# IRF1 SCEPTRE vs Seurat in monocytes

2023-04-04

## Introduction

This is a followup analysis of Gene's `IRF1-analysis-v2` writeup, with the main difference that the ChIP-seq data is from CD14+ monocytes rather than K562 is considered. Here, we consider two ways of determining transcription factor target genes: those that are in the hTFtarget database and those that have at least one ChIP-seq peak within 5kb of their TSS. These two align reasonably well; see the table below.

Table 1: Comparing target genes identified based on database and based directly on ChIP-seq data (at least one peak within 5kb).

| Database \ ChIP-seq | FALSE | TRUE |
|---|---|---|
| FALSE | 6787 | 740 |
| TRUE | 2727 | 2869 |

## 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 2: Comparing SCEPTRE versus Seurat discoveries.

| SCEPTRE \ Seurat | FALSE | TRUE |
|---|---|---|
| FALSE | 9935 | 422 |
| TRUE | 602 | 1819 |

This table suggests that SCEPTRE and Seurat results have decent, but imperfect, agreement. Next, let's look at how the SCEPTRE and Seurat discoveries align with the ChIP-seq target genes (as identified by either the hTFtarget database or directly from the ChIP-seq data).

Table 3: Comparing SCEPTRE to database.

| Database \ SCEPTRE | FALSE | TRUE | Prop |
|---|---|---|---|
| FALSE | 6260 | 1007 | 0.14 |
| TRUE | 4097 | 1414 | 0.26 |
| Prop | 0.4 | 0.58 | |

Table 4: Comparing Seurat to database.

| Database \ Seurat | FALSE | TRUE | Prop |
|---|---|---|---|
| FALSE | 6382 | 885 | 0.12 |
| TRUE | 4155 | 1356 | 0.25 |
| Prop | 0.39 | 0.61 | |

Table 5: Comparing SCEPTRE to ChIP-seq binary scores.

| ChIP-seq \ SCEPTRE | FALSE | TRUE | Prop |
|---|---|---|---|
| FALSE | 7755 | 1462 | 0.16 |
| TRUE | 2602 | 959 | 0.27 |
| Prop | 0.25 | 0.4 | |

Table 6: Comparing Seurat to ChIP-seq binary scores.

| ChIP-seq \ Seurat | FALSE | TRUE | Prop |
|---|---|---|---|
| FALSE | 7906 | 1311 | 0.14 |
| TRUE | 2631 | 930 | 0.26 |
| Prop | 0.25 | 0.41 | |

The proportions are the proportion of `TRUE` values in each row or column. For example, 0.58 of the genes found by SCEPTRE are marked as IRF1 targets in the database (i.e. SCEPTRE has specificity 0.58). From these tables, we see that **SCEPTRE has slightly higher sensitivity but slighly lower specificity than Seurat.** We can summarize each 2-by-2 table via its odds ratio (the p-values are all extremely small). The resulting odds ratios are shown below.

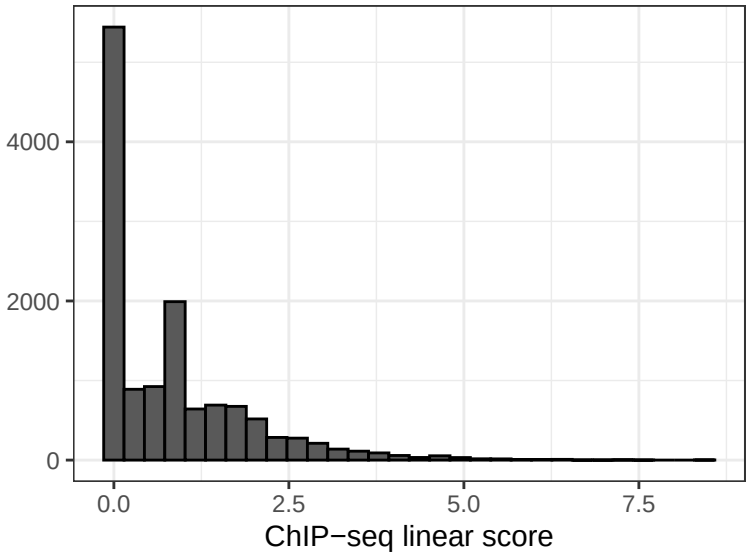
Table 7: Enrichment odds ratios, comparing to database and our ChIP-seq target assignments.

| Ground truth \ Method | SCEPTRE | Seurat |
|---|---|---|
| database | 2.15 | 2.35 |
| ChIP-seq | 1.95 | 2.13 |

These results suggest that **there is nontrivial (roughly two-fold) enrichment of ChIP-seq signal in both the SCEPTRE and Seurat discoveries, presumably because CD14+ cells are a better match to THP1 cells. Unfortunately, the Seurat discoveries have slightly higher enrichment, as measured by the odds ratio.** This result is somewhat contradictory to our results on control data, where SCEPTRE found fewer false positives and more true positives than Seurat.

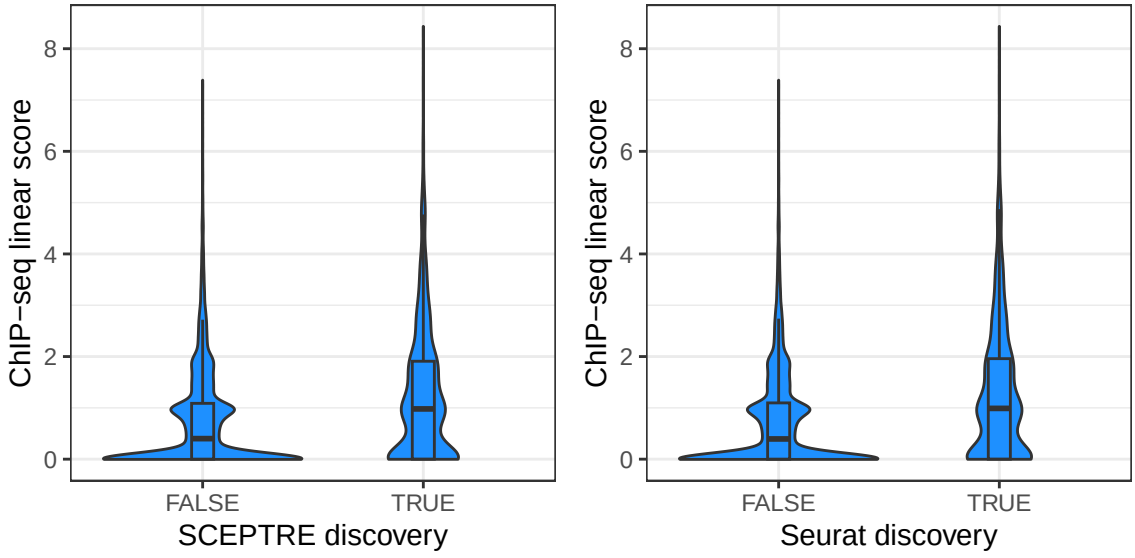
# Appendix

Another way of measuring the amount of ChIP-seq signal near a gene is the linear score proposed by Sikora-Wohlfeld et al, 2013. 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. Below is the distribution of the linear IRF1 ChIP-seq scores across genes:




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.

Now, let's see the distributions of these linear scores for genes detected by SCEPTRE and Seurat.




Again, we see some nontrivial enrichment for both SCEPTRE and Seurat, without a significant difference apparent between the two methods.