PeakMatcher: Matching Peaks Across Genome Assemblies

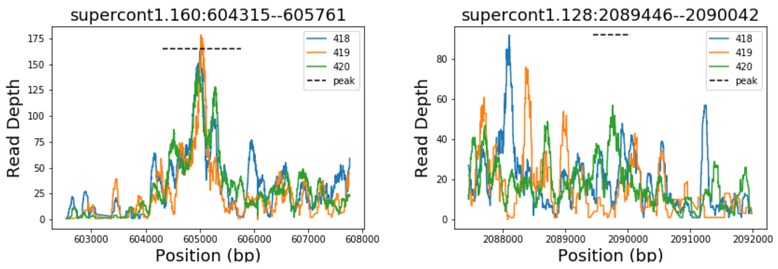
RJ Nowling[1], CR Beal[2], Scott Emrich[3], S. K. Behura[4], M. S. Halfon[5], and M. Duman-Scheel[6]

[1]Milwaukee School of Engineering, [2]Marquette University, [3]University of Tennessee—Knoxville, [4]University of Missouri—Columbia, [5]SUNY at Buffalo, [6]IU School of Medicine

Problem

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. Researchers need ways to compare the peaks called for the new genomes to understand what was reproduced with the new genome what is new, and what might have been missed.

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.



Our Solution

We introduce PeakMatcher, an open-source package that aids in validation by matching peaks across two genome assemblies. PeakMatcher uses the alignment of the same set of reads to two genomes to match peaks across the assemblies. PeakMatcher calculates recall and precision while also outputting lists of matched and unmatched peaks for further downstream analyses.

References

[1] S K Behura, J Sarro, P Li, K Mysore, D W Severson, S J Emrich, and M Duman-Scheel. 2016. High-throughput cis-regulatory element discovery in the vector mosquito *Aedes aegypti*. *BMC Genomics* 17 (May 2016), 341.

[2] O Dudchenko, S S Batra, A D Omer, S K Nyquist, M Hoeger, N C Durand, M S Shamim, I Machol, E S Lander, A Presser Aiden, and E L Aiden. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 6333 (April 2017), 92–95.

[3] B J Matthews, O Dudchenko, S B Kingan, S Koren, I Antoshechkin, J E Crawford, W J Glassford, M Herre, S N Redmond, N H Rose, G D Weedall, Y Wu, S S Batra, C A Brito- Sierra, S D Buckingham, C L Campbell, S Chan, E Cox, B R Evans, T Fansiri, I Filipović, A Fontaine, A Gloria-Soria, R Hall, V S Joardar, A K Jones, R G G Kay, V K Kodali, J Lee, Gareth J Lycett, S N Mitchell, J Muehling, M R Murphy, A D Omer, F A Partridge, P Peluso, A Presser Aiden, V Ramasamy, G Rašić, S Roy, K Saavedra-Rodriguez, S Sharan, A Sharma, M Laird Smith, J Turner, A M Weakley, Z Zhao, O S Akbari, W C Black, H Cao, A C Darby, C A Hill, J S Johnston, T D Murphy, A S Raikhel, D B Sattelle, I V Sharakhov, B J White, Li Zhao, E L Aiden, R S Mann, L Lambrechts, J R Powell, M V Sharakhova, Z Tu, H M Robertson, C S McBride, A R Hastie, J Korlach, D E Neafsey, A M Phillippy, and L B Vosshall. 2018. Improved reference genome of *Aedes aegypti* informs arbovirus vector control. *Nature* 563, 7732 (Nov. 2018), 501–507.

[4] K Mysore, P Li, and M 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.

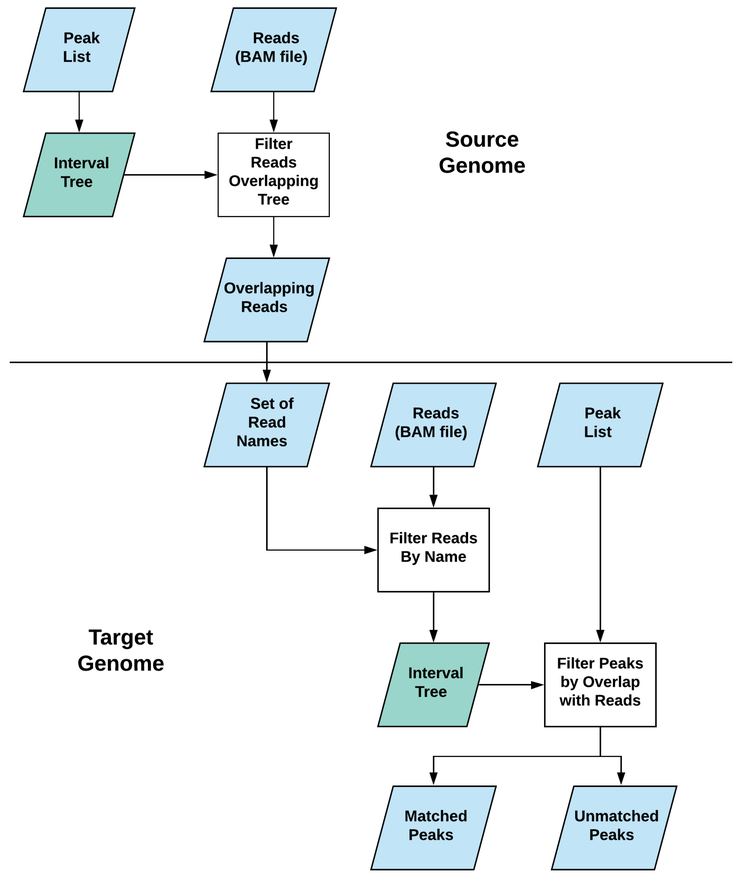
[5] V Nene, J R Wortman, D Lawson, B Haas, C Kodira, Z J Tu, B Loftus, Z Xi, K Megy, M Grabherr, Q Ren, E M Zdobnov, N F Lobo, K S Campbell, S E Brown, M F Bonaldo, J Zhu, S P Sinkins, D G Hogenkamp, P Amedeo, P Arensburger, P W Atkinson, S Bidwell, J Biedler, E Birney, R V Bruggner, J Costas, M R Coy, J Crabtree, M Crawford, B Debruyn, D Decaprio, K Eiglmeier, E Eisenstadt, H El-Dorry, W M Gelbart, S L Gomes, M Hammond, L I Hannick, J R Hogan, M H Holmes, D Jaffe, J S Johnston, R C Kennedy, H Koo, S Kravitz, E V Kriventseva, D Kulp, K Labutti, E Lee, S Li, D D Lovin, C Mao, E Mauceli, C F M Menck, J R Miller, P Montgomery, A Mori, A L Nascimento, H F Naveira, C Nusbaum, S O'leary, J Orvis, M Pertea, H Quesneville, K R Reidenbach, Y Rogers, C W Roth, J R Schneider, M Schatz, M Shumway, M Stanke, E O Stinson, J M C Tubio, J P Vanzee, S Verjovski-Almeida, D Werner, O White, S Wyder, Q Zeng, Qi Zhao, Y Zhao, C A Hill, A S Raikhel, M B Soares, D L Knudson, N H Lee, J Galagan, S L Salzberg, I T Paulsen, G Dimopoulos, F H Collins, B Birren, C M Fraser-Liggett, and D W Severson. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316, 5832 (June 2007), 1718–1723.

Our Method

PeakMatcher uses aligned reads to match peaks across genome assemblies. The user will need to align the same set of reads to both the source and target genomes.

For the source genome, PeakMatcher takes a peak list (e.g., from MACS) and aligned reads (e.g., from a BAM file) as input. PeakMatcher builds an interval tree from the peak list. The interval tree is used to select reads that overlap with the peaks. The names of those reads are then written out to a file.

For the target genome, PeakMatcher takes a peak list, aligned reads, and overlapping read names as input. PeakMatcher selects the aligned reads by name and constructs an interval tree from their alignment coordinates. Peaks are checked for overlaps with the reads using the interval tree and sorted into two categories: matched and unmatched.



Application to *Aedes aegypti* FAIRE Seq

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]. When applied to all of the peaks, Peak- Matcher matched 78.8% (precision) of the 124,959 AaegL3 peaks with 76.7% (recall) of the 128,307 AaegL5 peaks.

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.

Why Not Whole-Genome Alignment?

With the wide array of whole-genome alignments programs such as MUMmer available, why did we opt for a new approach? In our tests on the Aedes FAIRE-Seq peaks, we were only able to match (recall) 40% of peaks by coordinate from a whole-genome alignment between the AaegL3 and AaegL5 assemblies. Our approach (PeakMatcher) was able to match up 78.8% of the peaks. We do not consider this a fair "head-to-head" comparison, however.

Availability

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.