Supplementary Information for BayesPeak - An R package for analysing ChIP-seq data

Jonathan Cairns, Christiana Spyrou, Rory Stark, Mike Smith, Andy Lynch, Simon Tavaré

Contents

1 The NRSF/REST ChIP-seq data set

We used the data set described in Johnson et al. (2007), in which ChIP-seq was performed in Jurkat cells with an antibody specific to NRSF/REST. We then called peaks in the data, as is described in section 3.

The NRSF/REST protein has a well-characterized binding site motif that is known as NRSE. Previously, Mortazavi et al. (2006) performed a motif analysis in which they took a sample of 113 potential binding sites, chosen to represent a mixture of high-scoring and low-scoring for known NRSF/REST binding site motifs. The individual sites were subsequently validated using ChIP-qPCR with a site being defined as enriched if it exhibited a qPCR fold enrichment of greater than 2.44. (This threshold was chosen to be 3 standard deviations above the average of 5 negative controls.) Of the 113 sites assessed, 83 were found to be enriched and 30 were found not to be enriched.

The data used have possible limitations in the context of a peak-caller comparison. Since the sites chosen for validation in Mortazavi et al. (2006) were found by motif analysis, there is a potential bias, as we cannot account for binding sites without this motif present - these sites may have qualitatively different signatures in the data. This is not a significant problem in the context of NRSF/REST, as it is known to be highly specific to its binding site, but it has implications when generalizing these results to other transcription factors.

Additionally, the sites that returned negative for ChIP-qPCR were originally selected because they had lower motif scores, rather than because they are often incorrectly called as peaks. As a result, there is no reason to suspect that a given site will be called in ChIP-seq data (as a given motif need not be functional) and therefore it is difficult to draw firm conclusions about the specificity of peak-callers based on the absence of these peaks.

The analysis also depends on the somewhat arbitrary choice of threshold used to call true binding sites in the qPCR data.

However, the overall results give a good indication of the sensitivity of each peak-caller, and some indication of the comparative specificities based on the sites that are incorrectly called.

2 *BayesPeak* - over-fitting correction

Peak-caller	Correction applied?	Peaks called	Validated peaks (out of 83)	Invalid regions (out of 30)	Median call width (bp)
BayesPeak	No	11243	72	2	250
BayesPeak	Yes	3011	68	2	300

Table 1: The effect of applying an over-fitting correction to *BayesPeak*'s results. Performing the over-fitting correction greatly decreased the number of calls made with only a small associated decrease in sensitivity. As a result, the enrichment for validated peaks is far greater in the post-correction results. The particular correction that we used is described in Supplementary Figure 1.

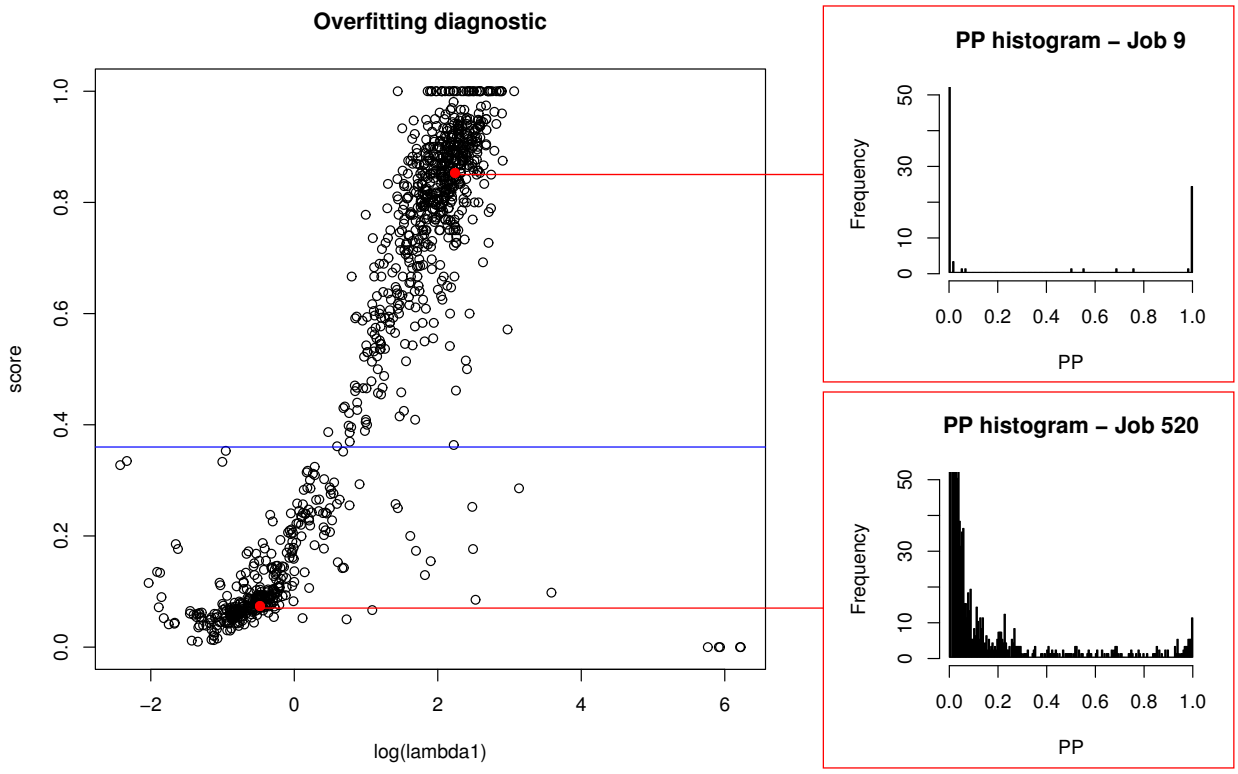

Figure 1: The threshold used in the over-fitting correction method, as applied to raw output from the NRSF data set. Each point corresponds to a job. The X-axis represents the logarithm of a particular job's average λ_1 parameter (where this parameter quantifies the mean increase in read count of an enriched bin compared to an unenriched bin). The Y-axis represents the "score" of a job, defined as the proportion of "potentially enriched" bins in that job with $PP > 0.5$ (where a bin is defined as "potentially enriched" if it has $PP > 0.01$). Two clear clusters of points are present in the job QC data, and thus we designate points in the region $score < 0.36$ (i.e. below the blue line) as representing over-fit jobs. The insets show histograms of the PP-values associated with each of two jobs, as a demonstration of the qualitative difference between the two clusters. We can see that job 9, in the cluster that is not over-fit, displays strong certainty about its peaks, as most of the PP-values are 0 or 1. On the other hand, job 520, in the over-fit cluster, displays more uncertainty about its results, as there is a great number of peaks with PP-values between 0 and 1. A full explanation of the over-fitting correction method can be found in the *BayesPeak* package vignette.

3 Peak-caller performance comparison

We compared the results of *BayesPeak* with those of 3 other peak-callers, each derived from analysis of the NRSF data set. In each case, the package's tutorial or vignette was followed, with default settings used otherwise. For *BayesPeak*, we applied an over-fitting correction as previously described in Supplementary Figure 1. For *PICS*, an FDR of 0.01 was applied, as described in the package vignette.

Peak-caller	Version	Peaks called	Validated peaks (out of 83)	Invalidated regions (out of 30)	Median call width (bp)
BayesPeak	1.1.3	3011	68	2	300
MACS	1.3.7.1	6486	74	3	447
PICS	1.0.6	4657	29	3	300
CSAR	1.1.0	1009	30	0	339
BayesPeak (high confidence)	1.1.3	1019	48	0	600

Table 2: Comparison between peak-caller results.

BayesPeak obtained a similar sensitivity to *MACS* whilst calling under half the number of peaks, suggesting that the enrichment for validated peaks in *BayesPeak*'s output will be much greater. Additionally, *BayesPeak* called 2 false positives as opposed to 3 false positives from *MACS*, mildly indicating a greater specificity from *BayesPeak*. The results from *CSAR* included fewer calls, more valid peaks and fewer invalid regions compared to those from *PICS*, with the former exhibiting a low number of calls and a high enrichment for validated peaks compared to *BayesPeak* and *MACS*.

However, the stringent results derived from *CSAR* were obtained without applying a cut-off to the scores associated with each call. We give an example of a high-confidence peak set that can be obtained from *BayesPeak*'s output, where we applied the filter $1 - PP < 10^{-12}$ to the set of calls obtained after applying the overfitting correction. This threshold was intentionally chosen to give a set of around 1000 peaks. This set of peaks shows a greater number of validated peaks than are present in *CSAR*'s output.

Note that filtering by PP-value increases the median length of *BayesPeak*'s calls from 300 to 600. This occurs because our method of summarizing PP-values is biased towards large regions of sustained enrichment. Intuitively, such a region has shown stronger evidence of binding site enrichment than a shorter region with similar PP values has.

4 Overlap between peak-callers

To assess the overlap between the results returned by each peak-caller, we took the union of all of the regions called by each of the peak-callers to form a set of "meta-calls". Each of these meta-calls was defined as being present in a particular peak-caller's output if that peak-caller returned a region within the meta-call. We plot the results as a Venn diagram in Figure 2.

As a result of combining the calls in this way, the numbers displayed in the figure do not add up to the exact numbers reported in Table 2. This analysis will not be robust to some unusual situations - for example, there may be instances where multiple nearby regions called by one peak-caller are merged into a single peak because another peak-caller cannot find the individual sub-peaks precisely and therefore calls the entire region as a single peak. However, we do obtain an indication of the agreement on enriched regions that is present between the different peak-callers. *BayesPeak* called very few regions that were not identified by at least one of the other methods. *CSAR*'s calls were a subset of those called by *MACS*.

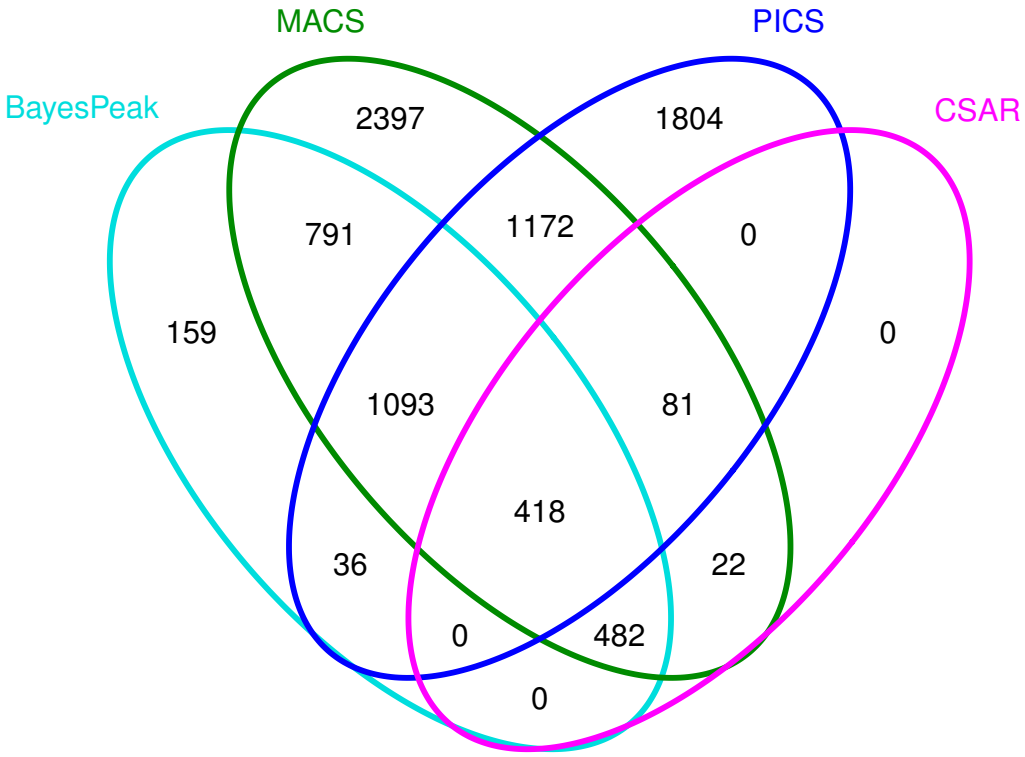

Figure 2: Venn diagram representing the overlap between the regions called by each algorithm.

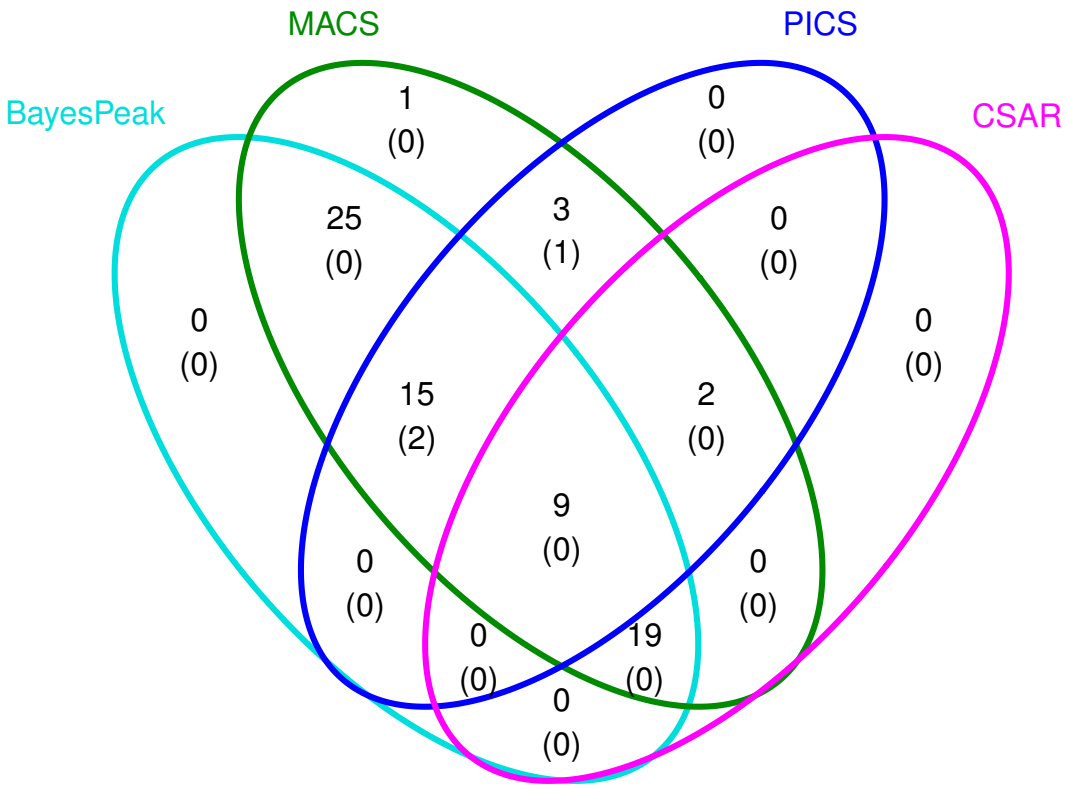

Figure 3: Venn diagram representing the presence of validated peaks or invalid regions in each algorithm's results. The numbers in brackets correspond to the number of invalid regions called.

5 R code used to obtain *BayesPeak*'s results

```
##directory contains the two files NRSF.treatment.bed and NRSF.control.bed.
library(BayesPeak)
library(multicore)
NRSF.bp <- bayespeak("NRSF.treatment.bed", "NRSF.control.bed",
use.multicore = TRUE, mc.cores = 8)

##no over-fitting correction
NRSF.bp.calls <- summarize.peaks(NRSF.bp)
nrow(NRSF.bp.calls)

##over-fitting correction
sel <- NRSF.bp$QC$score < 0.36
NRSF.bp.calls.OF <- summarize.peaks(NRSF.bp, exclude.jobs = sel)
nrow(NRSF.bp.calls.OF)

##further PP-value correction
sel.PP <- log(1 - NRSF.bp.calls.OF$PP, 10) < -12
NRSF.bp.calls.stringent <- NRSF.bp.calls.OF[sel.PP,]
nrow(NRSF.bp.calls.stringent)
```

References

David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316:1497–502, June 2007. ISSN 1095-9203. doi: 10.1126/science.1141319. URL `http://www.ncbi.nlm.nih.gov/pubmed/17540862`.

Ali Mortazavi et al. Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. *Genome Research*, 16:1208–1221, 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/16963704`.