STAT1 SCEPTRE vs Seurat in monocytes, with same QC

2023-04-05

Introduction

This is a followup analysis of Gene's `IRF1-analysis-v3` writeup, this time for STAT1. 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	6424	1407
TRUE	596	4280

This table suggests that SCEPTRE and Seurat results have decent, but imperfect, agreement. Also note that Seurat makes roughly 800 more discoveries than SCEPTRE. 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	2640	1259	0.323
TRUE	5191	3617	0.411
Prop	0.663	0.742	

Table 4: Comparing Seurat to database.

Database \ Seurat	FALSE	TRUE	Prop
FALSE	2282	1617	0.415
TRUE	4738	4070	0.462
Prop	0.675	0.716	

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

ChIP-seq \ SCEPTRE	FALSE	TRUE	Prop
FALSE	1888	927	0.329
TRUE	5943	3949	0.399
Prop	0.759	0.81	

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

ChIP-seq \ Seurat	FALSE	TRUE	Prop
FALSE	1684	1131	0.402
TRUE	5336	4556	0.461
Prop	0.76	0.801	

The proportions are the proportion of `TRUE` values in each row or column. For example, 0.742 of the genes found by SCEPTRE are marked as STAT1 targets in the database (i.e. SCEPTRE has specificity 0.742). From these tables, we see that **SCEPTRE has slightly lower sensitivity and slightly higher specificity than Seurat.** We can summarize each 2-by-2 table via its p-value and odds ratio.

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

Ground truth \ Method	SCEPTRE	Seurat
database	1.461	1.212
ChIP-seq	1.353	1.271

Table 8: Enrichment p-values, comparing to database and our ChIP-seq target assignments.

Method \ Ground truth	SCEPTRE	Seurat
database	4.3×10^{-21}	7.3×10^{-7}
ChIP-seq	1.3×10^{-11}	3.0×10^{-8}

These results suggest that **there is moderate enrichment of ChIP-seq signal in both the SCEPTRE and Seurat discoveries, presumably because CD14+ cells are a better match to THP1 cells. The enrichment odds ratios for SCEPTRE are modestly higher than those of 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:
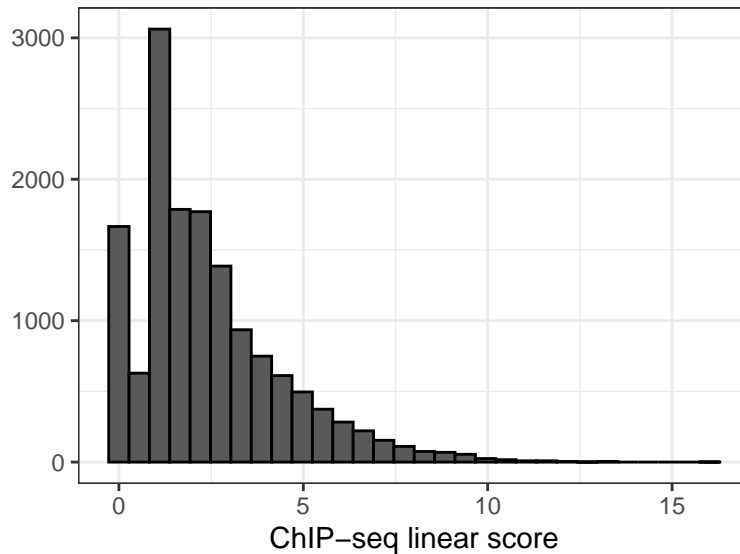
We see that there are modes at 0 (no peaks within the window width) and 1 (one peaks near the TSS). There is also a long right tail.

Now, let's see the distributions of these linear scores for genes detected by SCEPTRE and Seurat.
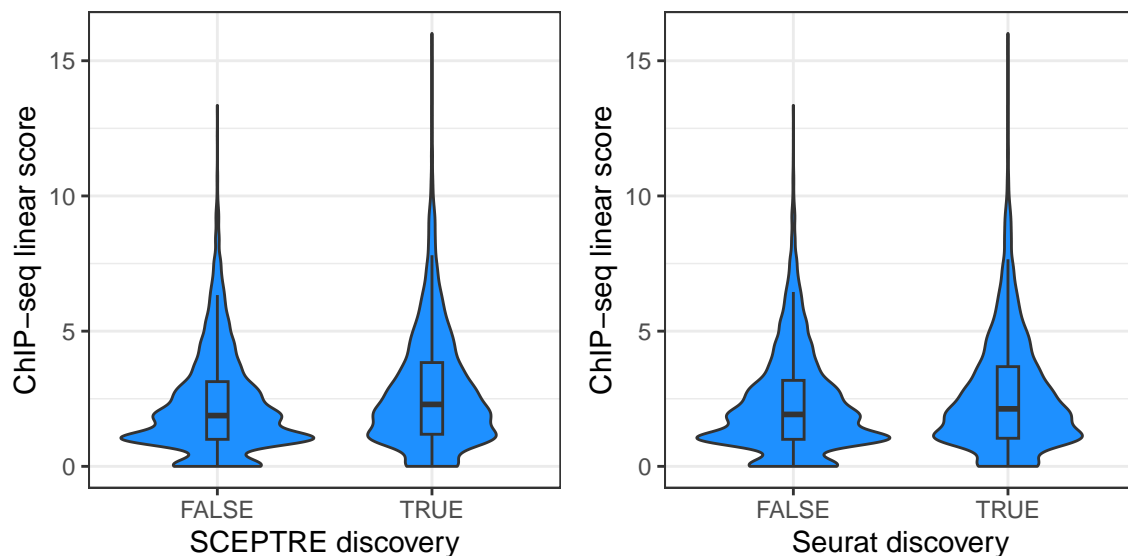
Again, we see some nontrivial enrichment for both SCEPTRE and Seurat, without a significant difference apparent between the two methods.