PeakMatcher: Matching Peaks Across Genome Assemblies

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

Christopher R. Beal
Marquette University
Milwaukee, Wisconsin, USA
christopher.beal@marquette.edu

Scott Emrich
University of Tennessee–Knoxville
Knoxville, Tennessee, USA
semrich@utk.edu

Susanta K. Behura
University of Missouri
Columbia, Missouri, USA
behuras@missouri.edu

Marc S. Halfon
SUNY at Buffalo
Buffalo, New York, USA
mshalfon@buffalo.edu

Molly Duman-Scheel
Indiana University School of Medicine
South Bend, Indiana, USA
mscheel@nd.edu

ABSTRACT

When reference genome assemblies are updated, the peaks from DNA enrichment assays such as ChIP-Seq and FAIRE-Seq need to be called again using the new genome assembly. PeakMatcher is an open-source package that aids in validation by matching peaks across two genome assemblies using the alignment of reads or within the same genome. PeakMatcher calculates recall and precision while also outputting lists of peak-to-peak matches.

PeakMatcher uses read alignments to match peaks across genome assemblies. PeakMatcher finds all read aligned to one genome that overlap with a given list of peaks. PeakMatcher uses the read names to locate where those reads are aligned against a second genome. Lastly, all peaks called against the second genome that overlap with the aligned reads are found and output. PeakMatcher groups uses the peak-read-peak relationships to discover 1-to-1, 1-to-many, and many-to-many relationships. Overlap queries are performed with interval trees for maximum efficiency.

We evaluated PeakMatcher on two data sets. The first data set was FAIRE-Seq (Formaldehyde-Assisted Isolation of Regulatory Elements Sequencing) of DNA isolated embyros of the mosquito *Aedes aegypti* [2, 4]. We implemented a peak calling pipeline and validated it on the older (highly fragmented) AaegL3 assembly [5]. PeakMatcher matched 92.9% (precision) of the 121,594 previously-called peaks from [2, 4] with 89.4% (recall) of the 124,959 peaks called with our new pipeline. Next, we applied the peak-calling pipeline to call FAIRE peaks using the newer, chromosome-complete AaegL5 assembly [3]. PeakMatcher found matches for 14 of the 16 experimentally-validated AaegL3 FAIRE peaks from [2, 4]. We validated the matches by comparing nearby genes across the genomes. Nearby genes were consistent for 11 of the 14 peaks; inconsistencies for at least two of the remaining peaks were clearly attributable to differences in assemblies. When applied to all of the peaks, PeakMatcher matched 78.8% (precision) of the 124,959 AaegL3 peaks with 76.7% (recall) of the 128,307 AaegL5 peaks.

The second data set was STARR-Seq (Self-Transcribing Active Regulatory Region Sequencing) of *Drosophila melanogaster* DNA in S2 culture cells [1]. We called STARR peaks against two versions (dm3 and r5.53) of the *D. melanogaster* genome [6]. PeakMatcher matched 77.4% (precision) of the 4,195 dm3 peaks with 94.8% (recall) of the 3,114 r5.53 peaks.

PeakMatcher and associated documentation are available on GitHub (https://github.com/rnowling/peak-matcher) under the open-source Apache Software License v2. PeakMatcher was written in Python 3 using the `intervaltree` library.

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

CCS CONCEPTS

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

KEYWORDS

peak calling, DNA enrichment assays, genome assembly

ACM Reference Format:
Ronald J. Nowling, Christopher R. Beal, Scott Emrich, Susanta K. Behura, Marc S. Halfon, and Molly Duman-Scheel. 2020. PeakMatcher: Matching Peaks Across Genome Assemblies. 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

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] Susanta K Behura, Joseph Sarro, Ping Li, Keshava Mysore, David W Severson, Scott J Emrich, and Molly Duman-Scheel. 2016. High-throughput cis-regulatory element discovery in the vector mosquito Aedes aegypti. *BMC Genomics* 17 (May 2016), 341.

[3] Olga Dudchenko, Sanjit S Batra, Arina D Omer, Sarah K Nyquist, Marie Hoeger, Neva C Durand, Muhammad S Shamim, Ido Machol, Eric S Lander, Aviva Presser Aiden, and Erez Lieberman Aiden. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 6333 (April 2017), 92–95.

[4] Keshava Mysore, Ping Li, and Molly Duman-Scheel. 2018. Identification of Aedes aegypti cis-regulatory elements that promote gene expression in olfactory receptor neurons of distantly related dipteran insects. *Parasit. Vectors* 11, 1 (July 2018), 406.

[5] Vishvanath Nene, Jennifer R Wortman, Daniel Lawson, Brian Haas, Chinnappa Kodira, Zhijian Jake Tu, et al. 2007. Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* 316, 5832 (June 2007), 1718–1723.

[6] Jim Thurmond, Joshua L Goodman, Victor B Strelets, Helen Attrill, L Sian Gramates, Steven J Marygold, Beverley B Matthews, Gillian Millburn, Giulia Antonazzo, Vitor Trovisco, Thomas C Kaufman, Brian R Calvi, and FlyBase Consortium. 2019. FlyBase 2.0: the next generation. *Nucleic Acids Res.* 47, D1 (Jan. 2019), D759–D765.