functions of 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

Results

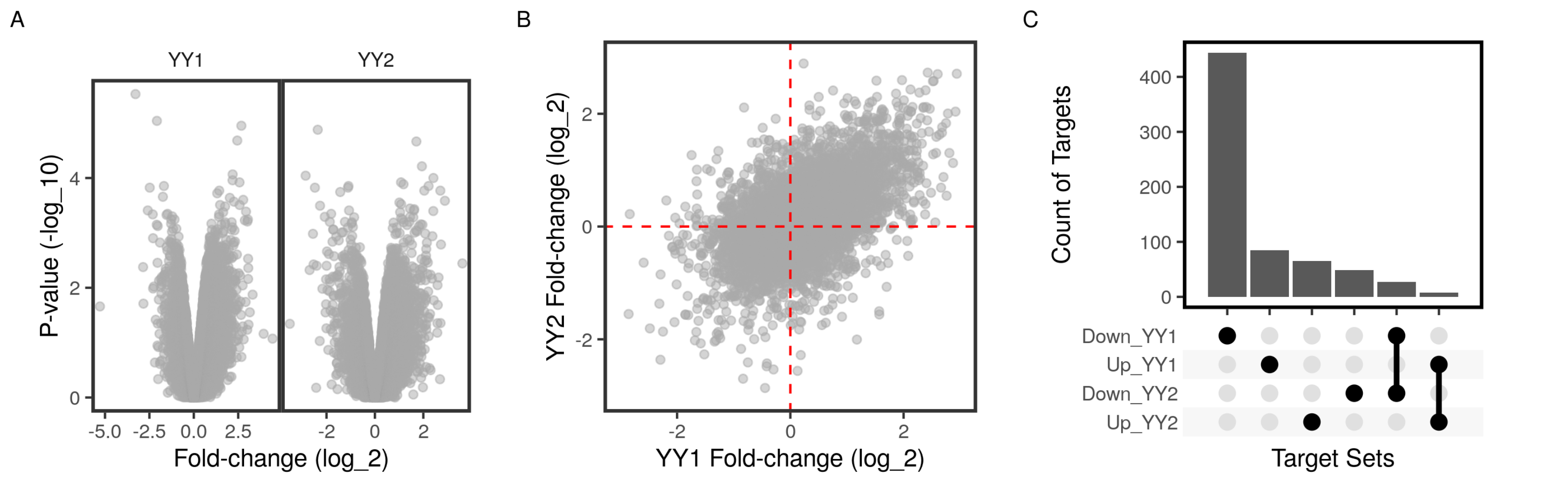


Figure 2: Differential expression of YY1 and YY2 in knockdown vs control HeLa cells. Probe intensities from microarrays of YY1 or YY2 (n = 3) knockdown and control (n = 3) were aggregated by gene and used to perform differential expression analysis (GSE14964). The gene expression in the YY1 and YY2-knockdown samples was compared to the control samples individually. A) Volcano plots show the fold-change ($\log_2$) and p-values ($-\log_1 0$) in each comparison. B) The fold-change ($\log_2$) of the YY1 and YY2-knockdown are shown as scatter plot. C) The count of regulated (Up/Down) genes in by YY1 or YY2-knockdown and their intersections are shown as bars.

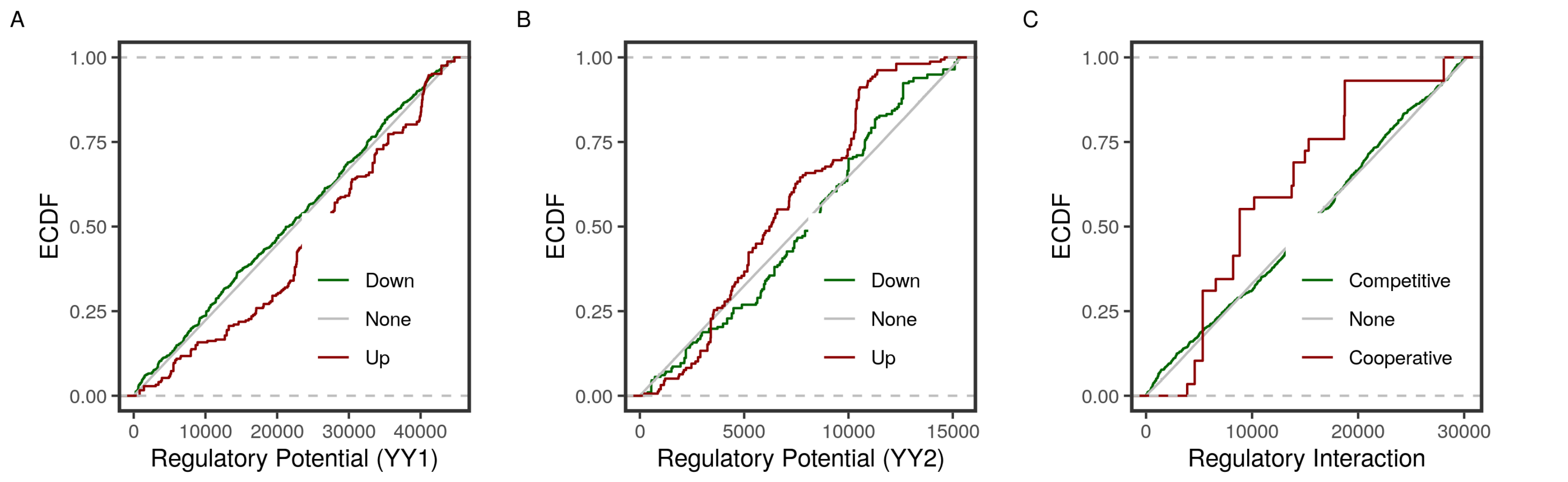


Figure 3: Predicted function of YY1 and YY2 on specific and shared targets in HeLa cells. The target analysis was applied using two sets of data from the HeLa cells; expression data in YY1 and YY2-knockdown (GSE14964) and two sets of ChIP peaks using antibodies for YY1 (GSE31417) and YY2 (GSE96878). Predicted targets were ranked based on their distance to the transcription start sites (TSS) and their fold-change. The empirical distribution function (ECDF) of each group of targets (Down, None or Up-regulated genes) of A) YY1 and B) YY2 was calculated. C) The shared targets were ranked based on their distance to the TSS in which they had overlapping peaks and the product of the corresponding fold-changes. The ECDF of each group of targets (Competitively, None or Cooperatively regulated genes) was calculated.

Table 2: Testing YY1 and YY2 target groups.

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

References

[1] R. Breitling et al. "Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments". In: *FEBS Letters* (2004).

[2] C. Hernandez-Munain, J. L. Roberts, and M. S. Krangel. "Cooperation among Multiple Transcription Factors Is Required for Access to Minimal T-Cell Receptor α-Enhancer Chromatin In Vivo". In: *Molecular and Cellular Biology* (1998).

[3] L. J. Norton et al. "Direct competition between DNA binding factors highlights the role of Krüppel-like Factor 1 in the erythroid/megakaryocyte switch." In: *Scientific reports* 7.1 (2017), p. 3137.

[4] A. Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (Oct. 2005), pp. 15545–15550.

[5] Q. Tang et al. "A comprehensive view of nuclear receptor cancer cistromes". In: *Cancer Research* (2011).

[6] S. Wang et al. "Target analysis by integration of transcriptome and ChIP-seq data with BETA". In: *Nature Protocols* (2013).

The scripts to reproduce this analysis, figures and tables are available here `https://github.com/BCMSLab/target_ranking` or by directly scanning the QR code. The github repository contains the instructions for setting up a software environment, obtaining the data and running the analysis.