A scalable machine learning pipeline for joint peak calling reveals differences between cell and experiment types

Toby Dylan Hocking and Guillaume Bourque

April 11, 2019

1 Results

2 Time and memory requirements

PeakSegPipeline uses algorithms with log-linear time complexity in the number of samples and base pairs. The PeakSegFPOP algorithm for peak prediction in a single sample with n coverage data (lines in the bedGraph file) uses $O(\log n)$ memory, $O(n \log n)$ disk space, and $O(n \log n)$ time [Hocking et al., 2017]. In practice we assign each job 10GB RAM and 24 hours of compute time. The total pipeline takes several hours or days to run (depending on the number of labels and samples).

PeakSegPipeline analyzes the data at single base pair resolution, and has no arbitrary bin or window size parameters.

JAMM ran out of memory (over 20GB) when analyzing the tcell group, which had the most samples (15 H3K36me3 tcell samples, 19 H3K4me3 tcell samples). So for the tcell group we only used five samples, which took 16GB of RAM and 11-16 hours of compute time.

References

T. Hocking, G. Rigaill, P. Fearnhead, and G. Bourque. A log-linear time algorithm for constrained changepoint detection. arXiv:1703.03352, 2017.

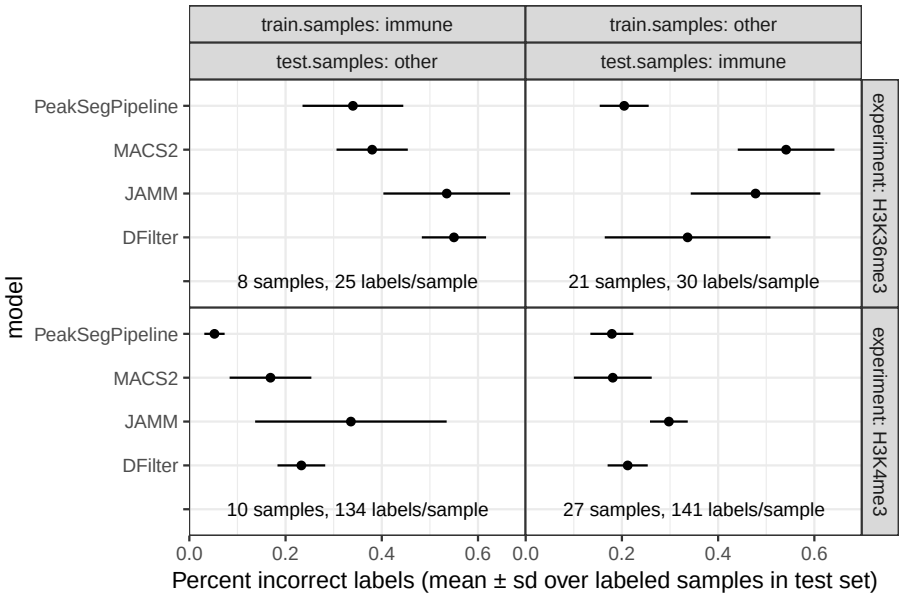

Figure 1: Comparison of peak detection error rate in two histone modification ChIP-seq experiments (panels from top to bottom) and two sample groups (panels from left to right). MACS, JAMM, and DFilter were run with parameters indicated by the authors for either sharp H3K4me3 or broad H3K36me3 data; parameters of PeakSegPipeline were learned using the train samples, and error rates were computed using the test samples. It is clear that the error rate of PeakSegPipeline is competitive with the other tools, and sometimes significantly lower.

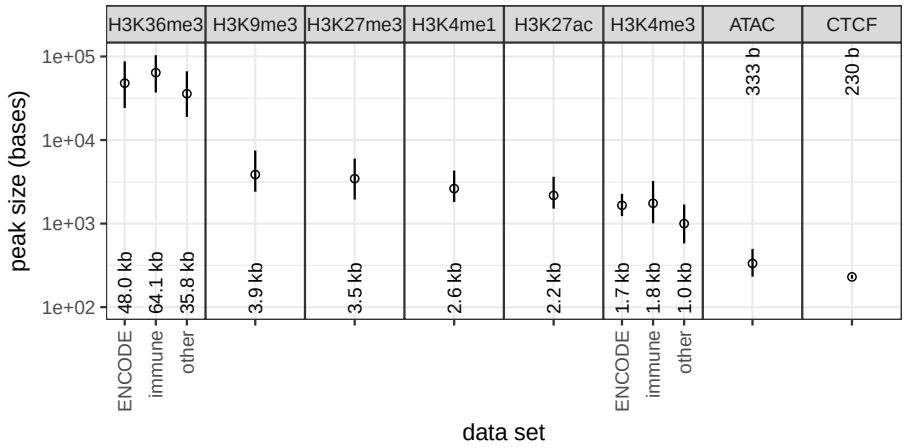

Figure 2: Peak size varies with epigenome experiment type. PeakSegPipeline was trained on each data set, and used to predict peaks for each data set. It is clear that there are several distinct categories of peak sizes: large 30-50 kb (H3K36me3), medium 1–5 kb (H3K9me3, H3K27me3, H3K4me1, H3K27ac, H3K4me3), and small 200–400 b (ATAC, CTCF).

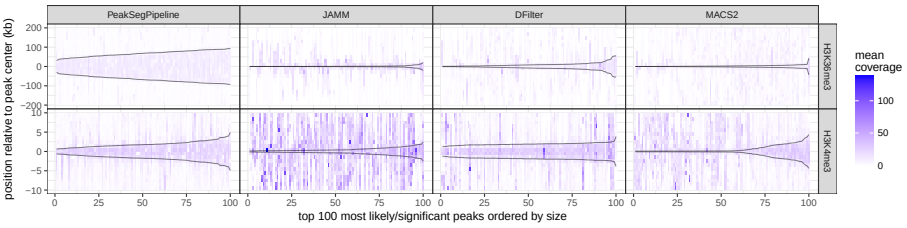

Figure 3: PeakSegPipeline detects peaks which are more well-defined relative to background than other tools (relative position of peak border shown in grey lines). The top 100 most likely/significant peaks are shown for each tool (panels from left to right) for one monocyte sample. **Top**: broad H3K36me3 data, for which the most likely peaks are $\approx$ 100 kb for PeakSegPipeline, but much smaller for the other tools. **Bottom**: sharp H3K4me3 data, for which PeakSegPipeline and DFilter recover peaks with relatively large signal, but JAMM and MACS2 detect many noisy regions.