Response to the reviewers – *Alignment-free detection and seed-based identification of multi-loci V(D)J recombinations in Vidjil-algo*

Dear Editor,

We thank you, and the reviewers, for your positive appreciations on our manuscript and the very constructive comments *"Alignment-free detection and seed-based identification of multi-loci V(D)J recombinations in Vidjil-algo"*. Compared to our initial submission, we improved Vidjil-algo to fix the issue we had with TRA+D/TRD+. This had a small impact on time consumption. We also noticed a small issue in benchmark D (LIGM-DB) where some sequences were duplicated. To prevent any bias, we removed the duplicated sequences. Finally we fixed the issues that were raised with Snakemake.

Below we comment in more details some remarks from you and from the reviewers.

Editor

This work contributes an efficient algorithm to extract the so-called "V(D)J junctions" from raw sequencing data, using an approach based on the Aho-Corasick automaton and spaced seeds.

Both reviewers found merits in the technical constribution of this paper and agreed that the paper is generally well-written and easy to follow, even for non experts. Furthermore, the experiments are convicing and the code is open-source. (...)

We are very grateful for the kind appreciations.

While one of them has minor comments and suggestions on how to improve some script in the software repository, the other reviewer asks for some clarifications regarding the experiments. In particular, the authors are kindly requested to address the following points in the revised version of their work: - A discussion on how the seeds are built.

See below, in answer to Reviewer 1

- Inclusion of the method RTCR as another suitable baseline to compare against.

See below, in answer to Reviewer 2

- A discussion on why the previous version of the algorithm performs better in

some circumstances.

See below, in answer to Reviewer 2

Reviewer 1

The evaluation gives the impression that the challenges concerning [VDJ] recombination are now essentially solved; perhaps the authors can comment on whether this is true or not, or what else should be done in the future in this field (except further small optimizations).

This is a very interesting remark. As for genome read mapping, the issue is not in mapping the read themselves but in obtaining meaningful results the more efficiently possible in order to be able to process huge datasets frugally. The ultimate goal would be to have an a linear analaysis of sequences, independently of the number of recombinations systems, and, with this contribution, we are not far from that. This was slightly discussed in the perspectives and we added a paragraph to make that clearer.

- It may be better to move Fig. 1 to the bottom of the page.

Done.

- 46: "Afzal et al. (2019) did a comparison of several of those software ." -¿ software tools.

Done.

- The term "affectation vector" does not sound right to me, but then I am not a native speaker. Maybe one could use "label sequence"?

Thank you. We now use "label sequence", and streamlined these and other terms across the papers. At each accepting state, there is a "list of gene labels".

- Notation: Please write p-value instead of p-value and E-value instead of e-value in running text.

Done.

- 175: the word p-value is not correct here. You mean a 99.9% confidence

interval?

Right, thanks for pointing out this.

- Fig. 5: The titles of the subfigures are too far away from the actual figure. I first mis-interpreted the leftmost figure as the detection results and could not make sense of the statements on lines 291.

Thanks for the suggestion, we shifted the titles down.

- Figures: It should also be said that vidjil-old is probably 2018.02 and -new is the "development" version.

Done.

- Fig. 8: I suggest to place the color legend to the right of the figures, not below.

Done.

- Fig. 8: IGK/vidjil-new: How/why is the designation bar higher than the detection bar?

This was because we launched the designation step independently of the detection step in order to only assess the designation. However your comment made us realize that it could be misleading and that it is unfair as we recommend not using the designation step alone. Thus we now changed this.

The config file could have reasonable defaults for the directories (results/, benchmarks/, software/).

Thanks for the suggestion, this is fixed, now.

MixCR can be skipped, e.g. by using –keep-going on Snakemake, but it would be nicer if there was an option to disable it.

Thanks for the suggestion, now if no *MIXCR_LICENSE* is provided, MiXCR is quietly skipped.

I do get warnings during compilation.

Warnings when compiling vidjil-old are coming from an external library (nlohmann/json.hpp). A warning in the Vidjil-new version has been fixed.

The workflow outputs a lot in information to the screen. This could be written into log files (using a logs/ directory).

Done, thank you for the suggestion.

There are many errors during the Snakefile. Not all are related to MixCR. The software doesn't seem to build properly. Here is a list of the jobs still to be done, i.e. none of the jobs below completes successfully, in spite of –keep-going.

Thanks a lot for pointing out this issue and sorry for the inconvenience. It appeared that our Snakefile was not working properly with relative paths. This is fixed now. We also added a few dependencies that were missing in the Conda environments.

Reviewer 2

Overall the manuscript is very well written, extremely clear, and easy to follow in all of its sections. Results are generally convincing and the datasets and experimental benchmarks seem to have been adequately designed in order to evaluate the performance of the algorithms.

Thanks a lot for these kind appreciations.

One of the core components of the method is an automaton built from spaced seeds extracted from V, D, and J genes. However, it is not clear how these seeds were extracted. I think it is important to discuss it a bit more in depth.

We made this that clearer in the paper. Seed occurrences are extracted at each position of the V, (D), and J genes. We detailed the reason to use spaced seeds in the first paragraphs of the section 2. The very choice and optimization of spaced seeds was not a focus of this paper, we underline that in the perspectives. The actual seeds are now described in the paper. We also detailed the Figure 2 to include different labels for different seeds.

It seems that experiments are focused on data affected by substitution errors *(dataset B). Are datasets C, D, and E affected by indels? I guess the method is mainly thought to be used with sequences that are primarily affected by substitution errors but I think it could be interesting to see how it performs in presence of indels.*

Indeed, this is correct. Regarding dataset C, the article by Afzal *et al* is unclear about whether indels have been included or not. However, after digging into the data, it seems that no indel was included.

As datasets D and E are real sequences, they likely have indels, but at an unknown rate. We have added in dataset B, a dataset with 1% indels. The Figure has been added in the Supplementary Material (Figure 3): the results are very similar.

Since also the designation algorithm has been improved it could also make sense to briefly discuss its complexity with respect to the previous implementation.

The complexity is discussed in the "Selecting candidate genes with seed-based heuristics". We made this clearer, and reminded thee result in the introduction. We also clarified the complexity on the detection algorithm relative to a preprocessing step.

As a matter of fact RTCR also seems to offer a good balance but is also the one with the most consistent performance across datasets. For this reason, I would suggest also to include RTCR in the comparison, in particular to see how it compares against Vidjil-algo in the challenging dataset E.

We apologize to the reviewers and to the editor because we overlooked to make explicit why RTCR was not included. Actually, RTCR is specialized to only consider TCR sequences, it will not process Ig sequences (ie. IGHV, IGK, IGL), this is why we excluded it at first and this is why we mentioned in the article that according to Afzal *et al*, MiXCR is the most balanced *generic* tools. Thus we didn't include RTCR because we didn't consider it to be generic enough. Again, we apologize for not clearly stating our inclusion criteria.

However after your comment we looked more closely at RTCR. It appears that RTCR doesn't seem to be under active development (apart from minor fixes) for at least six years. On top of that, RTCR is actually less generic than we thought as the chain to be analyzed (*eg.* TRA, TRB or TRG) has to be specified beforehand, which is not the case for MiXCR or Vidjil. This doesn't allow us to assess whether the detection phase attributes the sequences to the correct locus, especially in dataset E. Finally, RTCR doesn't give access to the detection or designation for each read, which is required for our assessment, but only to the designation of the clonotypes.

The authors showed the new version of Vidjil-algo provides better results in almost all datasets considered. However in the most challenging one (dataset E), the previous approach actually performs better in some cases. I wonder if the authors investigated this behavior and could discuss the possible reasons.

The main difference (on TRA+D and TRD+) was explained in the article as being an engineering issue to differentiate genes coming from TRA or from TRD as some of them are shared. We acknowledge this appeared as a glitch in our results. Therefore we continued our developments in order to solve this issue. This required to split some germlines in "sub-germlines", hence the time consumption is a bit longer compared to our first submission as we have more germlines, but the results are now good for all germlines, without the exception of TRA+D/TRD+. We removed the sentence from the article.

It might be worthwhile also to mention the specific algorithmic differences between the old and new method in the introduction.

The end of the introduction states now more clearly what are the differences.

The authors