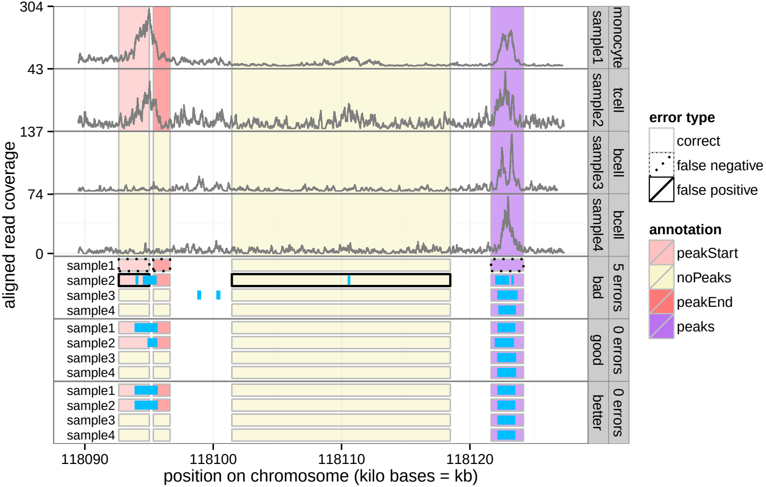
Machine learning algorithms for simultaneous supervised detection of peaks in multiple samples and cell types

Toby Dylan Hocking, toby.hocking@nau.edu
School of Informatics, Computing, and Cyber Systems at
Northern Arizona University
and
Guillaume Bourque
Department of Human Genetics, McGill University

January 4, 2020

Peak detectors should predict differences between samples

Context: detecting presence/absence of peaks (active regions) in epigenomic data profiles such as ChIP-seq, ATAC-seq, ...

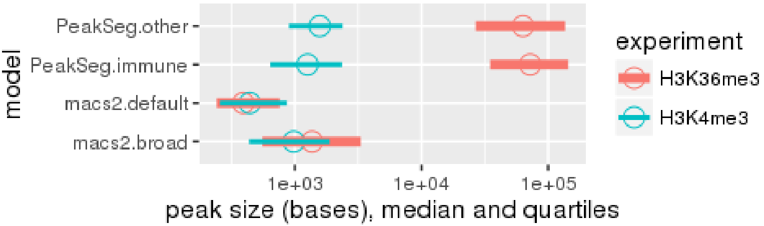


Algorithm and pipeline with two main novel ideas

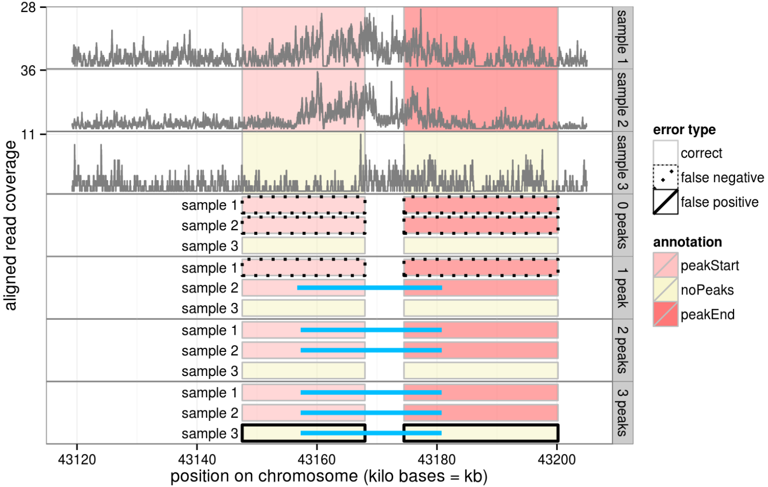
Supervised machine learning. Labels that indicate presence/absence of peaks in specific samples/regions are used to train model parameters (users do not need to know how to tune p-value thresholds, bin size parameters, etc).

Joint model predicts differences between any number of samples/groups. Unlike previous methods, not limited to one or two groups with replicates.

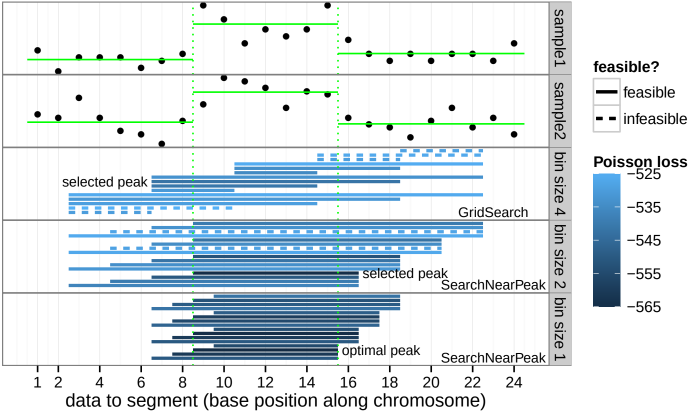
Results: state-of-the-art peak models with predicted sizes that are consistent with biological expectation.



Joint peak model for 3 labeled samples



Approximate algorithm for computing optimal peak boundaries for piecewise constant Poisson mean model

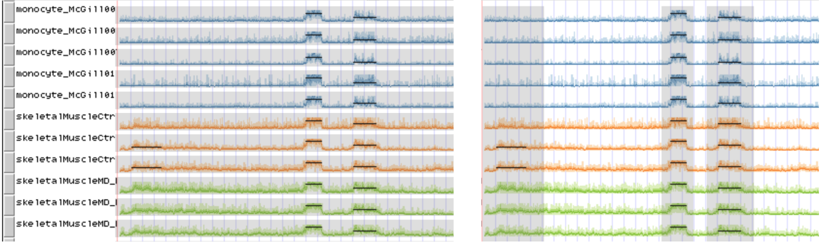


R package **PeakSegJoint**.

Two steps of the proposed peak prediction pipeline

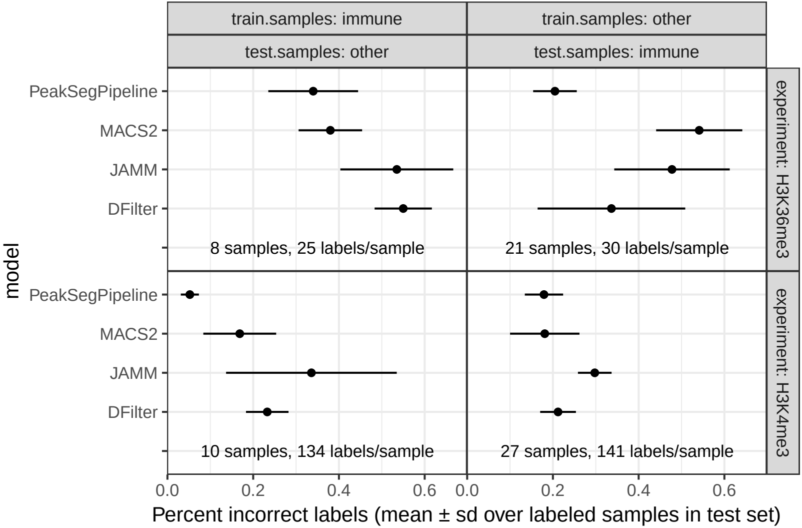
Preliminary peak predictions for each sample/contig

Joint prediction of presence/absence of a peak
with common boundaries in each sample/group,
for every genomic region with preliminary peaks



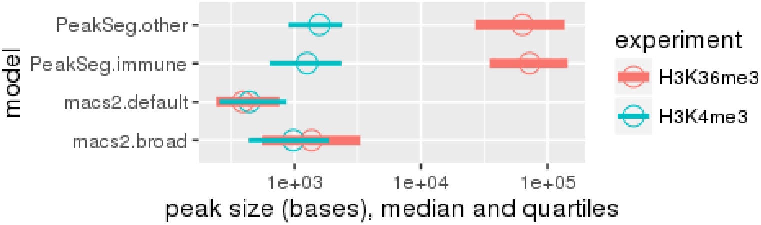
R package **PeakSegPipeline**.

Test error comparable to or better than baselines

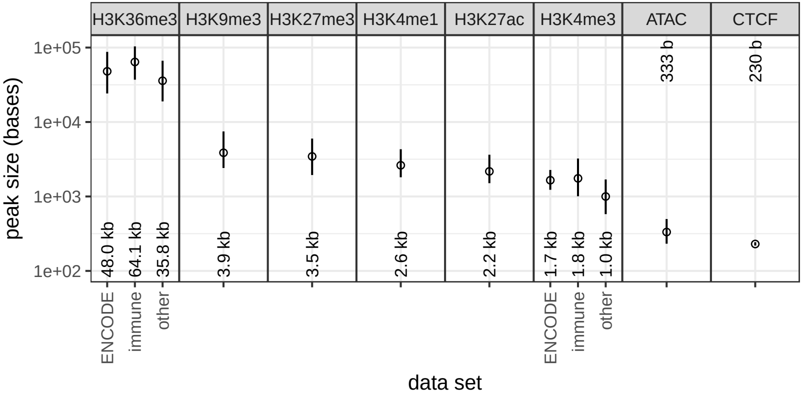


PeakSeg predicted sizes depend on experiment, macs2 baseline sizes depends on parameter settings

- Train PeakSeg models on either labeled samples of immune cell types (B cells, T cells, monocytes) or other cell types.
- Use macs2 baseline with either default parameter settings or broad command line argument.
- PeakSeg model has learned that H3K4me3 has smaller peaks than H3K36me3.



Predicted peak sizes vary by experiment type



Conclusions

- New algorithm for joint peak detection, R package **PeakSegJoint**.
- New pipeline for supervised joint analysis, R package **PeakSegPipeline**.
- State-of-the-art peak prediction accuracy and interpretable peak sizes.
- Recruiting graduate students for research projects! **toby.hocking@nau.edu**