ImmGen TF Score Cutoff

Caleb Lareau

2017-01-16

The Question

There are 30,218,042 peak-motif pairs in `fimo_cisBP_jmotifs_p4.txt` using `1e-04` as a p-value cutoff. Using a slightly more stringent threshold of `1e-05`, we yield 3,552,978 peak-motif pairs. Does our choice of cutoff affect analysis/interpretation?

Analysis strategy

- Import TF binding hits from `fimo_cisBP_jmotifs_p4.txt` into the `chromVAR` analysis environment.
- Per sample/per motif/per threshold, compute deviation z-scores for comparison

Weak motif associations

Of the 981 unique motifs in our data base, 164 only have association p-values p_a such that $10e-04 > p_a > 10e-05$. In other words, these 164 motifs have exclusively weak associations and thus would be eliminated (TF scores are `NA`) if we used a more stringent threshold. Below, I show the extreme values of the TF-sample scores for $10e-04 > p_a$ (remember, these scores are `NA` if using a more stringent threshold)–

```
M6353_1.02  lib_7   77.31294
M6353_1.02  lib_216 73.25797
M5273_1.02  lib_10  71.92814
M5273_1.02  lib_207 70.64803
M5273_1.02  lib_13  69.6737
M6353_1.02  lib_142 68.33025

M5273_1.02  lib_133 -55.00627
M5273_1.02  lib_135 -51.01178
M5273_1.02  lib_134 -49.59398
M5273_1.02  lib_148 -48.29713
M5273_1.02  lib_175 -46.00564
M5273_1.02  lib_136 -45.72158
```

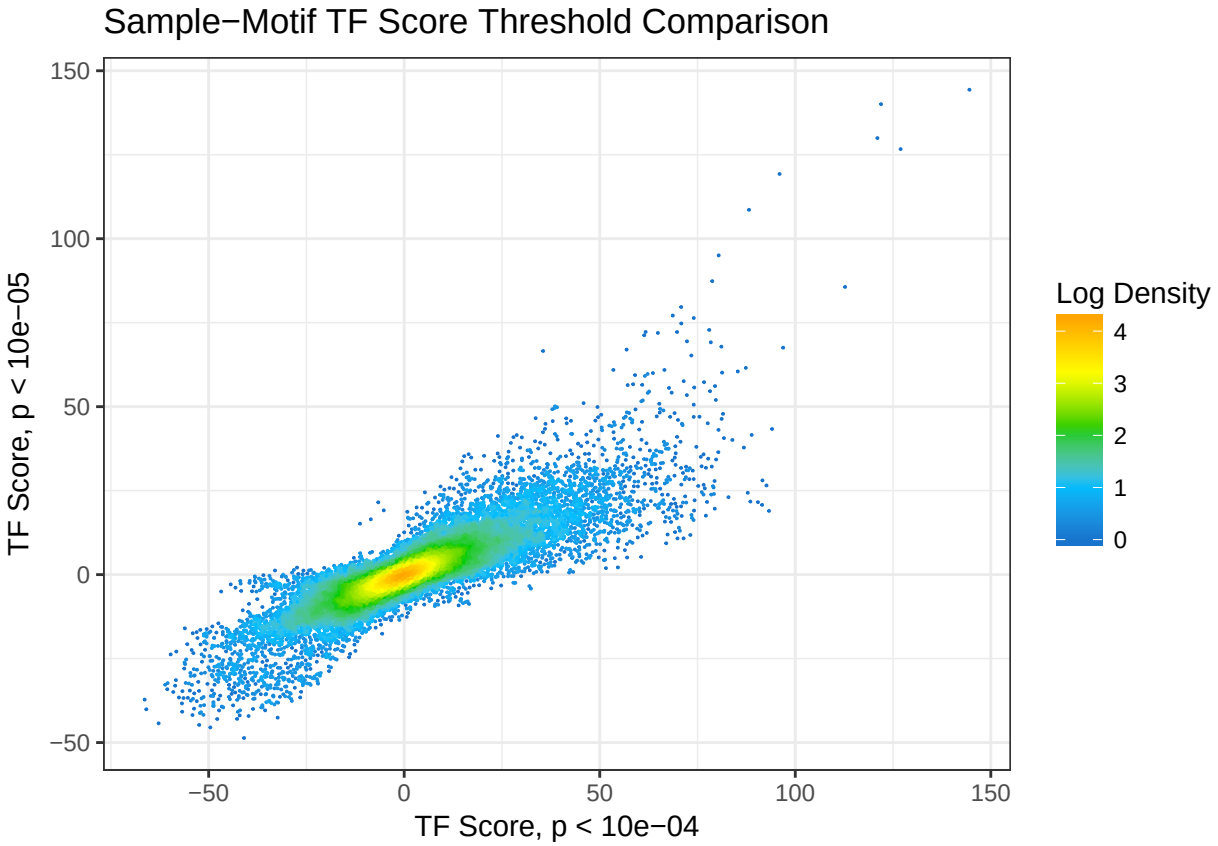
So, seemingly `M5273_1.02` explains a lot of variability in the data based on the extremes. This motif maps to *Snai2*, which has been shown to function in hematopoietic progenitor biology (source). One can verify that this motif has only association p-values p_a in the specified range $10e-04 > p_a > 10e-05$ (sanity check).

```
grep M5273_1.02 fimo_cisBP_jmotifs_p4.txt
```

From this relatively obvious example, it appears that for some motifs, the choice of cutoff does matter, and we may be missing out on relevant biology by selecting a more stringent threshold.

Large-scale comparison

Of the 827 motifs that do have at least 1 association p-value p_a such that $p_a < 10e-05$, we see a very nice correlation between the scores as shown in the plot below.



Each point is a sample-motif pair with TF scores at different binding significance p-values as noted on the axis labels. The log density of the points is denoted in the color.

Conclusion

Using the more liberal threshold for the peak-motif associations seemingly is the way to go. We'll recover more motifs, and the scores are largely invariant when we use a more stringent threshold when at least 1 motif-association p-value is $p_a < 10e - 05$.