Response letter to the recommender and reviewers

Dear recommender,

We greatly appreciate the comments and suggestions of the reviewers, which were very helpful to improve the quality of our article. Please find bellow our detailed responses.

Kind regards,

Lucie Tamisier

Round #1

by Hadi Quesneville, 2021-01-19 14:03

Manuscript: https://zenodo.org/record/4293594#.X8D6GLPjJEY

Article needs revision

Dear authors,

The two referees found your article interesting and potentially of great value. However, it can be still improved according to their suggestions. I recommend you to take into account their suggestions and to re-submit it for a second evaluation round.

Best regards,

Hadi Quesneville

Reviews

Reviewed by anonymous reviewer, 2020-12-18 08:30

In this manuscript, the authors aim at describing several semi-artificial and artificial dataset of plant virus that could be used to benchmark bioinformatic pipelines for virus identification, allowing the assessment of their performance.

The initiative is very commendable and truly necessary with the number of bioinformatics tools developed today in all fields of biology. However, I have a real problem with this manuscript which seems to me insufficiently accomplished with a lack of information and precision.

The subject of the article is very specialized as it concerns the detection of plant viruses, this is why it is important to better introduce the subject.

A new paragraph was added at the beginning of the introduction, in order to better introduce the plant virus diagnostic field.

There is a problem in the lack of explanation concerning the type of data allowing these detection or how they are obtained (from which biological data). Are they RNA-seq or DNA-seq data, or both? Do they come from purified extract from tissues (meaning are there steps of filtration to enrich in virus sequences or is there also host sequences)?

The reviewer is correct that this crucial information was lacking in the paper. A table was now added to the supplementary material (Table S1, https://zenodo.org/record/4584967#.YEIku-fjKUk) to present the initial data including details on how it was obtained. Seven samples out of 8 were obtained following total RNA extraction, with or without ribodepletion. Therefore, a high background of host sequences is present in these samples, as it is often the case for infected plant samples analysed in diagnostic or in virus discovery. The 4th real dataset (potato leaves infected with PVY, used to create semi-artificial datasets 5 and 6) was obtained following purification of the PVY virions. However, a small fraction of host sequences is also present in this dataset.

Likewise, it would be desirable to recall the existing bioinformatic tools or at least the approaches used depending on the questions asked to have an idea about the difficulties of these approaches.

A non-exhaustive list of bioinformatics tools allowing to perform a complete analysis was added to the text. We also refer to more complete reviews of such tools.

The approach used to analyse HTS samples infected with viruses/viroids always followed the same main steps as explained below. We briefly included these main steps in the introduction of the paper.

- First, quality control and pre-processing of the reads (trimming, adapter removal, optional merging of forward and reverse reads, etc) are performed.
- In some cases, a plant host removal is done by mapping the reads against the plant host genome (if available), and subsequently removing those mapped reads from the data.

- Then, a *de novo* assembly can be done, in order to obtain contigs (but it is not mandatory).
- Several approaches can then be used to identify viral sequences, using either contigs or reads as input, such as mapping against known viral/viroid references, similarity searches against sequence or domain databases, protein domain searches or K-mer based approaches.
- Finally, additional analyses can be done, like more in-depth viral population analyses (SNP calling) and haplotype reconstruction.

Currently, many tools are available to perform each step of the analysis. “All-in-one” pipelines able to perform all the different steps of the analysis also exist (with more or less options to change the parameters). In the end, researchers can choose many different strategies, tools and parameters to analyse their data. All these different methods will impact their results (as shown by Massart *et al.*, 2019) and can lead to a lack of repeatability, which can be a real problem in the diagnostic field, where a repeatable diagnostic is needed. This is why having resources allowing an objective comparison of all these pipelines is of crucial importance.

The proposed dataset are also not very detailed nor the way they have been constructed. Especially concerning the real data. Sometimes figures would be useful to illustrate the text.

Our goal was to write a “resource announcement” type of article, in order to let the scientific community know about the existence of our datasets, in the form of a short announcement. The details regarding the methods are given on the GitLab page. However, we agree that more explanations can be given in the article too. We have added a supplementary text to clarify our method (Text S1, https://zenodo.org/record/4584967#.YEIku-fjKUk). A figure summarizing the challenges and the corresponding datasets has also been added (Figure 1).

Another missing point is the lack of proof of principle to show examples in the use of at least some of these dataset and how they really allow a good benchmarking process.

The reviewer is correct. As mentioned in one of the previous remarks, we see this paper more as a “resource announcement”, in which we want to inform the scientific community that these datasets are available. One of the aims of this publication is also to attract and engage the community to evaluate the datasets by their own approach. In the current form of the paper, it has not been shown that the datasets are good benchmarking datasets. Nevertheless, we expect them to be good starting points for benchmarking since they were constructed to address each at least one challenge that could prevent a correct virus detection. These challenges correspond to real problems faced by the community when analysing HTS data, and should therefore be useful to compare pipelines. For instance, Massart et al., (2019) have compared the ability of 21 plant virology laboratories, each employing a different bioinformatics pipeline, to detect plant viruses in 10 completely artificial small RNA datasets. This study revealed that some pipelines performed poorly if the virus was novel or if the virus

concentration was low. These two challenges (low virus concentration and new virus strain/species) are addressed by our datasets.

The use of the datasets by researchers will allow to evaluate their benchmark ability. Indeed, the VIROMOCK challenge we launched on GitLab is actually a follow-up effort we would like to initiate within the community. People are encouraged on the GitLab website to submit their results for comparison. This will allow researchers to compare their pipelines and possibly to identify key parameters influencing their results and to increase discussion among virologists and bioinformaticians.

To make it clearer, we went through the text and carefully checked our wording to emphasize that these datasets are rather a starting point for benchmarking analyses, rather than the ideal datasets for benchmarking.

Finally, the authors argue about the fact that having semi-artificial dataset allow to bypass the drawbacks of having either only real dataset or completely artificial dataset. This seems contradictory with the fact that the authors propose 3 real dataset and 9 artificial ones among the 18 dataset. Moreover, I think the semi-artificial dataset may also have some drawbacks that could be discussed. It could be possible that the drawbacks of both artificial and real dataset add up.

We agree with the reviewer that each type of datasets can have specific pros and cons and none of them is ideal for any purpose. This is why we attempted to select or develop the most appropriate type of the dataset (real, artificial or semi-artificial) for specific problems put forward in the text.

For example, the 9 completely artificial datasets are proposed for a very specific goal: benchmarking viral haplotype reconstruction software. Currently, viral haplotype reconstruction is one of the most challenging problems in bioinformatics. Indeed, several issues are faced when trying to convert reads from HTS into viral haplotypes. First, for low frequency haplotypes, it is difficult to discriminate between sequencing error and real SNP. Then, if the distance between 2 SNPs exceeds the read length, it is particularly difficult to determine which SNPs occur on the same haplotype (Schirmer *et al.*, 2014). A recent study has compared 12 of the most commonly used haplotype reconstruction software using completely artificial HIV-1 viral populations (Eliseev *et al.*, 2020). Drastically different results have been obtained between the software. Most methods worked well when viral genetic diversity was low, but performed poorly when viral genetic diversity was high (which is the case for intra-host HIV population, and, to a higher extent, for most plant RNA virus populations). Therefore, our ability to reconstruct viral haplotypes can still be improved. No dataset with mix of plant RNA viral haplotypes is currently available to test such software, haplotype callers being mainly developed using HIV data. Since viral haplotype reconstruction is a hard task, producing completely artificial datasets composed of plant RNA viral haplotypes

already constitutes a useful and challenging resource. We agree that this choice can be surprising if this information is not provided, and some sentences have been added to explain our choice in the article.

We also propose 3 real datasets without modifications because their compositions were already challenging. The challenge of these datasets is always to test the ability of the pipeline to detect one or several viruses (either a cryptic virus, a defective variant mixed with a normal length variant, or a segmented virus). For each real dataset, the viruses have been detected with at least 2 independent methods: a serological or molecular one (ELISA or PCR) and through High-Throughput Sequencing. Moreover, the composition of each real HTS dataset has been analysed by at least 2 independent laboratories (the one that gave us the real dataset and our laboratory). This means that our real datasets have been deeply analysed, and that we have a high level of confidence regarding the composition of these real datasets. The same procedure has been performed for all the real datasets. Therefore, the semi-artificial datasets are composed of deeply-analysed real data mixed with completely known artificial reads. In our opinion, the benefits associated with this kind of semi-artificial datasets are greater than the disadvantages.

In sum, I think this work is needed since benchmarking bioinformatic tools is of utmost importance. However, this manuscript does not meet, at this stage, standards of scientific publications.

Reviewed by Alexander Suh, 2021-01-19 10:29

Tamisier et al. provide a combination of real and semi-artificial datasets with high relevance for benchmarking detection and analysis approaches in plant virus detection. The manuscript is succinct and well written, accompanied by a detailed GitLab repository, and proposes the VIROMOCK challenge as a community-driven effort to benchmark virus detection and analysis.

Below are some minor suggestions for improved clarity that the authors may want to implement to help a broad readership.

1. **Line 86: It is unclear whether the read lengths vary within or between each data set. Table 1 suggests that the latter is mostly the case, however, then it would help the reader if the distinct sets of read lengths were stated here in the text.**

   The read lengths vary between each dataset. It was clarified in the text.

2. **Lines 91-94: Both for the real and artificial dataset, I recommend briefly discussing the potential issues arising from Illumina's recent shift from a four-channel system (e.g., HiSeq X) to a two-channel system (e.g. NovaSeq). A recent opinion piece by De-Kayne et al. (https://onlinelibrary.wiley.com/doi/epdf/10.1111/1755-0998.13309) reviewed evidence for T>G errors in NovaSeq data and provided suggestions for how to deal with this. I assume this does not affect the datasets presented in the present manuscript (assuming all data here are based on HiSeq data or simulated on these), but this may be important to be pointed out for readers using NovaSeq data or HiSeq/NovaSeq combinations after benchmarking with the present datasets. Please also clarify in the text what system the present datasets are based or simulated on.**

   All the datasets were sequenced or simulated using an Illumina four-channels system (either HiSeq or Miseq), except the datasets 9 and 10 which were sequenced on an Illumina two-channels system (NextSeq). Therefore, most of our datasets are not impacted by the potential bias produced by the two-channel system.

   However, this potential issues due to Illumina's shift from four-channel system to two-channel system is indeed important to raise. The sequencing platforms used to sequence our real datasets are now mentioned in Table S1 (https://zenodo.org/record/4584967#.YEIku-fjKUk) and in Table 1, so that readers know on which platform the datasets have been generated. Sentences about the potential bias you mentioned have also been added in the text.

3. **Line 101: Here and throughout, it may be unclear to some readers whether "non-complete genome" refers to the virus or the host.**

   Modifications were made in the text, the table and the GitLab to clarify that we talk about an incomplete virus genome coverage.

4. **Line 113: I commend the authors on preparing a very detailed GitLab repository. The Dryad download links appear to be working here, unlike the DOIs stated in Table 1. Please make sure that the DOIs stated in Table 1 are accessible, I was unable to have a look at the datasets through the Table 1 DOI links.**

   A column with the DOI URL was added to the Table 1.

5. **Line 145: Did the authors double-check that the random removal of reads led to complete absence of coverage for some genomic regions of these viruses, rather than reduced coverage for these regions?**

All the semi-artificial datasets were analysed after their creation. The absence of coverage for some genomic regions has been confirmed for the 3 concerned viruses (GRSPaV, GRVFV and GLRaV2) and is shown in a plot in the GitLab page:

https://gitlab.com/ilvo/VIROMOCKchallenge/-/blob/master/Datasets/Dataset3.md

6. **Line 216: I like the diversity of challenging datasets discussed in the text and the authors' idea for the VIROMOCK challenge, however, for visual learners it might help to summarize key points in a figure. If the authors agree that this would help, consider providing simplified illustrations of virus detection/analysis challenges (with pointers to datasets 1-18), and/or the suggested community-driven approach of the VIROMOCK challenge.**

   A figure (Figure 1) summarizing the challenges and the corresponding datasets was added.

7. **Table 1: In the modification column, consider stating the number of reads (or read pairs) added, and possibly also the number of strains.**

   The number of strains and reads added (or removed) was added in the “Modification” column. Note that all this information is provided with more details on the GitLab page.

8. **Table 1: In the "Challenge" column, it is not always clear which virus a specific "mutation" or "strain" refers to. Please revise for clarity by adding as much information as space allows.**

   The viruses targeted by the challenge was added in the “Challenge” column for each dataset.

References

Eliseev, A., Gibson, K.M., Avdeyev, P., Novik, D., Bendall, M.L., Pérez-Losada, M., Alexeev, N. and Crandall, K.A. (2020) Evaluation of haplotype callers for next-generation sequencing of viruses. *Infect. Genet. Evol.*, 104277.

Massart, S., Chiumenti, M., De Jonghe, K., et al. (2019) Virus detection by high-throughput sequencing of small RNAs: Large-scale performance testing of sequence analysis strategies. *Phytopathology* **109**, 488–497.

Schirmer, M., Sloan, W.T. and Quince, C. (2014) Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. *Brief. Bioinform.* **15**, 431–442.