Filtering STARR-Seq Peaks for Enhancers with Sequence Models

Ronald J. Nowling
Milwaukee School of Engineering
Milwaukee, Wisconsin
nowling@msoe.edu

Rafael Reple Geromel
Milwaukee School of Engineering
Milwaukee, Wisconsin
geromelrr@msoe.edu

Benjamin Halligan
Milwaukee School of Engineering
Milwaukee, Wisconsin
halliganbs@msoe.edu

ABSTRACT

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 depth is low but the variance is high. Fortunately, enhancer activity is strongly correlated with sequence content. We propose using sequence-based machine learning models in a semi-supervised framework to filter peaks. 501-bp sequences centered on the ≈11k STARR peaks from [1] were extracted from the *Drosophila melanogaster* dm3 genome. Randomly-sampled 501-bp sequences were used as a negative set. Peaks were filtered using a Bonferroni-corrected significance value ($\alpha = 0.05$) to create a "high-confidence" subset of ≈2.2k peaks. A Logistic Regression model with k-mer count features was trained on the high-confidence peak sequences and their negatives and used to classifying the remaining ≈8.8k peak sequences. The self-trained, sequenced-based model identified an additional ≈3.7k candidate enhancers ("medium confidence"). The remaining ≈5k STARR peaks were considered "low confidence" peaks.

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.

We called peaks for the 4 *D. melanogaster* FAIRE-Seq data sets from [2]. Sequencing data were cleaned with Trimmomatic, aligned to the dm3 genome with bwa backtrack, and filtered for mapping quality (q < 10) with samtools. MACS2 called ≈61k FAIRE peaks. The STARR peaks overlapped with the FAIRE peaks with precisions of 52.7% (high-confidence peaks), 40.6% (medium-confidence peaks), and 22.5% (low-confidence peaks).

ACM-BCB'20, September 21–24, 2020, Virtual

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

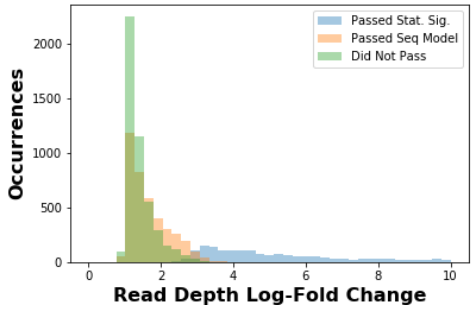

Figure 1: Read Depth Log-Fold Changes for High-, Medium-, and Low-confidence STARR Peaks

CCS CONCEPTS

• **Applied computing** → **Computational genomics**; **Bioinformatics**; **Recognition of genes and regulatory elements**.

KEYWORDS

STARR-Seq, peak calling, machine learning, k-mers

ACM Reference Format:
Ronald J. Nowling, Rafael Reple Geromel, and Benjamin Halligan. 2020. Filtering STARR-Seq Peaks for Enhancers with Sequence Models. In *ACM-BCB'20: 11th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, September 21–24, 2020, Virtual*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/1122445.1122456

ACKNOWLEDGMENTS

We would like to thank Dr. Michelle Riehle and Dr. Scott Emrich for insightful discussions.

REFERENCES

[1] Cosmas D Arnold, Daniel Gerlach, Daniel Spies, Jessica A Matts, Yuliya A Sytnikova, Michaela Pagani, Nelson C Lau, and Alexander 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] Kristofer Davie, Jelle Jacobs, Mardelle Atkins, Delphine Potier, Valerie Christiaens, Georg Halder, and Stein 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] Donghoon Lee, Manman Shi, Jennifer Moran, Martha Wall, Jing Zhang, Jason Liu, Dominic Fitzgerald, Yasuhiro Kyono, Lijia Ma, Kevin P White, and Mark Gerstein. 2020. STARRPeaker: Uniform processing and accurate identification of STARR-seq active regions. (May 2020), 694869 pages.

[4] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, 9 (Sept. 2008), R137.