Peer Community In Genomics

How to best call the somatic mosaic tree?

Nicolas Bierne *based on reviews by 2 anonymous reviewers*

A recommendation of:

Somatic mutation detection: a critical evaluation through simulations and reanalyses in oaks

Sylvain Schmitt, Thibault Leroy, Myriam Heuertz, Niklas Tysklind *(2022) bioRxiv, 2021.10.11.462798, ver. 4 peer-reviewed and recommended by Peer Community in Genomics* ***https://doi.org/10.1101/2021.10.11.462798***

Open Access

Data used for results

- ***https://www.ncbi.nlm.nih.gov/bioproject/327502***
- ***https://www.ebi.ac.uk/ena/browser/view/PRJEB8388***
- ***https://doi.org/10.5281/zenodo.7274868***
- ***https://doi.org/10.5281/zenodo.7274872***

Published: 2022-11-08

Submission: posted 28 April 2022
Recommendation: posted 25 October 2022, validated 08 November 2022

Cite this recommendation as:
Nicolas Bierne (2022) How to best call the somatic mosaic tree?. *Peer Community in Genomics, 100024.* ***10.24072/pci.genomics.100024***

Recommendation

Any multicellular organism is a molecular mosaic with some somatic mutations accumulated between cell lineages. Big long-lived trees have nourished this imaginary of a somatic mosaic tree, from the observation of spectacular phenotypic mosaics and also because somatic mutations are expected to potentially be passed on to gametes in plants (review in Schoen and Schultz 2019). The lower cost of genome sequencing now offers the opportunity to tackle the issue and identify somatic mutations in trees.

However, when it comes to characterizing this somatic mosaic from genome sequences, things become much more difficult than one would think in the first place. What separates cell lineages ontogenetically, in cell division number, or in time? How to sample clonal cell populations? How do somatic mutations distribute in a population of cells in an organ or an organ sample? Should they be fixed heterozygotes in the sample of cells sequenced or be polymorphic? Do we indeed expect somatic mutations to be fixed? How should we identify and count somatic mutations?

To date, the detection of somatic mutations has mostly been done with a single variant caller in a given study, and we have little perspective on how different callers provide similar or different results. Some studies have used standard SNP callers

that assumed a somatic mutation is fixed at the heterozygous state in the sample of cells, with an expected allele coverage ratio of 0.5, and less have used cancer callers, designed to detect mutations in a fraction of the cells in the sample. However, standard SNP callers detect mutations that deviate from a balanced allelic coverage, and different cancer callers can have different characteristics that should affect their outcomes.

In order to tackle these issues, Schmitt et al. (2022) conducted an extensive simulation analysis to compare different variant callers. Then, they reanalyzed two large published datasets on pedunculate oak, *Quercus robur.* The analysis of in silico somatic mutations allowed the authors to evaluate the performance of different variant callers as a function of the allelic fraction of somatic mutations and the sequencing depth. They found one of the seven callers to provide better and more robust calls for a broad set of allelic fractions and sequencing depths. The reanalysis of published datasets in oaks with the most effective cancer caller of the in silico analysis allowed them to identify numerous low-frequency mutations that were missed in the original studies.

I recommend the study of Schmitt et al. (2022) first because it shows the benefit of using cancer callers in the study of somatic mutations, whatever the allelic fraction you are interested in at the end. You can select fixed heterozygotes if this is your ultimate target, but cancer callers allow you to have in addition a valuable overview of the allelic fractions of somatic mutations in your sample, and most do as well as SNP callers for fixed heterozygous mutations. In addition, Schmitt et al. (2022) provide the pipelines that allow investigating in silico data that should correspond to a given study design, encouraging to compare different variant callers rather than arbitrarily going with only one. We can anticipate that the study of somatic mutations in non-model species will increasingly attract attention now that multiple tissues of the same individual can be sequenced at low cost, and the study of Schmitt et al. (2022) paves the way for questioning and choosing the best variant caller for the question one wants to address.

References

Schoen DJ, Schultz ST (2019) Somatic Mutation and Evolution in Plants. Annual Review of Ecology, Evolution, and Systematics, 50, 49–73. https://doi.org/10.1146/annurev-ecolsys-110218-024955

Schmitt S, Leroy T, Heuertz M, Tysklind N (2022) Somatic mutation detection: a critical evaluation through simulations and reanalyses in oaks. bioRxiv, 2021.10.11.462798. ver. 4 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2021.10.11.462798

Conflict of interest:
The recommender in charge of the evaluation of the article and the reviewers declared that they have no conflict of interest (as defined in the code of conduct of PCI) with the authors or with the content of the article.

Reviews

Toggle reviews

Reviewed by anonymous reviewer, 12 Sep 2022

I thank the authors for their patience. The manuscript is much improved. The authors have engaged fully with the suggested revisions. In particular, the paper now contains a more balanced and informed discussion of the current and future usefulness of somatic variant callers for analysis of plant genomes and of the limitations and scope of the preprint and overall presents a fairer picture of this study's contribution. One further suggestion, though minor, is that the appearance of the abstract might be improved by removing the numbering. Overall, I think this paper should be recommended for publication and will be interested to follow developments in this area.

Reviewed by anonymous reviewer, 23 Aug 2022

After having read the answers of the authors and the new version of the manuscript, I am satisfied by their modifications and answers.

I do not have further comments and recommend the acceptance of this work.

Evaluation round #1

DOI or URL of the preprint: *https://www.biorxiv.org/content/10.1101/2021.10.11.462798v2*

Author's Reply, 17 Aug 2022

Download author's reply

Decision by Nicolas Bierne, posted 07 Jul 2022

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 analyzed, it is clear that cancer callers are more designed to do that while generic/SNP callers aren’t (they are rather designed to assigned low VAF mutations to sequecing 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 (eg Octopus) have different models depending if the user wants to call germline mutations, somatic mutations, or depending on the ploidy and you can even analysed poolseq 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 to have submitted your preprint to PCI Genomics peer reviewing and I’m looking forward to reading your revised version.

Best regards,
Nicolas Bierne

Reviewed by anonymous reviewer, 06 Jul 2022

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 manuscipt 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.

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 that authors show that Strelka2 actually outperforms other callers when applied to plant data specificaly.

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?

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 shoud 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.

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

- 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.

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

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

- 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.

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

- 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?

- 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.

Reviewed by anonymous reviewer, 28 Jun 2022

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.

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 caller 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.

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.

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.

- 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".

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

- 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?

- 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.

- 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).

- 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".

- 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?).

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

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

- L253: "truly simulated" -> true simulated

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

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