



SOFTWARE TOOL ARTICLE

REVISED target: an R package to predict combined function of transcription factors [version 3; peer review: 2 approved with reservations]

Mahmoud Ahmed , Deok Ryong Kim 

Department of Biochemistry and Convergence Medical Sciences and Institute of Health Sciences, Gyeongsang National University School of Medicine, Jinju, Gyeongnam, 52727, South Korea

V3 First published: 05 May 2021, 10:344
<https://doi.org/10.12688/f1000research.52173.1>
 Second version: 10 Nov 2021, 10:344
<https://doi.org/10.12688/f1000research.52173.2>
 Latest published: 16 Nov 2021, 10:344
<https://doi.org/10.12688/f1000research.52173.3>

Abstract

Researchers use ChIP binding data to identify potential transcription factor binding sites. Similarly, they use gene expression data from sequencing or microarrays to quantify the effect of the transcription factor overexpression or knockdown on its targets. Therefore, the integration of the binding and expression data can be used to improve the understanding of a transcription factor function. Here, we implemented the binding and expression target analysis (BETA) in an R/Bioconductor package. This algorithm ranks the targets based on the distances of their assigned peaks from the transcription factor ChIP experiment and the signed statistics from gene expression profiling with transcription factor perturbation. We further extend BETA to integrate two sets of data from two transcription factors to predict their targets and their combined functions. In this article, we briefly describe the workings of the algorithm and provide a workflow with a real dataset for using it. The gene targets and the aggregate functions of transcription factors YY1 and YY2 in HeLa cells were identified. Using the same datasets, we identified the shared targets of the two transcription factors, which were found to be, on average, more cooperatively regulated.

Keywords

transcription-factors, DNA-binding, gene-expression, r-package, bioconductor, workflow

Open Peer Review

Reviewer Status ? ?

Invited Reviewers

1 2

version 3

(revision)
16 Nov 2021

version 2


(revision)
10 Nov 2021

version 1

05 May 2021

?
report

?
report

1. **Shulan Tian**, Mayo Clinic, Rochester, USA
Yan Huihuang, Mayo Clinic, Rochester,, USA
2. **Mireia Ramos-Rodríguez** , Pompeu Fabra University, Barcelona, Spain

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the RPackage gateway.



This article is included in the **Bioconductor** gateway.

Corresponding author: Deok Ryong Kim (drkim@gnu.ac.kr)

Author roles: **Ahmed M:** Conceptualization, Formal Analysis, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kim DR:** Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) of the Korea government [2015R1A5A2008833 and 2020R1A2C2011416].

Copyright: © 2021 Ahmed M and Kim DR. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ahmed M and Kim DR. **target: an R package to predict combined function of transcription factors [version 3; peer review: 2 approved with reservations]** F1000Research 2021, 10:344 <https://doi.org/10.12688/f1000research.52173.3>

First published: 05 May 2021, 10:344 <https://doi.org/10.12688/f1000research.52173.1>

REVISED Amendments from Version 2

This revised version of the article contains a few minor corrections

- The regions of interest were extended to 100kb upstream and 200bp downstream from the transcription start sites.
- The section "Predicting gene targets of individual transcription factors", was updated to briefly describe what the two main functions do.
- On several occasions, we used the term "transcription factor" instead of "factor".

Any further responses from the reviewers can be found at the end of the article

Introduction

The binding of a transcription factor to a genomic region (e.g., gene promoter) can have the effect of inducing or repressing its expression Latchman¹. The binding sites can be identified using ChIP experiments. High through-put ChIP experiments produce hundreds or thousands of binding sites for most transcription factors Johnson *et al.*². Therefore, methods to determine which of these sites are true binding sites and whether they are functional or not are needed Ucar *et al.*³. On the other hand, perturbing the transcription factor by over-expression or knockdown and measuring the gene expression changes provide valuable information on the function of the transcription factor Tran *et al.*⁴. Methods exist to integrate the binding data and the factor perturbation gene expression to predict the real target regions (e.g., genes)^{5,6}. This article presents a workflow for using the target package to integrate binding and expression data to predict the shared targets and the combined function of two transcription factors.

To illustrate the utility of this workflow, we applied it to the binding and expression data of the transcription factors YY1 and YY2. We asked whether the two factors cooperate or compete on their shared targets in HeLa cells.

Methods

Implementation

We developed an open-source R/Bioconductor package target to implement BETA for predicting direct transcription factor targets from binding and expression data. The details of the algorithm were described here Wang *et al.*⁶. In addition, our implementation extends BETA to apply for transcription factor combinations (Ahmed *et al.*⁷). Briefly, we identify the transcription factor potential binding sites by ChIP-sequencing and gene expression under factor perturbation by microarrays or sequencing. Next, we score the peaks based on their distances to the transcription start sites. The sum of the scores of the individual peaks in a certain region of interest is the region's regulatory potential. The signed statistics (fold-change or t-statistics) from the differential gene expression of the transcription factor perturbation reflect the transcription factor effects. The product of the ranks of the regulatory potential and the signed statistics is the final rank of the regions.

To predict the combined function of two transcription factors, two sets of data are required. The overlapping peaks are the potential binding sites. The product of the two signed statistics is the transcription factor function. When the two transcription factors agree in the direction of the regulation of a region where they both bind, they could be said to cooperate on this region. When the sign is opposite, they could be said to regulate that region competitively.

The package leverages the Bioconductor data structures such as GRanges and DataFrame to provide fast and flexible computation on the data Huber *et al.*⁸. Similar to the original python implementation, the input data are the identified peaks from the ChIP-Seq experiment and the expression data from RNA-Seq or microarrays perturbation experiment. The final output is the peaks associated with the transcription factor binding and the predicted direct targets. We use the terms "peaks" to refer to the GRanges object that contains the coordinates of the peaks. Likewise, we use the term "region" to refer to a similar object that contains the information on the regions of interest; genes, transcripts, promoter regions, etc. In both cases, additional information on the ranges can be added to the object as metadata.

Operation

The algorithm was implemented in R (≥ 3.6) and should run on any operating system. Libraries required for running the workflow are listed and loaded below. Alternatively, a docker image is available with R and the libraries installed on an Ubuntu image: https://hub.docker.com/r/bcmslab/target_flow

```
# load required libraries
library(GenomicRanges)
library(Biostrings)
library(rtracklayer)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
library(BSgenome.Hsapiens.UCSC.hg19)
library(org.Hs.eg.db)
library(tidyverse)
library(BCRANK)
library(seqLogo)
library(target)
```

Use case

YY1 and YY2 belong to the same family of transcription factors. YY1 is a zinc finger protein that directs histone deacetylase and acetyltransferases of the promoters of many genes. The protein also binds to the enhancer regions of many of its targets. The binding of YY1 to the regulatory regions of genes results in the induction or repression of their expression. YY2 is a paralog of YY1. Similarly, it is a zinc finger protein with both activation or repression functions on its targets. We will attempt to answer the following questions using the target analysis: Do the two transcription factors share the same target genes? What are the consequences of the binding of each transcription factor on its targets? If the two transcription factors share binding sites, what is the function of the two transcription factors binding to these sites?

To answer these questions, we use publicly available datasets to model the binding and gene expression under the transcription factors perturbations (Table 1). This dataset was obtained in the form of differential expression between the two conditions from KnockTF Feng *et al.*⁹. The first dataset is gene expression profiling using microarrays of YY1/YY2 knockdown and control HeLa cells. Next, the binding sites of the transcription factors in HeLa cells were determined using two ChIP-Seq datasets. The ChIP peaks were acquired in the form of bed files from ChIP-Atlas Oki *et al.*¹⁰. Finally, we used the UCSC hg19 human genome to extract the genomic annotations.

Briefly, we first prepared the three sources of data for the target analysis. Then we predict the specific targets for each individual transcription factor. Third, we predict the combined function of the two transcription factors on the shared target genes. Finally, we show an example of a motif analysis of the competitively and cooperatively regulated targets.

```
if(!file.exists('data.zip')) {
  # download the manuscript data
  download.file('https://ndownloader.figshare.com/articles/10918463/versions/1',
               destfile = 'data.zip')

  # decompress file
  unzip('data.zip', exdir = 'data')
}
```

Table 1. Expression and binding data of YY1 and YY2 in HeLa cells.

GEO ID	Data Type	Design	Ref.
GSE14964	Microarrays	YY#-knockdown	Chen <i>et al.</i> ¹¹
GSE31417	ChIP-Seq	YY1 vs input	Michaud <i>et al.</i> ¹²
GSE96878	ChIP-Seq	YY2 vs input	Wu <i>et al.</i> ¹³

Preparing the binding data

The ChIP peaks were downloaded in the form of separate bed files for each transcription factor. We first locate the files in the `data/` directory and load the files using `import.bed`. Then the data is transformed into a suitable format, `GRanges`. The resulting object, `peaks`, is a list of two `GRanges` items, one for each factor.

```
# locate the peaks bed files
peak_files <- c(YY1 = 'data/Oth.Utr.05.YY1.AllCell.bed',
               YY2 = 'data/Oth.Utr.05.YY2.AllCell.bed')

# load the peaks bed files as GRanges
peaks <- map(peak_files, ~GRanges(import.bed(.x)))
```

Preparing the expression data

The differential expression data were downloaded in tabular format. After locating the files in `data/`, we read the files using `read_tsv` and select and rename the relevant columns. The resulting object, `express`, is a list of two `tibble` items.

```
# locate the expression text files
expression_files <- c(YY1 = 'data/DataSet_01_18.tsv',
                    YY2 = 'data/DataSet_01_19.tsv')

# load the expression text files
express <- map(expression_files,
              ~read_tsv(.x, col_names = FALSE) %>%
                dplyr::select(2, 3, 7, 9) %>% #9
                setNames(c('tf', 'gene', 'fc', 'pvalue')) %>%
                filter(tf %in% c('YY1', 'YY2')) %>%
                na.omit())
```

The knockdown of either transcription factor in HeLa cells seems to change the expression of many genes in either direction (Figure 1A&B). Moreover, the changes resulting from the separate knockdown of the transcription factors are correlated ($r = 0.56$, $P < 0.0001$) (Figure 1C). These observations suggest that many of the regulated genes are shared targets of the two transcription factors, or they respond similarly to their perturbation of either factor.

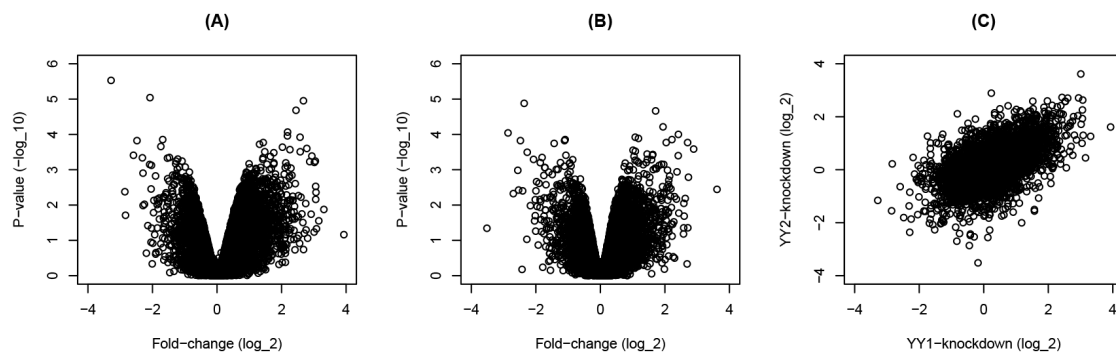


Figure 1. Differential expression between transcription factor knockdown and control HeLa cells. Gene expression was compared between transcription factors knockdown and control HeLa cells. The fold-change and p-values of (A) YY1- and (B) YY2-knockdown are shown as volcano plots. (C) Scatter plot of the fold-change of the YY1- and YY2-knockdown.

```

# Figure 1
par(mfrow = c(1, 3))

# volcano plot of YY1 knockdown
plot(express$YY1$fc,
      -log10(express$YY1$pvalue),
      xlab = 'Fold-change (log_2)',
      ylab = 'P-value (-log_10)',
      xlim = c(-4, 4), ylim = c(0, 6))
title('(A)')

# volcano plot of YY2 knockdown
plot(express$YY2$fc,
      -log10(express$YY2$pvalue),
      xlab = 'Fold-change (log_2)',
      ylab = 'P-value (-log_10)',
      xlim = c(-4, 4), ylim = c(0, 6))
title('(B)')

# plot fold-change of YY1 and YY2
plot(express$YY1$fc[order(express$YY1$gene)],
      express$YY2$fc[order(express$YY2$gene)],
      xlab = 'YY1-knockdown (log_2)',
      ylab = 'YY2-knockdown (log_2)',
      xlim = c(-4, 4), ylim = c(-4, 4))
title('(C)')

```

Preparing genome annotation

express records the gene information using the gene Symbols. We mapped the Symbols to the Entrez IDs before extracting the genomic coordinates. To do that, we use the `org.Hs.eg.db` to convert between the identifiers. Next, we use the `TxDb.Hsapiens.UCSC.hg19.knownGene` to get the genomic coordinates for the transcripts and extend them to 100kb upstream and 200bp downstream from the transcription start sites.

```

# load genome data
symbol_entrez <- AnnotationDbi::select(org.Hs.eg.db,
                                     unique(c(express$YY1$gene)),
                                     'ENTREZID', 'SYMBOL') %>%
  setNames(c('gene', 'gene_id'))

# format genome to join with express
genome <- promoters(TxDb.Hsapiens.UCSC.hg19.knownGene,
                   upstream = 100000, # (default) downstream = 200,
                   columns = c('tx_id', 'tx_name', 'gene_id')) %>%
  as_tibble() %>% mutate(gene_id = as.character(gene_id))

```

The resulting object, `genome`, from the previous step is a tibble that shares the column `gene_id` with the expression data `express`. Now the two objects can be merged. The merged object, `regions`, is similarly a tibble containing genome and expression information of all common genes.

```

# make regions by merging the genome and express data
regions <- map(express,
              ~inner_join(genome, symbol_entrez) %>%
                inner_join(.x) %>%
                makeGRangesFromDataFrame(keep.extra.columns = TRUE))

```

Predicting gene targets of individual transcription factors

The standard target analysis identifies associated peaks using `associated_peaks` and direct targets using `direct_targets`. `associated_peaks` calculates and transforms the distances between the peaks and TSSs. Then it assigns the peaks to the nearest transcript. `direct_targets` calculates the final gene ranks based on the distances and the change in gene expression. The inputs for these functions are the objects `peaks` and `regions` from the previous steps in addition to the column names for regions `regions_col` or the region and the statistics column `stats_col`, which is the fold-change in this case. The resulting objects are `GRanges` for the identified peaks assigned to the regions, `ap`, or the ranked targets. Several columns are added to the metadata objects of the `GRanges` to save the output.

```
# get associated peaks
ap <- map2(peaks, regions,
          ~associated_peaks(peaks=.x,
                           regions = .y,
                           regions_col = 'tx_id'))

# get direct targets
dt <- map2(peaks, regions,
          ~direct_targets(peaks=.x,
                          regions = .y,
                          regions_col = 'tx_id',
                          stats_col = 'fc'))
```

To determine the dominant function of a transcription factor, we divide the targets by the direction of the effect of transcription factor knockdown. We group the targets by the change in gene expression (regulatory potential). We use the empirical distribution function (ECDF) to show the fraction of targets with a specified regulatory potential or less. Because we use the ranks rather than the absolute value of the regulatory potential, the lower the rank, the higher the potential. Then, {we compare} the groups of targets to each other or to a theoretical distribution.

```
# Figure 2
par(mfrow = c(1, 3))

# plot distance by score of associate peaks
plot(ap$YY1$distance, ap$YY1$peak_score,
     xlab = 'Distance', ylab = 'Peak Score',
     main = '(A)')
points(ap$YY2$distance, ap$YY2$peak_score)

# make labels, colors and groups
labs <- c('Down', 'None', 'Up')
cols <- c('green', 'gray', 'red')

# make three groups by quantiles
groups <- map(dt, ~{
  cut(.x$stat, breaks = 3, labels = labs)
})

# plot the group functions
pmap(list(dt, groups, c('(B)', '(C)')), function(x, y, z) {
  plot_predictions(x$score_rank,
                  group = y, colors = cols, labels = labs,
                  xlab = 'Regulatory Potential', ylab = 'ECDF')
  title(z)
})
```

The scores of the individual peaks are a decreasing function of the distance from the transcription start sites—the closer the transcription factor binding site from the start site, the higher the score. The distribution of these scores is very similar for both transcription factors (Figure 2A). The ECDF of the down-regulated of YY1 is higher than that of up- and none-regulated targets (Figure 2B). Therefore, the absence of YY1 on its targets results in aggregate in their downregulation. If indeed these are true targets, then we expect YY1 to induce their expression. The opposite is true for YY2, where more high-ranking targets are up-regulated by the transcription factor knockdown (Figure 2C).

```
# Table 2
# test individual factor functions
map2(dt, groups,
      ~test_predictions(.x$rank,
                       group = .y,
                       compare = c('Down', 'Up')))
```

To formally test these observations, we use the Kolmogorov-Smirnov (KS) test. First, we compare the distributions of the two groups for equality. If one lies on either side of the other, then they must be drawn from different distributions. Here, we contrast the up and down-regulated functions for both transcription factors (Table 2). In both cases, the distributions of the two groups were significantly different from one another.

Predicting the shared targets of two transcription factors

Using target to predict the shared target genes and the combined function of the two transcription factors is a variation of the previous analysis. First, the shared/common peaks are generated using the overlap of their genomic coordinates, subsetByOverlaps. Second, Instead of one, two columns for the differential expression statistics, one for each transcription factor is needed; these are supplied to the argument stats_col in the same way. Here, common_peaks and both_regions are the main inputs for the analysis functions.

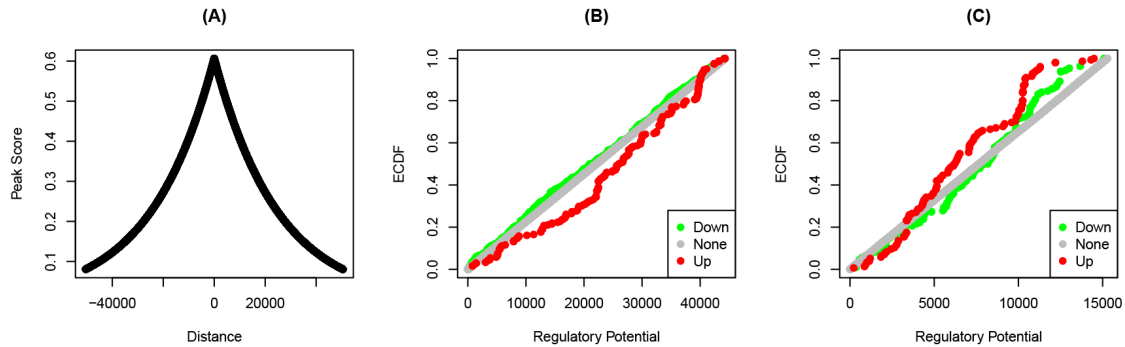


Figure 2. Predicted functions of YY1 and YY2 on their specific targets. Bindings peaks of the transcription factors in HeLa cells were determined using ChIP-Seq. Distances from the transcription start sites, and the transformed distances of the (A) YY1 and YY2 peaks are shown. The regulatory potential of each gene was calculated using target. Genes were grouped into up, none, or down-regulated based on the fold-change. The empirical cumulative distribution functions (ECDF) of the groups of (B) YY1 and (C) YY2 targets are shown at each regulatory potential rank.

Table 2. Testing for statistical significance of the regulated gene groups.

Factor	Statistic	P.value	Method	Alternative
YY1	0.224	2.2e-16	Two-sample KS test	two-sided
YY2	0.149	2.5e-15	Two-sample KS test	two-sided


```

# merge and name peaks
common_peaks <- GenomicRanges::reduce(subsetByOverlaps(peaks$YY1, peaks$YY2))
common_peaks$name <- paste0('common_peak_', 1:length(common_peaks))

# bind express tables into one
both_express <- bind_rows(express) %>%
  nest(fc, pvalue, .key = 'values_col') %>%
  spread(tf, values_col) %>%
  unnest(YY1, YY2, .sep = '_')

# make regions using genome and expression data of both factors
both_regions <- inner_join(genome, symbol_entrez) %>%
  inner_join(both_express) %>%
  makeGRangesFromDataFrame(keep.extra.columns = TRUE)

# get associated peaks with both factors
common_ap <- associated_peaks(peaks = common_peaks,
                             regions = both_regions,
                             regions_col = 'tx_id')

# get direct targets of both factors
common_dt <- direct_targets(peaks = common_peaks,
                            regions = both_regions,
                            regions_col = 'tx_id',
                            stats_col = c('YY1_fc', 'YY2_fc'))

```

The output, `associated_peaks`, is similar to before. `direct_targets` is the same, but the `stat` and the `stat_rank` columns carry the product and the rank of the two statistics provided in the previous step.

We can also visualize the output in a similar way. The targets are divided into three groups based on the statistics product. When the two statistics agree in the sign, the product is positive. This means the knockdown of either transcription factor results in the same direction change in the target gene expression. Therefore, the two transcription factors would cooperate if they bind to the same site on that gene. The reverse is true for targets with oppositely signed statistics. The two transcription factors would be expected to compete on these targets for inducing opposing changes in the expression.

```

# Figure 3
par(mfrow = c(1, 2))

# plot distance by score for associated peaks
plot(common_ap$distance,
      common_ap$peak_score,
      xlab = 'Distance',
      ylab = 'Peak Score')
title('(A)')

# make labels, colors and groups
labs <- c('Competitive', 'None', 'Cooperative')
cols <- c('green', 'gray', 'red')

# make three groups by quantiles
common_groups <- cut(common_dt$stat,
                    breaks = 3,
                    labels = labs)

# plot predicted function
plot_predictions(common_dt$score_rank,
                group = common_groups,
                colors = cols, labels = labs,
                xlab = 'Regulatory Interaction', ylab = 'ECDF')
title('(B)')

```

The common peak distances and scores take the same shape (Figure 3A). Furthermore, the two transcription factors seem to cooperate on more of the common target than any of the two other possibilities (Figure 3B). This observation can be tested using the KS test. The curve of the cooperative targets lies above that of none and competitively regulated targets (Table 3).

```
# Table 3
# test factors are cooperative
test_predictions(common_dt$score_rank,
                 group = common_groups,
                 compare = c('Cooperative', 'None'),
                 alternative = 'greater')

# test factors are more cooperative than competitive
test_predictions(common_dt$score_rank,
                 group = common_groups,
                 compare = c('Cooperative', 'Competitive'),
                 alternative = 'greater')
```

Binding motif analysis

The users can perform any number of downstream analyses on the final output. For example, we could apply binding motif analysis to the groups of regulated targets. In this example, all the motif analysis itself is handled by the BCRANK package Ameer *et al.*¹⁴. Here, we explain how to prepare the input from the shared peaks and target objects produced in the last step.

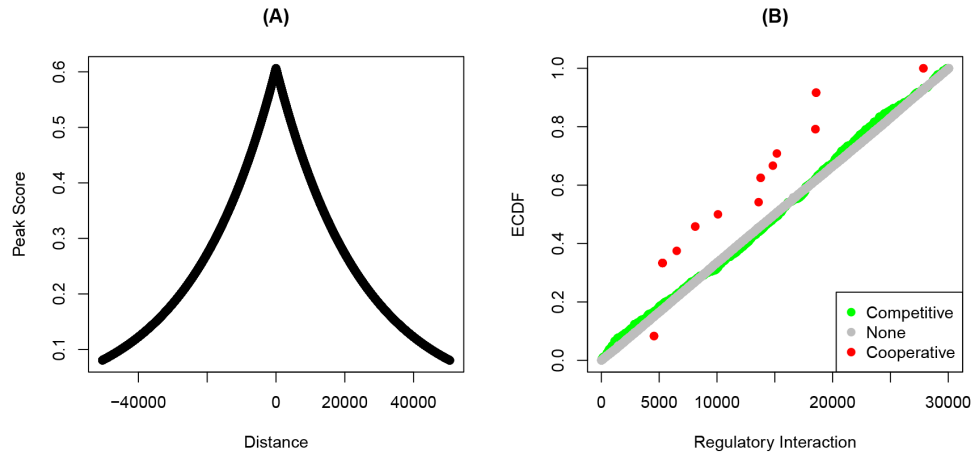


Figure 3. Predicted function of YY1 and YY2 on their shared targets. Shared bindings sites of YY1 and YY2 in HeLa cells were determined using the overlap of the individual transcription factor ChIP-Seq peaks. (A) Distances from the transcription start sites, and the transformed distances of the shared peaks are shown. The regulatory interaction of each gene was calculated using target. Genes were grouped into cooperatively, none, or competitively regulated based on the product of the fold-changes from YY1- and YY2-knockdown. (B) The empirical cumulative distribution functions (ECDF) of the targets groups are shown at each regulatory potential rank.

Table 3. Testing for statistical significance of combined functions of the two transcription factors.

Compare	Statistic	P.value	Method	Alternative
Coop vs None	0.168	1.5e-30	KS test	The CDF of x lies above that of y
Coop vs Comp	0.151	2.2e-16	KS test	The CDF of x lies above that of y

First, we extract the transcript IDs of the targets in their respective groups. Then the peaks assigned to these targets are ordered and sliced.

```
# group peaks by their assigned targets
peak_groups <- split(common_dt$tx_id, common_groups)

# reorder peaks and get top n peaks
peak_groups <- lapply(peak_groups, function(x) {
  # get peaks in x targets group
  p <- common_ap[common_ap$assigned_region %in% unique(x)]

  # order peaks by score
  p <- p[order(p$peak_score, decreasing = TRUE)]

  # get n top peaks
  p[seq_len(ifelse(length(p) > 50, 50, length(p)))]
})
```

The input for `bcrank` is a fasta file with the sequence of the regions to look for frequent motifs. We used the `BSgenome.Hsapiens.UCSC.hg19` to extract the sequences of the common peaks in the competitive and cooperative target groups. The sequences are first written to a temporary file and feed to the search function.

```
bcout <- map(peak_groups[c('Competitive', 'Cooperative')], ~{
  # extract sequences of top peaks from the hg19 genome
  pseq <- getSeq(BSgenome.Hsapiens.UCSC.hg19, names = .x)

  # write sequences to fasta file
  tmp_fasta <- tempfile()
  writeXStringSet(pseq, tmp_fasta)

  # set random see
  set.seed(1234)

  # call bcrank with the fasta file
  bcrank(tmp_fasta, silent = TRUE)
})
```

The sequences in the search path of the regions of interest are shown in (Figure 4). In the competitively regulated regions, one sequence was more frequent than all other sequences. By contrast, no sequence was uniquely frequent in the regions of cooperative targets.

```
# Figure 4
par(mfrow = c(1, 2))
# plot the occurrences of consensus sequence in the regions
map2(bcout, c('(A)', '(B)'), ~{
  plot(toptable(.x, 1))
  title(.y)
})
```

The most frequent motifs in the two groups are shown as seq logos using the `seqLogo` package (Figure 5).

```
# Figure 5
# plot the sequence of the predicted motifs
map(bcout, c('(A)', '(B)'), ~{
  seqLogo(pwm(toptable(.x, 1)))
  title(.y)
})
```

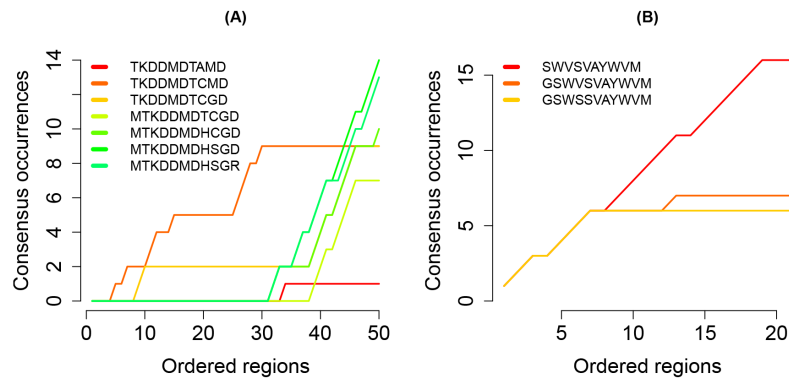


Figure 4. Occurrences of consensus sequences in the ranked regions. The number of occurrences of the sequences in the search path in the regions of (A) competitively and (B) cooperatively regulated regions.

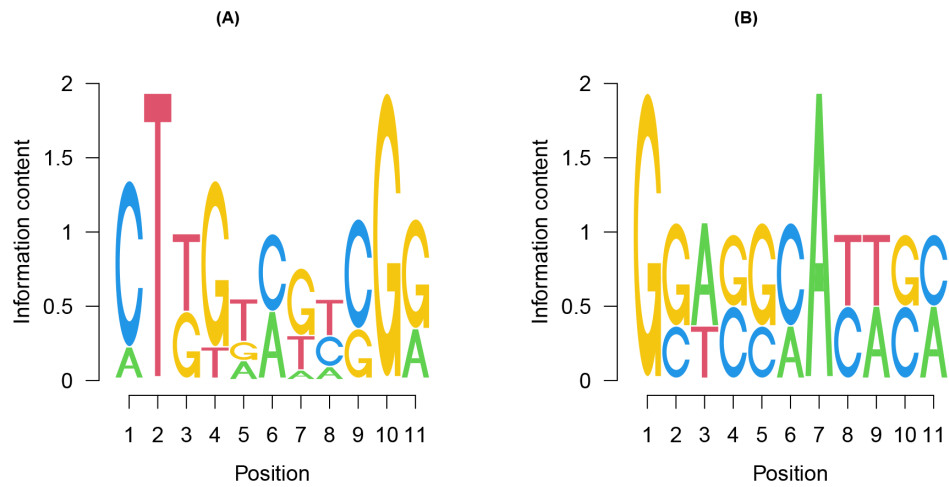


Figure 5. Predicted motifs of the cooperative and competitive binding sites. The position weight matrices of the most frequent motifs in the (A) competitively and (B) cooperatively regulated regions were calculated and shown as sequence logos. y-axis represents the information content at each position. The size of each letter represents the frequency in which the letter occurs at that position.

Summary

In this article, we present a workflow for predicting the direct targets of a transcription factor by integrating binding and expression data. The target package implements the BETA algorithm ranking gene targets based on the distances of the CHIP peaks of the transcription factor relative to the TSSs of the genes and the differential expression of the transcription factor perturbation. To predict the combined function of two transcription factors, two sets of data are used to find the shared peaks and the rank product of their differential expression statistics.

Data availability

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

Software available from: <https://doi.org/doi:10.18129/B9.bioc.target¹⁵>

Source code available from: <https://github.com/MahShaaban/target>

Archived source code as at time of publication: <https://doi.org/doi:10.18129/B9.bioc.target¹⁵>

License: GPL-3

Author information

MA. Convinced the idea and wrote the draft of the manuscript. DK. Contributed to writing and revising the manuscript.

Acknowledgments

We thank all lab members for the discussion and comments on the early drafts of the article.

References

- Latchman DS: **Transcription factors: Bound to activate or repress.** *Trends Biochem Sci.* 2001; **26**(4): 211–3.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Johnson DS, Mortazavi A, Myers RM, et al.: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science.* 2007; **316**(5830): 1497–502.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ucar D, Beyer A, Parthasarathy S, et al.: **Predicting functionality of protein-DNA interactions by integrating diverse evidence.** *Bioinformatics.* 2009; **25**(12): i137–44.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tran LM, Brynildsen MP, Kao KC, et al.: **gNCA: A framework for determining transcription factor activity based on transcriptome: Identifiability and numerical implementation.** *Metab Eng.* 2005; **7**(2): 128–41.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Subramanian A, Tamayo P, Mootha VK, et al.: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A.* 2005; **102**(43): 15545–15550.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang S, Sun H, Ma J, et al.: **Target analysis by integration of transcriptome and ChIP-seq data with BETA.** *Nat Protoc.* 2013; **8**(12): 2502–15.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ahmed M, Min DS, Kim DR: **Integrating binding and expression data to predict transcription factors combined function.** *BMC Genomics.* 2020; **21**(1): 610.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Huber W, Carey VJ, Gentleman R, et al.: **Orchestrating high-throughput genomic analysis with Bioconductor.** *Nat Methods.* 2015; **12**(2): 115–121.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Feng C, Song C, Liu Y, et al.: **KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors.** *Nucleic Acids Res.* 2020; **48**(D1): D93–D100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Oki S, Ohta T, Shioi G, et al.: **ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data.** *EMBO Rep.* 2018; **19**(12): e46255.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chen L, Shioda T, Coser KR, et al.: **Genome-wide analysis of YY2 versus YY1 target genes.** *Nucleic Acids Res.* 2010; **38**(12): 4011–4026.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Michaud J, Praz V, Faresse VJ, et al.: **HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy.** *Genome Res.* 2013; **23**(6): 907–16.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wu XN, Shi TT, He YH, et al.: **Methylation of transcription factor YY2 regulates its transcriptional activity and cell proliferation.** *Cell Discov.* 2017; **3**: 17035.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ameur A, Rada-Iglesias A, Komorowski J, et al.: **Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP.** *Nucleic Acids Res.* 2009; **37**(12): e85.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ahmed M: **target: Predict Combined Function of Transcription Factors.** R package version 1.8.0, 2021.
<http://www.doi.org/10.18129/B9.bioc.target>

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 08 November 2021

<https://doi.org/10.5256/f1000research.55413.r97730>

© 2021 Ramos-Rodríguez M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Mireia Ramos-Rodríguez** 

Endocrine Regulatory Genomics, Department of Experimental & Health Sciences, Pompeu Fabra University, Barcelona, Spain

In this paper, Ahmed and Kim present the target R package, which implements the BETA algorithm and extends its functionality to predict combined targets and functions of two different transcription factors (TF). By using transcription factor binding data (ChIP-seq) and gene expression data when the TF is perturbed, they are able to predict the gene targets of a single or a pair of TFs.

This paper raised some major concerns that need to be addressed by the authors:

1. Regarding the actual package code, the distance calculation doesn't measure the actual distance between peaks and TSS. The code of your `find_distance()` function subtracts the peak center from the region center to obtain the distance between these two features. This is not the same as the distance from the peak to the TSS, which is what this variable should be measuring according to your text "The scores of individual peaks are decreasing function of the distance from the transcription start sites", your vignette "find_distance: calculate the distance between the peaks and the regions of interest, e.g. transcription start sites (TSS)." and the original method publication " Δ is the exact distance between a binding site and the TSS". An easy way to fix this is to provide different arguments for peak and regions "how", so the user can for example select `how_peaks="center"` and `how_regions="end"`. In the specific case of the code you show in this paper, those parameters would return the actual distances between peaks and TSS.
2. Related to the previous point, the parameter `downstream` in the function `promoters()` is set to 200 by default, so the width of the regions you are generating in the first chunk of code in page 6 actually have a width of 100,200bp. You should set this argument to 0 to actually obtain 100kb windows upstream of TSS.
3. YY1 and YY2 might not be the best examples to use for extracting conclusions on gene targets and the combined action of both TFs. Besides YY1 activity as a TF, it can also interact

with chromatin modifiers and direct them to specific regions of the genome ¹. It has also been identified as a structural factor that regulates the formation and DNA loops ². Thus, the changes in gene expression observed when perturbing this TF might not all be associated with its activity as a TF, which is the main focus of this package.

4. The section on the binding motif analysis is quite interesting in terms of what to do after performing the analysis with the R target package. However, I think it would be interesting to develop it a little bit more, maybe associate the sequences present in the different groups of regulated targets to actual transcription factors to see if there is a common regulatory pathway to these targets.
5. Regarding the general text, and specifically the section “Predicting gene targets of individual factors”, I feel that the description of the main package functionality is too technical and not very informative. The authors describe all the arguments that can be provided to the different functions and the object classes that come out of them, but this description is already in the package manual. Instead of talking about the arguments and object classes, I would briefly describe what they do and how they do it, so readers can easily follow the methods without the need to read the original BETA publication or the package vignette.
6. Related to point 3, the datasets used in this paper are different from the ones used in the vignettes and included within the package and I wasn’t able to find it on the GEO site either. I would recommend providing this data either within the package or in the docker image they already created. This would facilitate the reproduction of the results presented in this paper.
7. The authors keep referring to transcription factors as “factors”, which might induce confusion when reading the article. They write “Transcription Factors” (or the abbreviation TF) to differentiate them from the broad and diverse meanings of “factors”.
8. When revising the vignette I noticed that it’s missing the steps for preparing the data gene expression data, specifically the set to create windows upstream of TSS. When they load the gene expression object with (data(“real_transcripts”)) the windows are already present. This is misleading for users that are following the vignette as they might miss this specific step and they will not be able to get the correct results when reproducing it with their own data.
9. In page 5, the authors mention that the changes in expression resulting from separate KD of YY1 and YY2 are correlated, but they do not provide any statistical test to confirm this. They should at least perform a correlation test and show the p-value to make this affirmation.
10. There are some sentences in the text that are difficult to understand. The authors should rewrite them to ensure that the readers can follow the text. Some examples are:
 - Missing citations for “KnockTF” and “ChIP-Atlas” in p.4.
 - “We used the **USSC** hg19 human genome” [p. 4], should be “UCSC”.
 - “We first **locate** the files in the data/ directory” [p. 4], shouldn’t it be “save”?

- “They respond similarly to **their** perturbation of either factor” [p.5] should be “the perturbation”.
- “They need to be mapped to the **EN-TREZIDS**” [p.5] should be “ENTREZ IDs”
- “get the genomic coordinates for the transcripts and resize them to 100kb upstream” [p.5]. The authors should rephrase this, as they are not resizing the transcripts but rather generating 100kb windows upstream of their TSS.
- “we divide the targets by the direction of the effect of knock-down of the factor on the expression of the target” [p.6]. The authors should rephrase this sentence, as it is very long and difficult to follow.
- “The ECDF of the down-regulated of YY1 is higher than that [...]”. This whole paragraph it’s difficult to understand, the authors should rephrase it.

References

1. Wilkinson FH, Park K, Atchison ML: Polycomb recruitment to DNA in vivo by the YY1 REPO domain. *Proc Natl Acad Sci U S A*. 2006; **103** (51): 19296-301 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Weintraub AS, Li CH, Zamudio AV, Sigova AA, et al.: YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*. 2017; **171** (7): 1573-1588.e28 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics and Regulatory Genomics, Bioinformatics, R and Bioconductor.

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 09 Nov 2021

Mahmoud Ahmed, Gyeongsang National University School of Medicine, Jinju, South Korea

1. The solution referred to by the reviewer is already implemented in the `find_distance` function as an argument called `how` which defaults to 'center'. This is a link to the code (<https://github.com/MahShaaban/target/blob/9c6f869d794cfaff63310c5f79cea1e1095e2198/R/function>). I chose this as default since it would be neutral to the peak width. But of course, in a different use case, the user might be interested in the distance from the 'start' or 'end' of the peak.
2. We revised the text to state the correct downstream and upstream distances as pointed out by the reviewer.
3. We are aware of these facts. We would like to argue that using these transcription factors is still suitable. In fact, one of the main goals of the package is to distinguish between the direct targets of transcription factors and the ones that change indirectly, hence the reliance on both gene expression and peak binding. However, the definition of target here might be broadened to include cases like the ones pointed to by the reviewer.
4. The goal of this section, as correctly pointed out by the reviewer, is to give an example of further analysis of the target package output. There is a number of further analysis that could be performed using the found motifs, but it is beyond the purpose of this workflow article to get into.
5. We revised the section to briefly describe what each function does and how.
6. We added to the revised version of the manuscript a chunk of code to download the dataset.
7. We revised the text to use "transcription factor" instead of just "factor", when appropriate.
8. The data object `real_transcripts` is made up of the test data provided by BETA. The original files are too large to be included in the R package. The processing script is part of the package though, `inst/extdata/make-data.R``
9. We calculated the correlation coefficient for the fold-change of YY1 and 2 and added it to the text.
10. We revised the text to correct the errors and rephrase difficult sentences.

Competing Interests: No competing interests were disclosed.

Reviewer Report 18 October 2021

<https://doi.org/10.5256/f1000research.55413.r94695>

© 2021 Tian S et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Shulan Tian

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Yan Huihuang

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester,, MN, USA

Ahmed and Kim developed an R package to implement the binding and expression target analysis (BETA) package and extend the application to cases involving two transcription factors. The package predicts the potential target genes for binding sites from individual TFs or shared binding sites from two factors. There are some major concerns that need to be addressed.

1. p4: YY1 is a zinc finger protein that directs deacetylase and histone acetyltransferases of the promoters of many genes.

This is misleading. While YY1 binds preferentially to the promoter regions, it also binds to enhancers. YY1 binds to both and facilitates the structural interactions between regulatory elements [see Cell. 2017¹]. YY2 also binds to both promoters and enhancers [see Proc Natl Acad Sci U S A. 2016²]. It should be "directs histone deacetylase and acetyltransferases to .."

2. YY1 and YY2 have been well studied in terms of their regulatory roles. Both have dual activating and repressive roles in regulating target gene expression, with a lot of overlap between their binding sites. How about two factors both with overall activating roles? Are there public data available to better demonstrate the applications of this package?
3. The R code in the manuscript is obsolete. It will be great to make the coding workflow consistent with the one that is available in bioconductor. For example, I can't find the data files mentioned in the manuscript when the target package was installed from bioconductor. Clearly lay out the strength to implement an R version vs. the original python version [described by Wang *et al.* ³] will be helpful, like what the authors described in the bioconductor documentation. Also, emphasize the low- or high-level functions implemented here are beneficial to general users who don't have comprehensive programming background.
4. The authors illustrate how to identify the target genes for the shared binding sites between

two TFs. How about the gene targets of the factor-unique binding sites?

5. The authors set the maximum peak-to-gene TSS distance at 100kb, which is fine for the purpose of demonstration. Practically, the authors may want to provide recommendations or suggestions to the external users, since this is a very critical parameter. Based on the chromatin interaction data and co-accessibility data, peaks can target genes over a much larger distance. Alternatively, provide the option to incorporate topologically associating domains data, which will improve the detection of regulatory interactions.
6. p6: Because the ranks rather than the absolute value of the regulatory potential are used, the lower the value, the higher the potential

Based on the original paper of the BETA package [Figure 2], genes were ranked based on their regulatory potential score (from high to low), it should be "the lower the rank, the higher the regulatory potential"?, Please check

Similarly on P7: The scores of the individual peaks are a decreasing function of the distance from the transcription start sites— the closer the factor binding site from the start site, the lower the score.

Based on the original paper of the BETA package [Table], this should be "the closer the factor binding site to the start site, the higher the score"? Please check.

7. The authors need to check spelling and grammar more carefully and try to make it more readable. Below are some of the examples:

p3: Therefore, methods to determine which of these sites are true targets [should be true binding sites]

p3: A signed statistics (fold-change or t-statistics)

p4: YY2 is a parloge of YY1 [a paralog of YY1]

p4: This dataset was obtained in the form of differential expression between the two conditions from KnockTF [need citation]

p4: ChIP-Atlas, no citation

p4: USSC hg19 [UCSC hg19]

p5: Figure 1, The fold-change [knockdown/control?]

p5: EN-TREZIDS [Entrez IDs]

P5: resize? them to 100kb upstream from the transcription start site [extend to 100kb...]

p12: In Summary section: based on the distance of the ChIP peaks of the transcription factor in the genes and the differential expression of the factor perturbation.

based on the distances of the ChIP peaks of the transcription factor relative to the TSSs of the genes

two sets of data are used to find the shared peaks and the product of their differential expression

two sets of data are used to find the shared peaks and the rank product of their differential expression statistics?

Table 2. two-sided [two.sided]

Figure 2 legend:

Figure 2A: the same color was used to represent both YY1 and YY2 data.

Figure 2C should be Figure 2B. Figure 2D should be Figure 2C

YY1 and YY2 targets are shown at each regulatory potential value. the x-axis is the rank, not the regulatory potential value itself

the same for Figure 3, the x-axis is the rank of regulatory interaction

Figure 3 legend:

what are the transformed distances of the shared peaks? Need to explain whether it represents % of distance

occurrences [occurrences]

Figure 5 legend:

weight matrices, position weight matrices

seq logos, sequence logos

the letter occurs at that position, occurs at that position

References

1. Weintraub AS, Li CH, Zamudio AV, Sigova AA, et al.: YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*. 2017; **171** (7): 1573-1588.e28 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Tahmasebi S, Jafarnejad S, Tam I, Gonatopoulos-Pournatzis T, et al.: Control of embryonic stem cell self-renewal and differentiation via coordinated alternative splicing and translation of YY2. *Proceedings of the National Academy of Sciences*. 2016; **113** (44): 12360-12367 [Publisher Full Text](#)
3. Wang S, Sun H, Ma J, Zang C, et al.: Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat Protoc*. 2013; **8** (12): 2502-15 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, bioinformatics, epigenomics, data science, etc

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 28 Oct 2021

Mahmoud Ahmed, Gyeongsang National University School of Medicine, Jinju, South Korea

We would like to thank the reviewers for their effort and thoughtful comments. We addressed each point separately

1. We corrected the above sentence and added another sentence to mention that the YY1 binds to the enhancer regions of many of its targets.
2. Since this package aims to model the combined function of two factors from separate datasets, the data presented in the manuscript fit the goal well. That is, we want to see whether using data generated separately but in the same biological system we could suggest the true gene targets of the two factors. First by modeling the regulatory potential of each on the shared targets and then by considering the effect of their knockdown on the expression of the same targets.
3. The data we used in this article is available from Figshare (and not in the package). We added a chunk of code to the manuscript to download the data from the source. This workflow article focuses on the steps to perform the analysis enabled by the package. It is not intended to be a substitute for reading the package documentation for users who are interested in the low-level functions.
4. The code in section "Predicting gene targets of individual factors" does predict the targets of the individual transcription factors on their unique binding sites and the results are presented in Figure 2.
5. This model is based on the idea that the regulatory potential of a given factor decreases with the distance from the transcription start sites. It is not clear to use whether this holds at very large distances or for regulators other than transcription

factors. Therefore we used the distance recommended by the original paper and left the decision to the user to make depending on their case. Users can define their regions of interest in any way they like, for example, using TADs. Here, we used the simplest case of extracting TSSs and including stretches of the up and down streams.

6. We corrected the sentence referred to above.

7. We corrected the sentence referred to above and revised the manuscript for typos and grammatical errors.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research