STAT1 SCEPTRE vs Seurat in monocytes, with same QC

2023-04-04

Introduction

This is a followup analysis of Gene's `IRF1-analysis-v2` writeup, with two main differences: **The ChIP-seq data is from CD14+ monocytes rather than K562 and the Seurat QC is matched to that of SCEPTRE.** 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	2448	2441
TRUE	1357	8259

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	6482	1316
TRUE	619	4228

This table suggests that SCEPTRE and Seurat results have decent, but imperfect, agreement. Also note that the total numbers of discoveries made by the two methods are nearly the same. 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	2622	1242	0.321
TRUE	5176	3605	0.411
Prop	0.664	0.744	

Table 4: Comparing Seurat to database.

Database \ Seurat	FALSE	TRUE	Prop
FALSE	2978	1603	0.35
TRUE	5349	4041	0.43
Prop	0.642	0.716	

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

ChIP-seq \ SCEPTRE	FALSE	TRUE	Prop
FALSE	1871	911	0.327
TRUE	5927	3936	0.399
Prop	0.76	0.812	

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

ChIP-seq \ Seurat	FALSE	TRUE	Prop
FALSE	2365	1144	0.326
TRUE	5962	4500	0.43
Prop	0.716	0.797	

The proportions are the proportion of `TRUE` values in each row or column. For example, 0.576 of the genes found by SCEPTRE are marked as IRF1 targets in the database (i.e. SCEPTRE has specificity 0.576). From these tables, we see that **SCEPTRE has slightly lower sensitivity and 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.

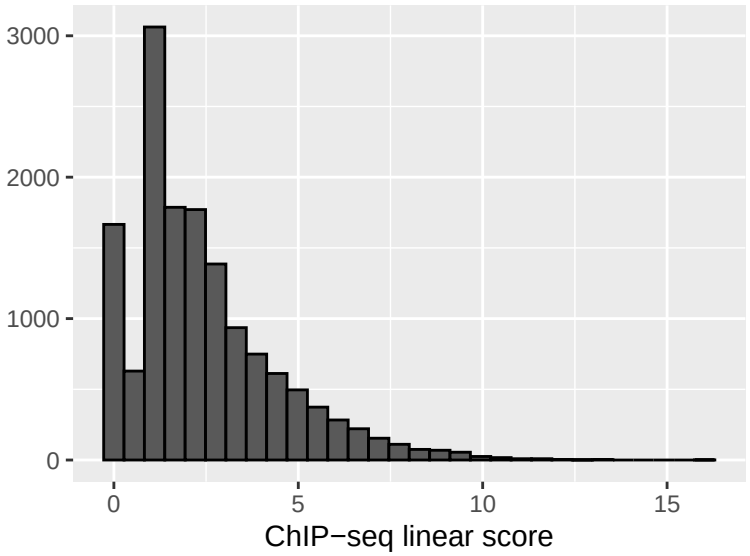
These results suggest that **there is nontrivial (roughly two-fold) enrichment of ChIP-seq signal in both the SCEPTRE and Seurat discoveries, presumably because CD14+ cells are a better match to THP1 cells. Unfortunately, the Seurat discoveries have slightly higher enrichment.** This result is somewhat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

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

Ground truth \ Method	SCEPTRE	Seurat
database	1.470	1.403
ChIP-seq	1.364	1.560

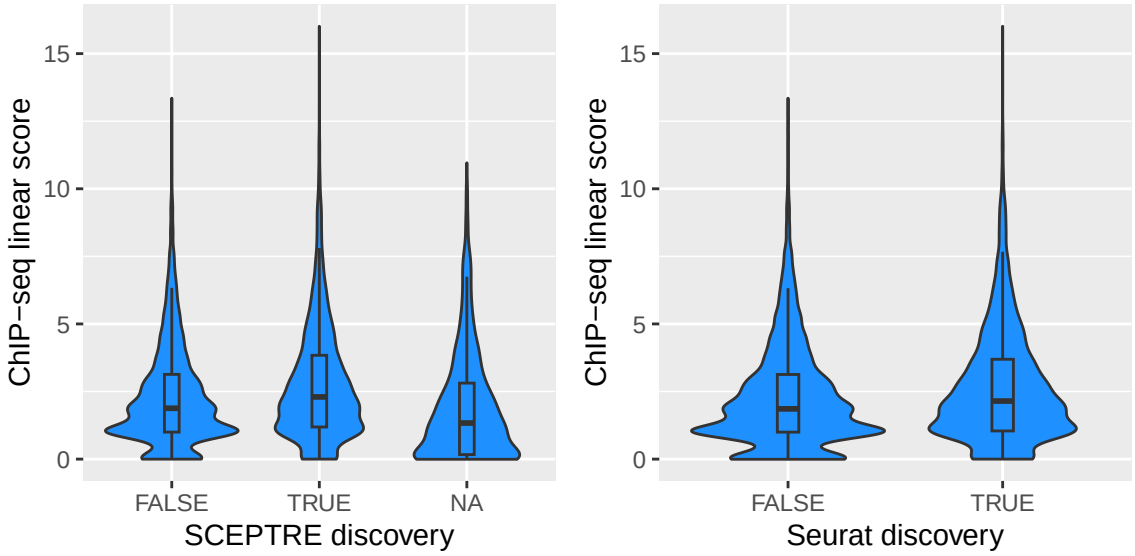
Appendix

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.