The PARA-suite: PAR-CLIP specific sequence read simulation and processing 2016:05:10695:0:0:REVIEW

1. BASIC REPORTING

1.1 The submission must adhere to all PeerJ policies.
The article meets PeerJ's standards and policies.

1.2 The article must be written in English using clear and unambiguous text and must conform to professional standards of courtesy and expression.

There are few grammatical and typing errors throughout the text as well. A good idea would be that a native English speaker edits the paper for grammar and other minor issues.

1.3 The article should include sufficient introduction and background to demonstrate how the work fits into the broader field of knowledge. Relevant prior literature should be appropriately referenced.

I would like to invite authors to consider some recent literature on this evolving and dynamic technology, for example:
Danan, PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites, 2016. http://www.ncbi.nlm.nih.gov/pubmed/26463383
In addition it should be an adding value to address a comparison with different methodologies, for this reason I suggest the authors to read some related review:
Wang, Design and bioinformatics analysis of genome-wide CLIP experiments, 2015
http://nar.oxfordjournals.org/content/early/2015/05/09/nar.gkv439.full
The advent of cross-linking immunoprecipitation coupled with high-throughput sequencing (genome-wide CLIP) technology has recently enabled the investigation of genome-wide RBP–RNA binding at single base-pair resolution. This technology has evolved through the development of three distinct versions: HITS-CLIP, PAR-CLIP and iCLIP.
Interesting comparison:
http://epigenie.com/quick-review-crosslinking-immunoprecipitation-clip-methods/

1.4 The structure of the submitted article should conform to one of the templates. Significant departures in structure should be made only if they significantly improve clarity or conform to a discipline-specific custom.
The article meets PeerJ's standards and policies.

1.5 Figures should be relevant to the content of the article, of sufficient resolution, and appropriately described and labeled.

I noted low-resolution figures within the text conforming the structure and the template of PeerJ on first submission, please let me ask the authors to prepare individual figure files in one of the following formats: PowerPoint (.ppt), Tagged Image File Format (.tif), Encapsulated PostScript (.eps), Adobe Illustrator (.ai) (please save your files in IIllustrator's EPS format), Portable Network Graphics (.png), or editable PDF. The resolution must be at a minimum resolution of 600 d.p.i. for line drawings (black and white) and 300 d.p.i. for colour or greyscale.

1.6 The submission should be 'self-contained,' should represent an appropriate 'unit of publication', and should include all results relevant to the hypothesis. Coherent bodies of work should not be inappropriately subdivided merely to increase publication count.
No comments.

1.7 All appropriate raw data has been made available in accordance with our Data Sharing policy.
No comments.

2. EXPERIMENTAL DESIGN

2.1 The submission must describe original primary research within the Scope of the journal.
The manuscript appears with an implementation of set of freely available tools analyzing and processing PAR-CLIP's data. In particular the authors included error model inference, PAR-CLIP read simulation based on PAR-CLIP specific properties, a full read alignment pipeline with a modified Burrows-Wheeler Aligner (BWA) algorithm and CLIP read clustering for binding site detection. The simulation's part sounds interesting and novel, but I invite the author to consider a comparison related to different simulation tool of PAR-CLIP's data from literature.
I can suggest for example:
https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-1-r18#MOESM1
http://www.ncbi.nlm.nih.gov/pubmed/26463385

In addition I would like to see differences and adding-value of the proposed methodology (PAR-CLIP simulator) considering for example
Kassuhn W1, Ohler U, Drewe P., CSEQ-SIMULATOR: A DATA SIMULATOR FOR CLIP-SEQ EXPERIMENTS., Pac Symp Biocomput. 2016;21:433-44.
http://www.ncbi.nlm.nih.gov/pubmed/26776207

It would be interesting to see comparison with performance evaluation of different PAR-CLIP's data simulator tools where in this context it would be useful and critical that those tools should have the same characteristics as real datasets.

2.2 The submission should clearly define the research question, which must be relevant and meaningful. The knowledge gap being investigated should be identified, and statements should be made as to how the study contributes to filling that gap.
I think it should be more useful for readers if the authors can explain in more details their research question. For some points it is not clear if the manuscript's message deals with a suite of different PAR-CLIP data analysis tools or with PAR-CLIP's read simulator. It is just a clearly definition of the message, the authors structured their abstract in sub-sections Background, Methods, Results, Availability they could define their aim, inspiring for example Cseq-Simulator's abstract:

"It has not been assessed which of the available tools are most appropriate for the analysis of CLIP-Seq data. This is because an experimental gold standard dataset on which methods can be accessed and compared, is still not available. To address this lack of a gold-standard dataset, we here present Cseq-Simulator"

2.3 The investigation must have been conducted rigorously and to a high technical standard.
No comments.

2.4 Methods should be described with sufficient information to be reproducible by another investigator.
The authors provided two github repositories such as:
https://github.com/akloetgen/PARA-suite
https://github.com/akloetgen/PARA-suite_aligner
In particular it would be interesting and useful in the context of reproducibility to add supplementary sections within the manual and the examples providing codes, parameters and figures to reproduce the results, reported in the Table1, in the PeerJ's manuscript.
https://github.com/akloetgen/PARA-suite/blob/master/Manual.pdf
https://github.com/akloetgen/PARA-suite/tree/master/examples

2.5 The research must have been conducted in conformity with the prevailing ethical standards in the field.
The article meets PeerJ's standards and policies.

3. VALIDITY OF THE FINDINGS

3.1 The data should be robust, statistically sound, and controlled.

Robust

I would like to see if the performances of proposed methodology such as Para-suite are influenced on data re-sampling and bootstrapping. It could be interesting to apply a cross-validation procedure and fitting the model on training data and then predict the results on testing data. In that context I would consider robust results.

Statistically sound:

In the supplementary table S1, they dind't reported any pvalues or corrected FDR.
I would like to see differences (increasing FDR) applying any statistical hypothesis testing, for example Fisher's exact test (FET) that authors prefer.
Without any FDR how can they assess the significance of their results in table S1?
They should also provide the same, a FET for the venn diagram in Supplementary Figure S6.
See for example table1 and table 2 of *Chen,PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis, Genome Biology, 2014*
https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-1-r18#Tab2
https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-1-r18#Tab1

'Controlled'

I have some concerns regarding supplementary tableS4, please the authors can double-check for which methodologies they reported TN and FN equal to zero, because it sounds strange.
If their findings are consistent, please the authors can provide further explanations.

3.2 The data on which the conclusions are based must be provided or made available in an acceptable discipline-specific repository.

See section Experimental design 2.4 for repository methodologies. In addition data used for method section 2.1 of the manuscript, should be summarized in a table reporting platforms (PAR-CLIP, HITS-CLIP, etc), samples, experiments, datafile, accession numbers, etc.

3.3 The conclusions should be appropriately stated, should be connected to the original question investigated, and should be limited to those supported by the results.

In the discussion section (row 530-531) the authors asserted that "*We characterized some of the unique properties of PAR-CLIP sequence datasets that have, to our knowledge, so far not been analyzed, such as preferred read positions for T–C conversion sites and their frequencies per read position.*" This point is interesting but not original, can the authors consider novel CSEQ-simulator findings, as I mentioned in section 2.1 and dealing with the results comparison.
I reported here an extract of CSEQ-simulator's manuscript:

“Finally, in order to generate the CLIP-Seq reads, we induce the diagnostic events (e.g. T-C conversions, deletions and truncations) in the raw reads. To this end, we sample the diagnostic events in the reads according to user specified distribution (diagnostic event profile) that is centred on the binding site.” From Kassuhn W1, Ohler U, Drewe P., CSEQ-SIMULATOR: A DATA SIMULATOR FOR CLIP-SEQ EXPERIMENTS., Pac Symp Biocomput. 2016;21:433-44.

3.4 Speculation is welcomed, but should be identified as such.
No comments.

3.5 Decisions are not made based on any subjective determination of impact, degree of advance, novelty, being of interest to only a niche audience, etc. Replication experiments are encouraged (provided the rationale for the replication, and how it adds value to the literature, is clearly described); however, we do not allow the ‘pointless’ repetition of well known, widely accepted results.
I have a concern related to methodologies used for read aligners.
I invite the authors to provide further proofs as a reference a table for example, from where they can show why they selected Bowtie, Bowtie2, BWA PSSM, STAR, etc.
(http://www.hindawi.com/journals/bmri/2014/309650/tab1/)
The reason could be dual-fold: (i) in terms of fitting approaches with different performances depending on different data-type (i) approached mostly used and cited as well as reduced computation time for example [http://www.hindawi.com/journals/bmri/2014/309650/fig2/#a] .

3.6 Negative / inconclusive results are acceptable.
No comments.