STAT1 SCEPTRE vs Seurat

2023-04-03

Introduction

This is a followup analysis of Gene's IRF1-analysis-v2 writeup
The goal of this writeup is to take a closer look at the ChIP-seq enrichment analysis, focusing on IRF1 in monocytes. In particular, I examine the following questions (same as Gene's):

- How do people typically determine transcription factor targets?
- How do ChIP-seq scores for potential target genes align with those in the hTFtarget database?
- How do SCEPTRE and Seurat IRF1 discoveries align with each other?
- How do SCEPTRE and Seurat IRF1 discoveries align with ChIP-seq scores?

Determining transcription factor targets

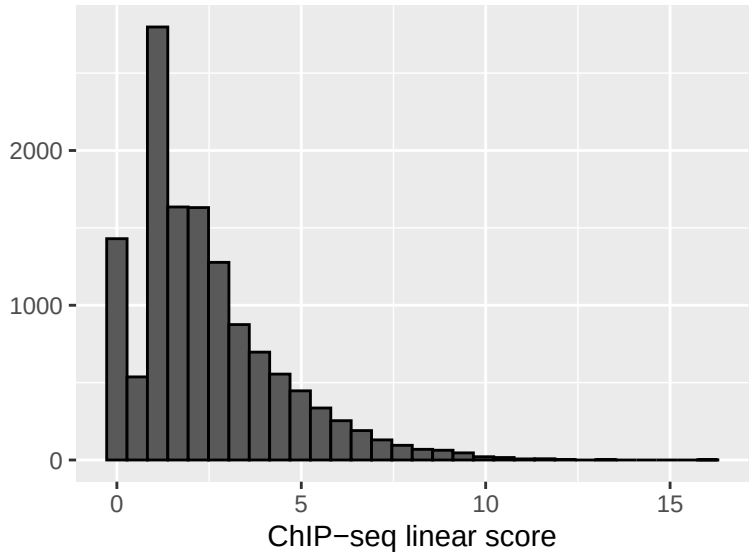
Sikora-Wohlfeld et al, 2013 discuss various computational methods to infer TF targets based on ChIP-seq data. This problem turns out to be not completely straightforward, with several competing methodologies available. The one we had been trying, which Sikora-Wohlfeld et al call the "binary" approach, is the most naive method. It has fairly poor performance. However, Sikora-Wohlfeld find that the best window width to use for this approach is 5kb, and they consider an interval of this width centered on the TSS rather than just upstream of it. An approach that performs better is called the "linear" approach. In this approach, the relative distances of ChIP-seq peaks to the TSS are summed, restricting attention to a 50kb window centered on the TSS. Another approach is TIP, which is the preferred approach of ENCODE. Unfortunately, the TIP results for ENCODE are available only for K562 and GM12878 cell types (probably without IFN-gamma stimulation), and the TIP software appears somewhat out of date. Hence, in this writeup, I will focus on the binary and linear scoring approaches outlined by Sikora-Wohlfeld et al.

Get ChIP-seq scores for each gene

```
# get binary score for each gene
window_width_binary <- 5e3
chipseq_scores_binary <- GRanges(
  seqnames = TSS_info$chr,
  ranges = IRanges(start = TSS_info$TSS-window_width_binary,
                   end = TSS_info$TSS+window_width_binary),
  gene = TSS_info$gene,
  TSS = TSS_info$TSS) |>
  join_overlap_left(chipseq_data) |>
  group_by(gene) |>
  summarise(binary_score = any(!is.na(score))) |>
  as_tibble()
```

```r
# get linear score for each gene
window_width_linear <- 5e4
chipseq_scores_linear <- GRanges(
  seqnames = TSS_info$chr,
  ranges = IRanges(start = TSS_info$TSS-window_width_linear,
                   end = TSS_info$TSS+window_width_linear),
  gene = TSS_info$gene,
  TSS = TSS_info$TSS) |>
  join_overlap_left(chipseq_data) |>
  group_by(gene) |>
  summarise(linear_score = sum((window_width_linear - abs(TSS - peak_position))/
                               window_width_linear)) |>
  as_tibble() |>
  mutate(linear_score = ifelse(is.na(linear_score), 0, linear_score))
```

Let's take a look at the distribution of the linear ChIP-seq scores.



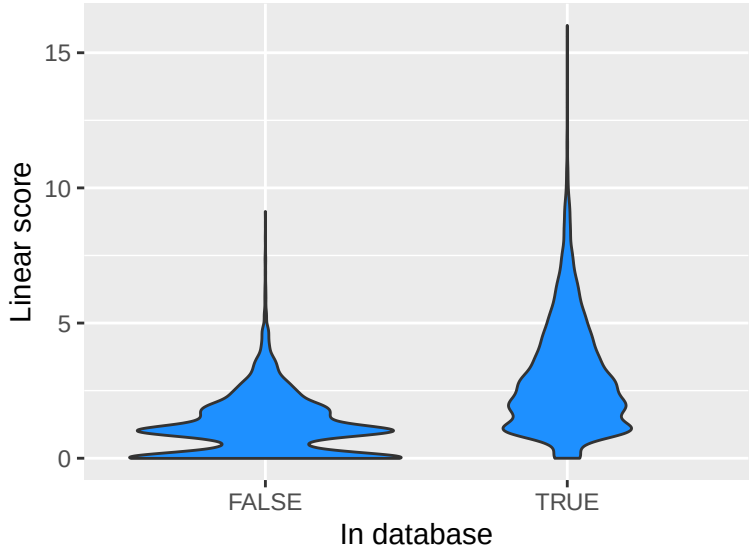
We see that there are modes at 0 (no peaks within the window width) and 1 (one peaks near the TSS). There is also a long right tail. Next, we check how these ChIP-seq scores align with the hTFtarget database.

Table 1: Comparing ChIP-seq binary score to database.

In database	Mean binary score
FALSE	0.52
TRUE	0.86

Table 2: Comparing ChIP-seq linear score to database.

In database	Median linear score
FALSE	1.00
TRUE	2.56



These results suggest that there is decent agreement between the hTFtarget database and the scores we derived from the ChIP-seq data.

Compare SCEPTRE and Seurat results with ChIP-seq scores

Let's first look at how the SCEPTRE and Seurat discoveries align with each other.

Table 3: Comparing numbers of discoveries made by SCEPTRE and Seurat.

SCEPTRE discovery	Seurat Discovery	Number
FALSE	FALSE	6128
FALSE	TRUE	1112
TRUE	FALSE	668
TRUE	TRUE	4033

This table suggests that SCEPTRE and Seurat results have decent, but highly imperfect, agreement. Next, let's look at how the SCEPTRE and Seurat discoveries align with the ChIP-seq signal (as measured by the hTFtarget database, the binary ChIP-seq scores, and the linear ChIP-seq scores).

Table 4: Comparing SCEPTRE discoveries to ChIP-seq binary scores.

In hTFtarget	Proportion SCEPTRE discoveries
FALSE	0.33
TRUE	0.42

Table 5: Comparing Seurat discoveries to ChIP-seq binary scores.

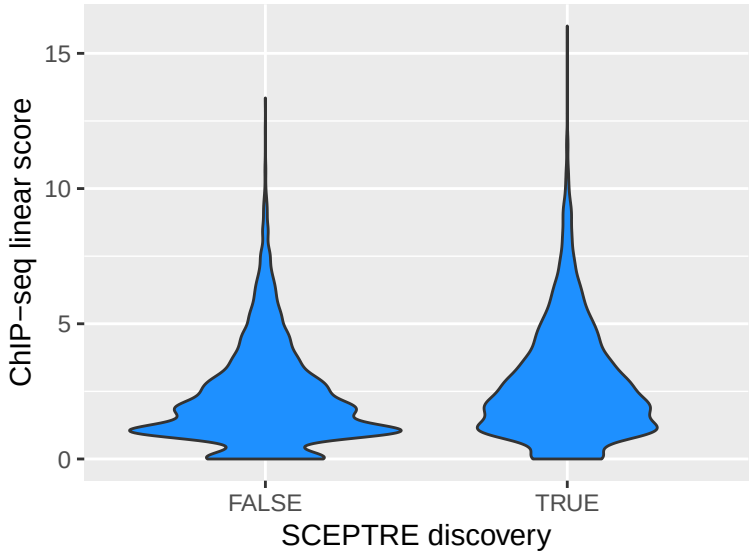
In hTFtarget	Proportion Seurat discoveries
FALSE	0.40
TRUE	0.44

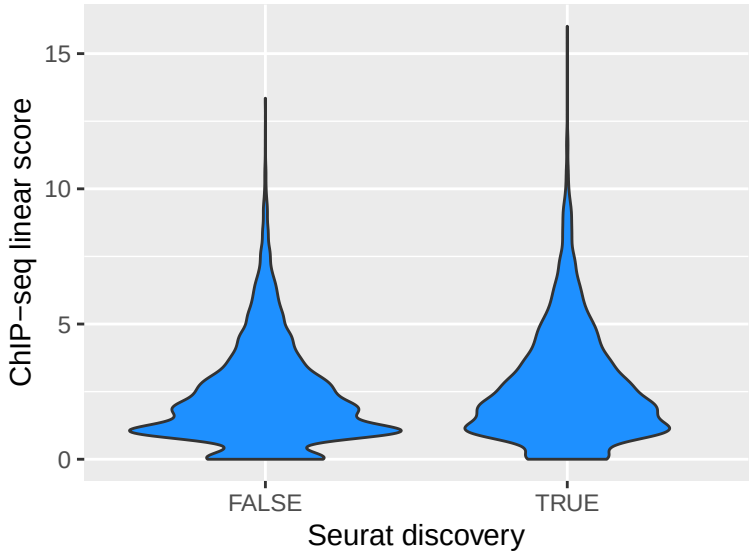
Table 6: Comparing SCEPTRE discoveries to ChIP-seq binary scores.

Binary ChIP-seq score	Proportion SCEPTRE discoveries
FALSE	0.34
TRUE	0.41

Table 7: Comparing Seurat discoveries to ChIP-seq binary scores.

Binary ChIP-seq score	Proportion Seurat discoveries
FALSE	0.39
TRUE	0.44





These results suggest that there is some enrichment of ChIP-seq signal in either the SCEPTRE or Seurat discoveries. However, the difference is not nearly as stark as in the htftarget database.

Conclusions

Below are the answers to the questions posed at the beginning:

- How do people typically determine transcription factor targets? **There isn't just one way of doing this. One lesson learned from the literature is that the window sizes used are larger than our original one (500bp) and are centered on the TSS rather than just upstream of it. Perhaps ChromHMM would help, but no one appears to have used this approach so perhaps neither should we.**
- How do ChIP-seq scores for potential target genes align with those in the hTFtarget database? **The alignment is pretty decent, suggesting that we're not making a big mistake processing the ChIP-seq data.**
- How do SCEPTRE and Seurat IRF1 discoveries align with each other? **The alignment is decent but very imperfect, as Kaishu has found before.**
- How do SCEPTRE and Seurat IRF1 discoveries align with ChIP-seq scores? **There appears to be some enrichment of ChIP-seq signal in the SCEPTRE and Seurat discoveries.**