Filtering STARR-Seq Peaks for Enhancers with Sequence Models

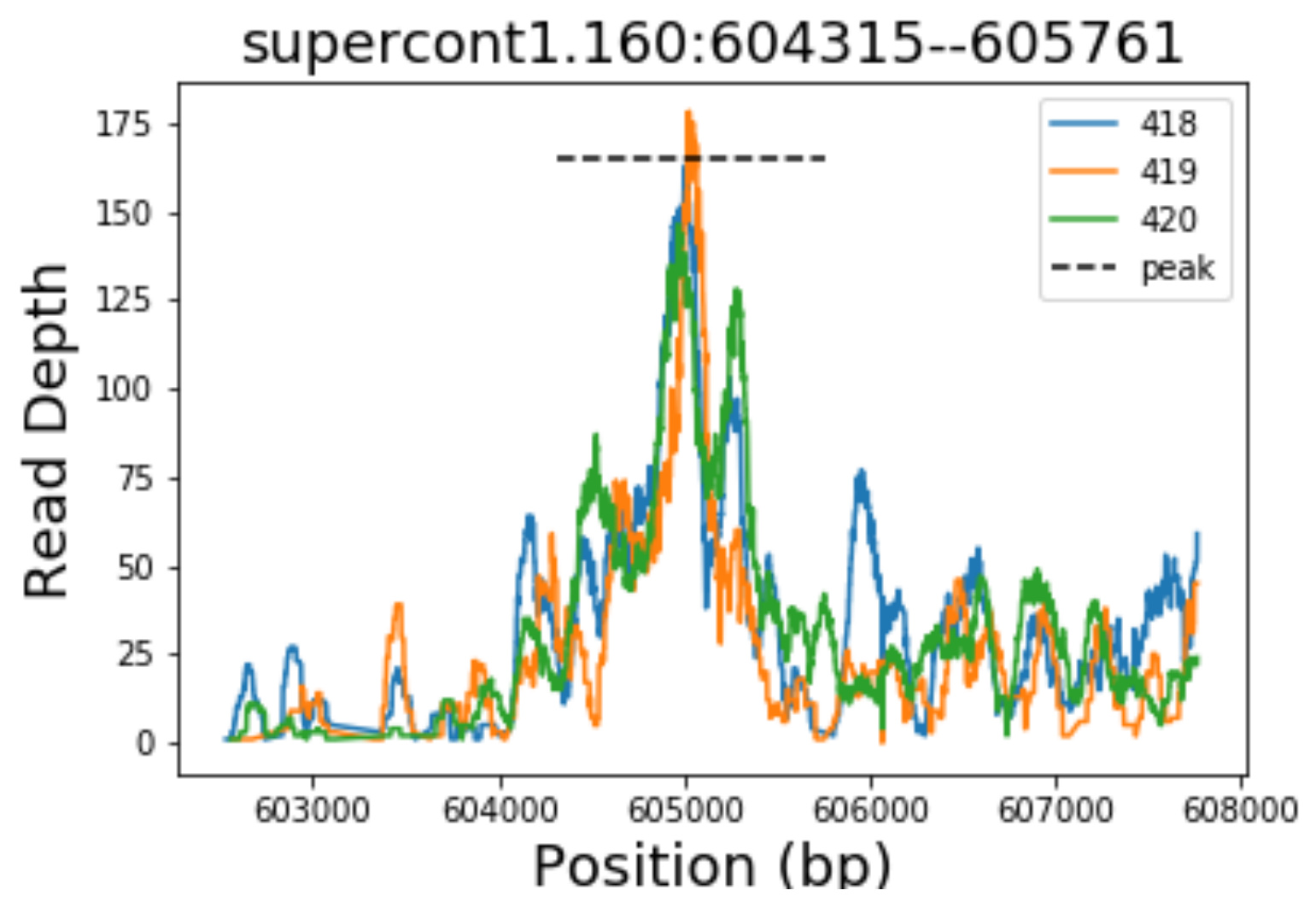
RJ Nowling[1], RR Geromel[1], and BS Halligan[1]

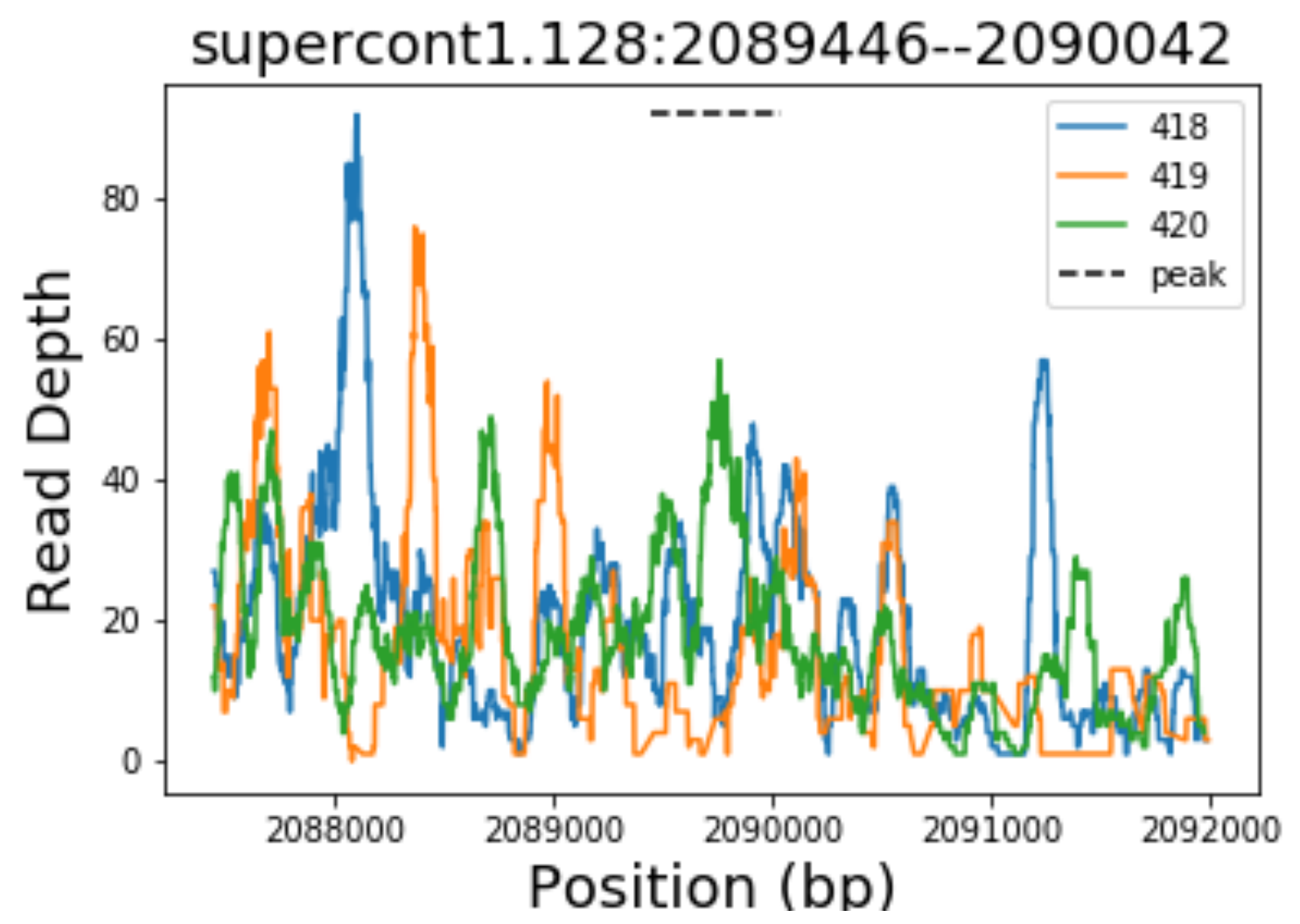
[1]Electrical Engineering and Computer Science, Milwaukee School of Engineering

Problem

STARR-Seq is a high-throughput technique for directly identifying genomic regions with enhancer activity [1]. Genomic DNA is sheared, inserted into artificial plasmids designed so that DNA with enhancer activity trigger self-transcription, and transfected into culture cells. The resulting RNA is converted back into cDNA, sequenced, and aligned to a reference genome. “Peaks” are called by comparing observed read depth at each point to an expected read depth from control DNA using a statistical test. Examples of peak calling methods based on read depth include MACS2 [4], basicSTARRSeq, and STARRPeaker [3].

It is challenging to accurately distinguish between real peaks and artifacts in regions where mean read count is low, but the variance is high (see figure below).





Our Hypothesis

We hypothesize that we can employ additional information about the genome to increase recall of called peaks. Enhancer activity is strongly correlated with sequence content. We focus on using sequence-derived features in machine learning models as an alternative to filtering peaks based on read counts alone.

Acknowledgments

We gratefully acknowledge Michelle Riehle for thoughtful discussions that provided the motivation for this effort and Scott Emrich for feedback on our results.

References

[1] C D Arnold, D Gerlach, D Spies, J A Matts, Y A Sytnikova, M Pagani, N C Lau, and A Stark. 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat. Genet*. 46, 7 (July 2014), 685–692.

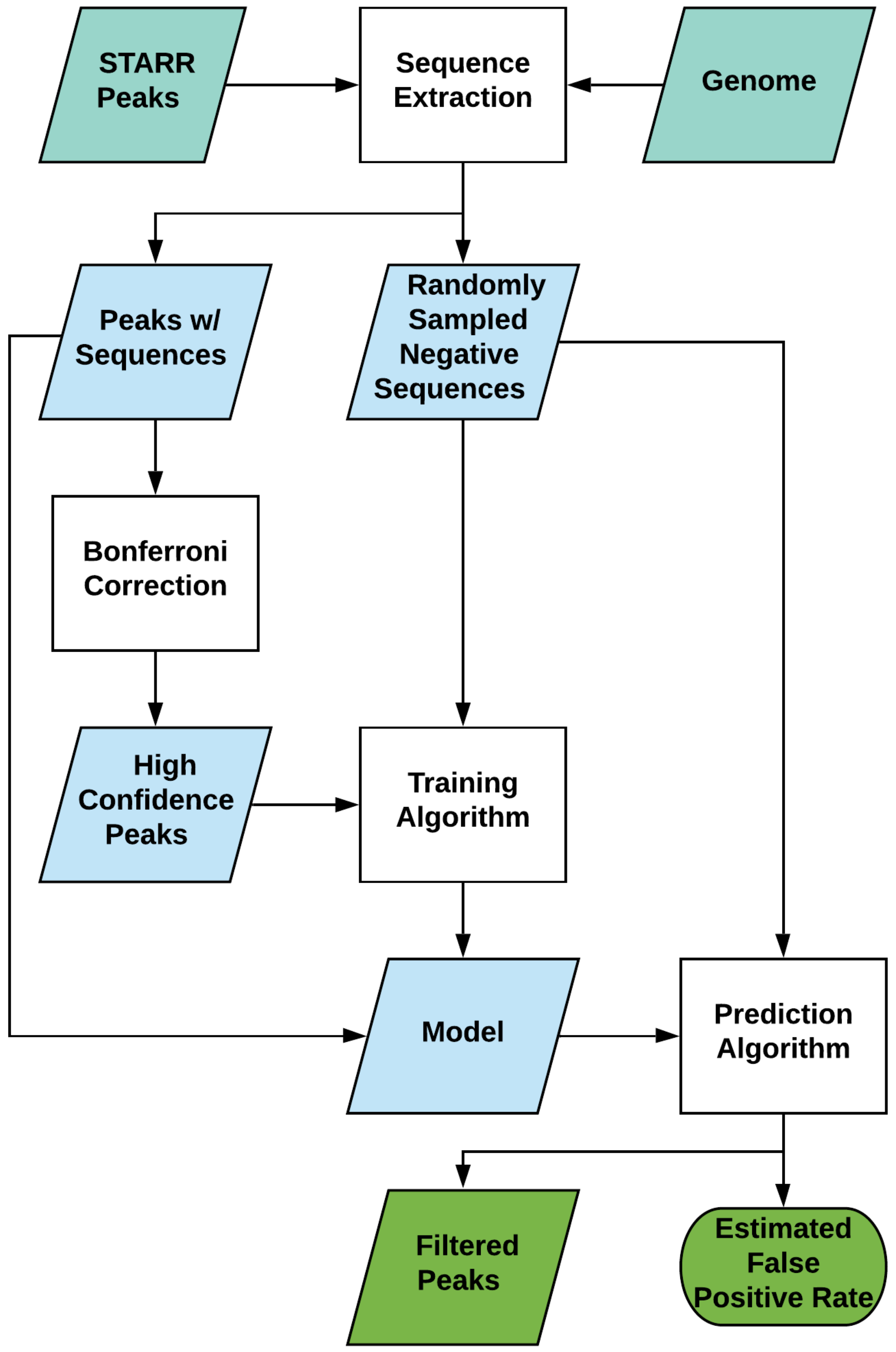
[2] K Davie, J Jacobs, M Atkins, D Potier, V Christiaens, G Halder, and S Aerts. 2015. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS Genet*. 11, 2 (Feb. 2015), e1004994.

[3] D Lee, M Shi, J Moran, M Wall, J Zhang, J Liu, D Fitzgerald, Y Kyono, L Ma, K P White, and M Gerstein. 2020. STARRPeaker: Uniform processing and accurate identification of STARR-seq active regions. *BiorXiv*. (May 2020).

[4] Y Zhang, T Liu, C A Meyer, J Eeckhoute, D S Johnson, B E Bernstein, C Nusbaum, R M Myers, M Brown, W Li, and X S Liu. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 9, 9 (Sept. 2008), R137.

Our Method

We propose using sequence-based machine learning models in a semi-supervised framework to filter peaks (see figure below). The Bonferroni correction is applied to p-values from statistical tests on read depth to identify a "high-confidence" set of training peaks. The sequences for those peaks along with an equal number of randomly-sampled negatives are used to train a ML model (an ensemble of 50 logistic regression models with k-mer count features (k=3 to 9)). The model is then applied to filtering all of the input peaks by sequence and an equal number of negative sequences to estimate false positive rates.



STARR Peaks

Sequence Extraction

Genome

Peaks w/ Sequences

Randomly Sampled Negative Sequences

Bonferroni Correction

High Confidence Peaks

Training Algorithm

Model

Prediction Algorithm

Filtered Peaks

Estimated False Positive Rate

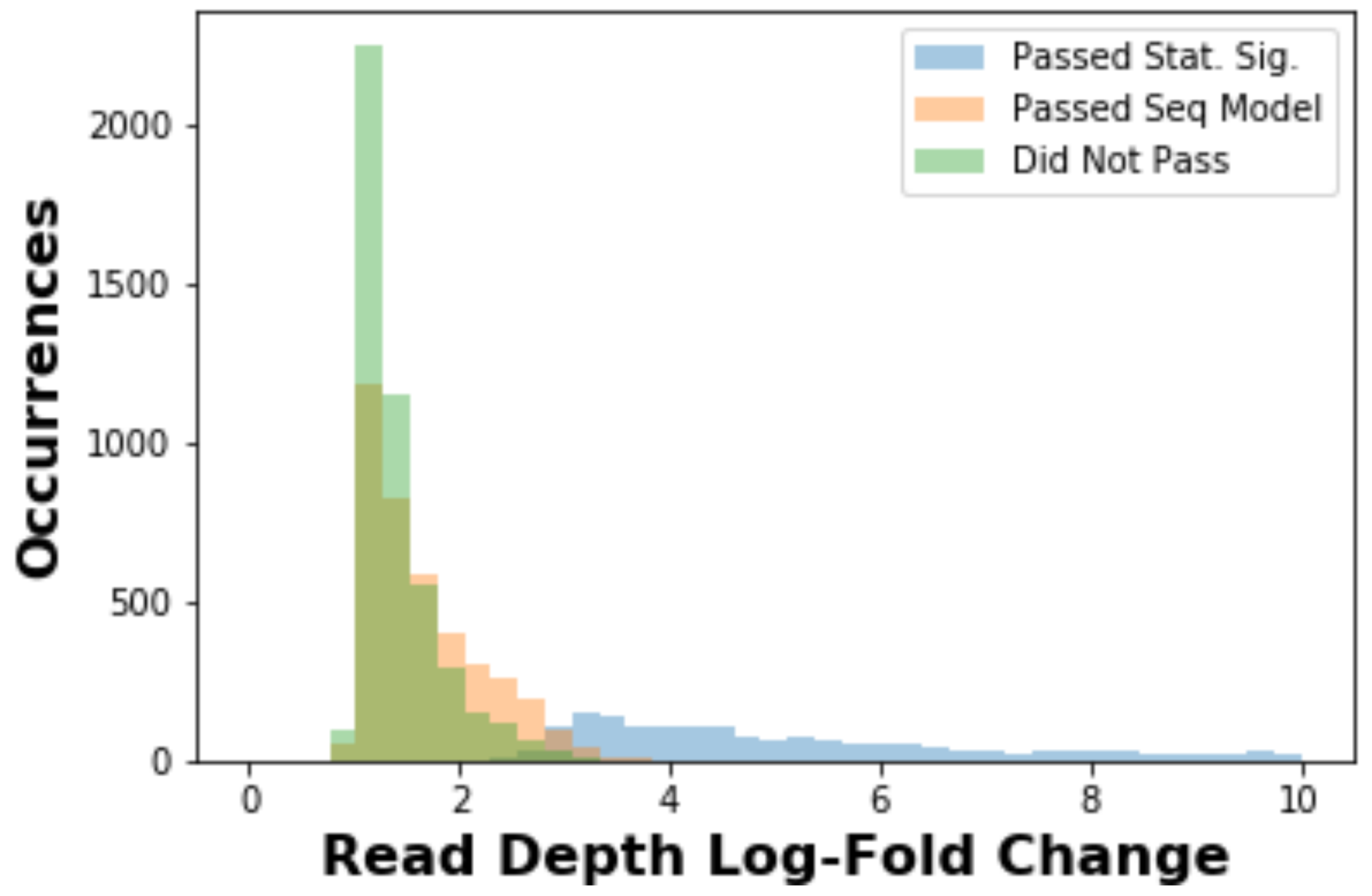
Sequence Model Doubles Called Peaks

501-bp sequences centered on the 10,996 STARR peaks from [1] were extracted from the *Drosophila melanogaster* dm3 genome. Randomly-sampled 501-bp sequences were used as a negative set. 2,214 training peaks were selected by the Bonferroni correction (α = 0.05).

5,801 STARR peaks passed the self-trained, sequenced-based model filter. 2,197 (or 99.2%) of the 2,214 "high-confidence" peaks were recalled. The remaining 3,604 "medium-confidence" peaks were new candidates identified by sequence.

Peak Read Depths

We plotted histograms of the read depth log-fold change for the three sets of peaks (high, medium, and low confidence) (see Figure 1). The distributions for the medium- and low-confidence peaks overlapped significantly. The sequence-based model identified enhancer candidates that would otherwise be filtered out using read depth alone.



Validation against FAIRE Peaks

We called peaks for the 4 *D. melanogaster* FAIRE-Seq data sets from [2]. FAIRE-Seq identified regions of chromatin that are open (accessible) in the tissue of interest under the experimental conditions. Active enhancers are located in open chromatin regions. While we not all of the predicted enhancers will overlap with the FAIRE peaks, we do expect to see a general trend.

FAIRE sequencing data were cleaned with Trimmomatic, aligned to the dm3 genome with bwa backtrack, and filtered for mapping quality (q < 10) with samtools. Alignments were pooled and 21,814 peaks were called by MACS2.

We compared the overlaps of the STARR peaks with the FAIRE peaks (see table below). The sequence model offered a compromise in recall and precision between the Bonferroni-corrected and the uncorrected peak lists. The sequence model more than doubled the recall compared with the Bonferroni correction at the cost of some precision.

STARR Peak Confidence	Peaks	FAIRE Recall	FAIRE Precision
High	2,214	5.7%	54.2%
High + Medium	5,801	12.8%	46.2%
All	10,996	18.2%	35.4%