Dear Dr. Bierne,

Please find enclosed the revised version of our manuscript entitled “Somatic mutation detection: a critical evaluation through simulations and reanalyses in oaks”, that we would like you to consider for recommendation in *PCI Genomics*. We thank you and the reviewers for their thorough comments on our manuscript. We better acknowledged the specificity of our virtual experiments and reanalyses and discussed other possible contexts that could be explored with the developed workflows in the revised discussion. As suggested by reviewer 2, we also assumed from the very beginning that cancer callers will be more suited for somatic mutation calling than SNP variant callers, and presented our work as a case study to advocate for their use. To avoid circular reasoning, we avoided any use of the term “robust”, we tried to clarify how *in silico* evidence was used to assume the better performance of *Strelka2* on real data, and how the increased number of low-frequency mutations seemed thus logical. We also tried to insist on the remaining uncertainties to avoid misunderstanding as these mutations stay candidate mutations to be validated (but see our additional section in the discussion as well as the Supplementary Note 2).

You will find below our responses to the reviewers’ comments, and in red in the linked file our changes to the previously submitted version of the manuscript. We hereby confirm that this work has not been published or accepted for publication, and is not under consideration for publication. Submission for publication has been approved by all co-authors and all persons entitled to authorship have been so named.

Yours sincerely,
The authors

Dear Dr. Schmitt,

I have received two thoughtful reviews of your preprint entitled “Somatic mutation detection: a critical evaluation through simulations and reanalyses in oaks”. The referees are globally positive and expressed mainly minor concerns though important to account for. I’d ask you to account for all these concerns in a revised version.

Basically both referees think that you interpreted as performance differences between callers what are indeed different purposes. If the objective is to identify somatic mutations in a low proportion of cells in the population of cells analysed, it is clear that cancer callers are more designed to do that while generic/SNP callers aren’t (they are rather designed to assign low VAF mutations to sequencing errors). So one part of the issue is not about some callers outperforming others but more about what we want to infer with these callers, and this should be made clearer. Some callers (e.g. Octopus) have different models depending if the user wants to call germline mutations, somatic mutations, or depending on the ploidy and you can even analyse pool seq data or low coverage data. You therefore have to make clearer that you are interested by datasets that consist of multiple tissues sampling of the same tree in order to detect somatic mutations and that it is different from standard SNP calling, and also different from standard cancer genomics experiments (referee 2 even suggested that ideally a caller should be designed or tuned for this type of data). Then you have the study of the performance of cancer callers to detect somatic mutations (under your given study design) that is very interesting to investigate. However, you should make clearer how you evaluated the robustness, in order to be sure you do not have circular reasoning (referee 1). You should clarify how sequencing errors and true low frequency somatic mutations can be sorted out in order to allow you to claim that Strelka2 outperforms other callers with true data. Is it more mutations the better or is it more subtle? You should also explore or discuss the effect of default parameters in the difference between callers (referee 2). At any rate, it would be better framing to argue that Strelka2 is better designed to the purpose and the data than to say it outperforms other callers. Mapping is also an important phase of the pipeline and it could be interesting to discuss that some studies used Bowtie and others BWA (referee 1).

I thank you for submitting your preprint to PCI Genomics peer reviewing and I'm looking forward to reading your revised version.

Best regards,

Nicolas Bierne

Reviews

Reviewed by anonymous reviewer, 06 Jul 2022 15:01

Thank you for your patience in awaiting this review.

The main conclusion of this manuscript is that somatic variant callers developing for cancer sequencing data can be used to improve de novo mutation calling in plants, and to reconcile empirical observations with theoretical expectations. This represents an interesting and worthwhile application of these methods and potentially creates an exciting interdisciplinary space within which cancer bioinformaticians and plant biologists might align interests. Furthermore, the new exploration of annotations and mutation spectra in previously published plant datasets presented by the authors is worthwhile.

My main concern on reading the manuscript relates to the reanalyses of the Plomion and Schmid-Siegert datasets which results in a number of claims - 1) that somatic variant callers outperform 'generic' callers when applied to plant data in terms of recall and precision and 2) that application of Strelka2 enabled to authors to identify a greater number of somatic mutations than in previously published datasets. However, the basis for these claims is uncertain.

> Thank you for this thorough review and justified concern. We claim that cancer callers outperform SNP callers based on the *in silico* virtual experiment "the best-performing caller based on our simulations". We have tried to clarify this in the new version. We agree that somatic mutations detected remain putative mutations at this stage. We have tried to discuss this further in the revised manuscript. Nevertheless, based on the results of the virtual experiment and the frequency distribution of alleles indicating low allele frequency missed mutations, where *Strelka2* recovered the results that were most similar to the *in silico* simulated mutations, we therefore state that Strelka2 was the best caller at recovering simulated mutations. We still chose to use the most stringent filtering to further analyse the

annotations and mutation spectra. We therefore tried to better acknowledge the uncertainty in the revised manuscript while maintaining our message.

The authors state '*our analyses were able to detect far more robust candidate mutations than initially reported*.' How is this robustness being evaluated? Based on EV scores generated by Strelka2? If so, this seems to be circular reasoning i.e. evaluating Strelka2 against other variant callers using metrics generated by Strelka2.

'*Adding Strelka recommended filtering ... yielded ... a 2 to 7-fold increase compared to the original number of mutations*.' This is worth reporting, but again what is the specific evidence that these mutations are more likely to be true somatic mutations? VAF distribution? Spectra? Without any ground truth, it's difficult to see the extent to which the authors can argue that they have achieved better results than previous studies. I'm not convinced that the authors show that Strelka2 actually outperforms other callers when applied to plant data specifically.

> For the sake of clarity we stopped using “robust” in the revised manuscript. The EVS were computed only for *Strelka2* and thus not used as measurement of robustness but as an extra layer of stringent filtering based on machine learning. The probably most convincing argument for the better performance of *Strelka2* was the combination of the results of the virtual experiment and the frequency distribution of alleles indicating low frequency mutations were frequently missed by SNP callers. We tried to clarify the manuscript accordingly. We further acknowledged that these mutations stay candidate mutations to be validated (but see our additional section in the discussion as well as the Supplementary Note 2).

Also to what extent does mapping confound these results, the Schmid-Siegert paper uses Bowtie which has a lower mapping rate than BWA used in Plomion and Schmitt papers?

> We reanalysed both datasets from scratch. We downloaded the raw data, then applied the same strategy for the two datasets. We added the information in methods: “Sequenced reads of every library were quality checked, trimmed and mapped with the same strategy than previously described for the simulation work, i.e. using FastQC, Trimmomatic and BWA mem”.

Ultimately I think this section would need to be made more robust before recommending the paper. Some further minor revisions are suggested below.

- The authors should confirm at some point early in the text that the pedunculate oak is highly heterozygous and diploid. Ploidy should be relevant to the suitability of most somatic variant callers. I expect implications arising from this work would be limited to diploid plant species but the authors may wish to comment on this.

> Thank you for the thorough comment. We further highlighted that pedunculate oak is diploid in the revised manuscript, and we added a discussion on polyploidy with assumptions and future directions.

- At lines 39-42, it would be worth citing Alex Cagan's recent (2022) work on somatic mutation rates here.

> Thank you, as suggested we added a reference to his work in the introduction.

- At lines 47-8, "*The drivers of new mutations, previously thought to be simply due to DNA replication errors, are now also debated".* It is unclear what the authors mean by this.

> We clarified the sentence: “The relative contribution of DNA replication errors and DNA repair fidelity to mutation rates represents another timely evolutionary question (Gao et al., 2019).”

- The authors should refer to FreeBayes, GATK, Samtools/mpileup, VarScan etc. as "SNP variant callers" instead of "generic variant callers".

>Thank you for the suggestion that we integrated throughout the revised manuscript.

- At line 153, "*depth of sequencing depth*" should be "sequencing depth"

> Corrected.

- Line 345-7, the authors state "*Our simulation framework therefore provides general insights regarding the impact of allelic dosage in mutation detection which go beyond somatic mutation detection".* What are these insights? This is the first time that allelic dosage is mentioned in the text.

> Sorry for the inappropriate use of allelic dosage here, we replaced it by allele fraction for consistency.

- There is probably a better way to refer to the respective Plomion and Schmid-Siegert datasets than as the 'French' vs. 'Swiss' datasets

>This is indeed incorrect and we used “the data from Schmid-Siegert et al. (2017) and [...] the data from Plomion et al. (2018)” for clarity and precision.

- Concerning the cross-valiation approach used in analysing the Plomion datasets, the authors speculate that they are likely to lose some 'real' somatic mutations using this strategy. Can the authors test this by looking at the 'robust candidate somatic mutations' identified with and without cross-valildation?

>With *Strelka2* EVS candidate mutations, filtering with the cross-validation filtered out 27% of sites. We added these figures to the revised manuscript.

- Lastly, the authors don't discuss or consider the application of either subclonal or mosaic variant callers developed for 'normal' data e.g. deepSNV, MosaicForecast.

> Our virtual experiment was designed to take a step forward. But of course this experiment, as any experiment, could be improved. Regarding the use of *deepSNV*, it should be noted that the software is developed for high coverage data which we do not use here. *MosaicForecast* is developed for a "tumour only" mode which does not fit our sample design. The *detectMutations* pipeline we provided could however be extended to include these callers, as well as some new callers in the future, but we have chosen not to do so in the context of our study for the aforementioned reasons. Nevertheless, we discuss applications

for which *generateMutations* and *detectMutations* could be extended in future contributions in the revised manuscript, pointing to these tools.

Reviewed by anonymous reviewer, 28 Jun 2022 00:45

The authors investigate the use of cancer specific variant callers compared to generic variant callers for the discovery of somatic mutations in long-lived plants.

First, I would like to congratulate the authors for the effort they put into making the pipelines reproducible and their code accessible.

The manuscript raises an important issue, which I think useful in what seems to be a new area of research (intra-individual mutations in plants). They rightly argue that cancer variant callers are better adapted for this type of data and additionally compare several of them.

Their re-analysis of published data with the variant caller identified as best-performing provides a strongly increased set of mutation candidates. Those numbers better fit theoretical expectations.

I do not have any major issues with the re-analysis part, but have some with the first part of tool comparison.

It seems the manuscript was initially formatted for a "short format" journal. I feel that some parts of the Supplementary could be added into the main text given PCI and biorxiv do not have size limits.

> Thank you for your thorough review, including the reproducible pipeline codes. We have taken note of your comments and have tried to address them the best we could. A previous version of the manuscript was indeed formatted for a short paper but following your suggestion, we have now moved significant parts of the supplementary text into the manuscript.

Major Comments:

1) I feel that there is an obvious statement and explanation missing from the manuscript, which would require some re-framing.

First, the description of what a variant caller is doing should be made broader to encompass both types. A variant caller evaluates the probability of a genotype given the data and an underlying model with some **assumptions**, which vary between callers and might be closer or not to the data considered. Each variant caller is designed for a specific purpose, with choices made by the programmer on read filtering, models and thresholds of sensitivity in the output (one caller might decide to output more false positives, expecting post-hoc filtering by the user).

While I find the objective of the manuscript of showing that plant researchers should use cancer variant callers to look for somatic mutation important, the first part using simulated data seems to me somehow superficial in its present form.

Compared to generic variant callers, cancer callers are **designed** for the type of data you are giving them due to the underlying probabilistic models they use. This is the obvious statement that is missing in my opinion, that the difference is expected and that generic callers shouldn't be used for this type of data in the first place. So this is saying "only based on the design of the tools, there is no question that cancer callers will be a better fit to the data used here. For those that do not believe it, here is an *in silico* demonstration. But really, we shouldn't need that and looking at the models should be enough".

To me, the comparison of generic and cancer variant callers is similar to testing the fit of different models to data generated under one of those (or a close enough one). (Or similar to testing what tool between a hammer and a screwdriver will perform best at driving a nail). So I recommend that the main argument to use cancer callers against generic ones should be a verbal one based on the design of the tools. After this, it is indeed interesting to investigate which of the cancer variant callers best fit the type of data coming from the search for somatic mutations in plants. You could even end in calling for the design of a variant caller with a more specific model that fits somatic mutation search in plants.

> Thank you for the thorough review, and we acknowledge that logically cancer callers are more appropriate than SNP callers for detecting somatic mutations, even in a plant research context. Nevertheless as pointed out in the revised manuscript "*Despite the advantage of cancer callers to identify mutations, SNP callers have been the most frequently used method to detect somatic mutations in plant research (Schmid-Siegert et al., 2017; Watson et al., 2016; Hanlon et al., 2019; Orr et al., 2019)*". So we redefined the aim of the manuscript as "*advocate the use of cancer callers in somatic mutation detection for plant research*". We think that a quantified demonstration of the performance differences stays useful for the field as the most recent publications on plant mutations still relied on *GATK*. Furthermore, we only used two SNP callers, and as you pointed out it is "interesting to investigate which of the cancer variant callers best fit the type of data coming from the search for somatic mutations in plants". Therefore, we prefer to retain this section, albeit incorporating your suggested rephrasing.

2) As said above, each caller will have pre-set filters on both input quality it considers and what variants it outputs. While using default parameters or a given set of parameters seems to be the norm in variant caller comparison literature, I find it is an unfair practice to evaluate variant callers. Variant caller programmers have choices to make when it comes to choosing default parameters and the choice is subjective and dependent on the objective. End users are expected to adapt parameters to their use case.

A fair pipeline would explore the parameter space to find the set where each caller perform best, then compare callers on their best sets (of both input filtering and output filtering parameters).

This point is a comment not restricted to this study, as this seems to be common practice in the literature. While there might not be any good solution to this issue and a solution would require a large amount of work, I consider this calls at least for a discussion of the issue.

Pertaining to this issue, you should explain the choice of parameters made for each caller and why you think this is the set that could provide the best result for it.

> A virtual experiment can always be improved and we of course agree that testing the effect of additional parameters on caller performance could be important. Nevertheless, increasing the number of parameters exponentially increases the number of combinations of parameters. To obtain the current simulation results, the computational load was already substantial. We therefore discuss this limitation of our study in the revised manuscript and are open to future contributions which we hope could be made with the help of the *generateMutations* and *detectMutations* pipelines.

Minor comments:

- abstract: I suggest not including any "suspense" in the abstract and clearly stating which caller best performed in your analyses and used for the data reanalysis.

> We acknowledged in the abstract that *Strelka2* was “the best-performing caller based on our simulations”.

- L53: "herbs"? I am not a plant biologist and find this term cryptic, couldn't you use "short-lived species" or "annual plants" instead? In contrast to your use of "long-lived species".

> Thanks for the suggestion, we replaced “herbs” by “short-lived species”.

- L55: small number note 9 after "processes" does not refer to anything. A missed change of reference system reformatting I guess.

> Sorry for the mistake, it was a reference to (Schoen & Schultz, 2019) now properly formatted.

- L88-92: I would have expected some discussion of genome size in this paragraph. Genome size will strongly influence what is possible to obtain in terms of sequencing depth with a fixed sequencing budget. And this is especially true for some trees that have particularly large genomes compared to humans.

As an addition to this comment, maybe a discussion on the scalability of cancer callers to very large dataset could be included, as they were indeed designed to work with human genomes. Is using them with quite larger genomes practical?

> Thank you for the suggestion, indeed genome size of plants in terms of sequencing capacity and performance of variant detection tools is a relevant question, as it impacts on the experimental design and analytical strategy and capacity. We added a paragraph of discussion on the topic in the revised manuscript. As an extra argument for their use, cancer callers also outperform SNP callers in computing time.

- L96: I feel like something is missing in this sentence as you talk about *in silico* and empirical data, and only describing the empirical data. I suggest adding "[...], using simulated reads with known mutation and two large published [...]" or something similar.

> Thank you for the suggestion that we integrated in the revised manuscript.

- Methods in general: Please include versions for all softwares used in this manuscript. This is the only piece missing in your well done reproducible pipeline. (Especially that versions are not provided inside the pipelines as you use the :latest versions of the singularity containers).

> Thank you for the comment and the suggestion. Unfortunately several images have only a “latest” tag. Consequently, we set the tag when possible for all images in the GitHub repositories and we added the versions in the manuscripts for all softwares.

- L151: "(2) the reference haploid genome with heterozygous sites [...]" I am not sure I understand this sentence. Is it the original haploid sequence with only sites given as heterozygous that have been changed. Maybe rephrase to clarify, or maybe merge (1) and (2) to better explain, e.g. "the diploid genome as two sequences, one being the raw reference and the other being the same sequence only modified at heterozygous sites".

> We clarified the sentence in the revised manuscript.

- L153: From (3), it is difficult to understand if all mutations have the same AF and C or if each mutation have its own (drawn from a distribution?).

> In one simulation all mutations have the same AF and C as indicated in the revised manuscript.

- L192-193: There is a discrepancy between "simulate back one thousand het sites" and "N = 10^4" (which is ten thousand).

> This is indeed ten thousand heterozygous sites. Thanks for noticing the mistake.

- L211-212: Best-performing caller is dependent on the parameter space, it should be specified here (and you show this yourself in Fig S4).

> We precise now in which context “ at low coverage and allelic fraction”.

- L253: "truly simulated" -> true simulated

> Corrected.

- L326: Specify that what is compared is the reanalyzed data from those papers (if I'm not mistaken) and not the original data.

> We added a precision for the sake of clarity.

- Fig 3A: Is the N=510611 a mistake?

> This is not a mistake, GATK with the same filtering used with Strelka2 suggested 510 611 candidate mutations.