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

Sylvain Schmitt, Thibault Leroy, Myriam Heuertz, Niklas Tysklind

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

Sylvain Schmitt, Thibault Leroy, Myriam Heuertz, Niklas Tysklind

CNRS, UMR EcoFoG (Agroparistech, Cirad, INRAE, Université des Antilles, Université de la Guyane), Campus Agronomique, 97310 Kourou, French Guiana; Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, A-1030 Vienna, Austria; IRHS-UMR1345, Université d'Angers, INRAE, Institut Agro, SFR 4207 QuaSaV, 49071 Beaucouzé, France; Université Bordeaux, INRAE, BIOGECO, 69 route d'Arcachon, CS 80227, 33612 Cestas Cedex, France; INRAE, UMR EcoFoG (Agroparistech, CNRS, Cirad, Université des Antilles, Université de la Guyane), Campus Agronomique, 97310 Kourou, French Guiana

- **Corresponding author**: Sylvain Schmitt, +33 6 49 19 32 63, sylvain.m.schmitt@gmail.com
- **Running title:** Somatic mutation detection in plants

Abstract

1. Mutation, the source of genetic diversity, is the raw material of evolution; however, the mutation process remains understudied, especially in plants. Using both a simulation and reanalysis framework, we set out to test the performance of two types of variant callers, generic ones and those developed for cancer research, to detect *de novo* somatic mutations.
2. In an *in silico* experiment, we generated Illumina-like sequence reads spiked with simulated mutations at different allele frequencies to compare the performance of seven commonly-used variant callers to recall them. More empirically, we then reanalyzed two of the largest datasets available for plants, both developed for identifying within-individual variation in long-lived pedunculate oaks.
3. Even in plants, variant callers developed for cancer research outperform generic callers regarding mutation recall and precision, especially at low allele frequency. Such variants at low allele frequency are typically expected for within-individual *de novo* plant mutations. Reanalysis of published oak data with the best-performing caller based on our simulations identified up to 7x more somatic mutations than initially reported.
4. Our results advocate the use of cancer research callers to boost *de novo* mutation research in plants, and to reconcile empirical reports with theoretical expectations.

Introduction

DNA sequence mutation is the raw material for evolutionary change, but, despite its crucial role, many fundamental questions around the mutation process are still open. The understanding of mutation processes is one of the most common conceptual difficulties in biology (Smith & Knight, 2012; Prevost *et al.*, 2013). Mutations are often assumed to occur at a relatively constant pace (i.e. following the hypothesis of a 'perfect' molecular clock). Despite the extremely low number of direct mutation rates estimates, mutation rates are however known to be highly variable along the tree of life, differing by several orders of magnitude among species and kingdoms, and are considered as an evolvable trait *per se* Lynch *et al.,* (2016). Mutations are assumed to be random, but the rate at which different nucleotides mutate strongly depends on the genomic context, in particular the surrounding nucleotides (Martincorena & Campbell, 2015), hereafter referred to as a mutation spectrum. The mutation spectra themselves are now believed to evolve over time (Milholland *et al.,* 2017), even at relatively short evolutionary timescales (Harris & Pritchard, 2017). The drivers of new mutations, previously thought to be simply due to DNA replication errors, are now also debated (Gao *et al.,* 2019).

Unlike most animals that transmit to the next generation only mutations present in their germ cells (*i.e.* sperm and eggs), plants are expected to produce heritable somatic mutations as they grow throughout their lives, departing from the so-called Weismann's germ plasm theory (Weismann, 1893; but see also Lanfear, 2018). As a consequence, long-lived species, such as trees, are generally assumed to accumulate more heritable mutations than herbs

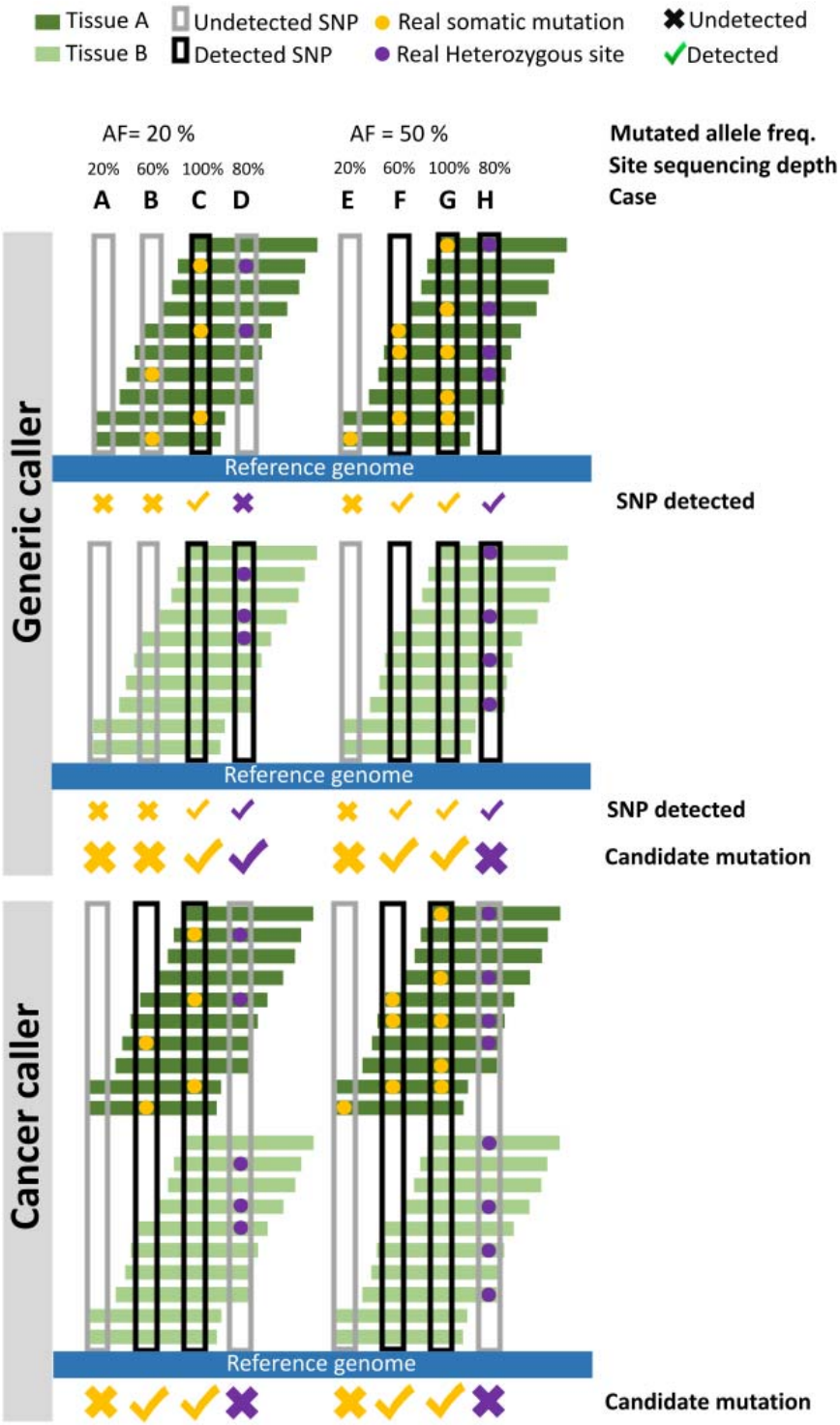
per generation (Hanlon *et al.,* 2019). To generate new knowledge on plant mutation processes[9], several studies examined within-individual variation in long-lived trees, whose individuals can live for more than a thousand years (Schöngart *et al.,* 2017). Two studies used the pedunculate oak (*Quercus robur*), a long-lived European tree species, as a plant model to identify somatic mutations. Schmid-Siegert *et al.,* (2017) identified 17 mutations by comparing sequencing data from two branches of a 234-year-old individual. The authors therefore argued that their results are consistent with a low mutation rate in pedunculate oak. Plomion *et al.,* (2018) identified 46 mutations using three branches of a younger (century-old) individual, which is an almost 10-fold higher rate after taking the tree age difference into account. Plomion *et al.,* (2018) also recovered these new mutations on acorn embryos collected on the same branches as those used for the *de novo* mutation identification, therefore producing empirical support for departure from Weismann's germ plasm theory in oaks. A shared limitation of both studies is that the authors have selected a single variant caller, without having investigated beforehand the robustness of the results from the selected method. The absence of a simulation work to identify the best suited detection method prior to the empirical investigations therefore represents a major limit with regards to the accuracy and completeness of the previously reported *de novo* mutations.

The development of tools to detect mutations in humans is rapidly expanding in cancer research (Kim *et al.,* 2018; Alioto *et al.,* 2015). Detecting mutations in cancers is conceptually similar to detecting somatic mutations in plants, *i.e.*, the aim is to detect mutations that potentially affect only a small fraction of the sequenced tissue. This specific

challenge is poorly addressed in plants, where mutation detection remains based on generic variant callers, which were initially designed to detect heterozygous sites, which have an expected frequency of 0.5 (Schmid-Siegert *et al.,* 2017; Watson *et al.,* 2016; Hanlon *et al.,* 2019; Orr *et al.,* 2019). Generic variant callers primarily detect candidate mutations per sample against the reference genome and validate mutation robustness by comparing results between sample pairs, while cancer callers identify mutations by comparing two samples, one mutated and one normal sample, against the reference genome (Fig. 1). The per-sample strategy used in generic variant callers carries the risk of overlooking low-frequency mutated reads in one or more samples that should invalidate the mutation in the other sample, whereas the consideration of paired samples in cancer variant callers instead better addresses low allelic frequency mutations in one or both samples. Transferring mutation detection tools from cancers to plants requires evaluating their performance in a plant research context. Cancer research frequently uses very high sequencing depths (100X - 1000X), while the depth available for plants is often considerably lower (e.g., 34X for Hanlon *et al.,* 2019; 40X for Wang *et al.,* 2018; or 70X for Schmid-Siegert *et al.,* 2017), bar a few exceptions (240X for Orr *et al.,* 2019; 250X for Plomion *et al.,* 2018; or 1000X for Watson *et al.,* 2016). To improve the detection of mutations for basic and applied plant research, a deep evaluation of the performance of variant callers is needed in relation to the biological features and quality of data typical of plant studies.

Here, we performed both an *in silico* and an empirical data-based evaluation of the performance of variant callers to detect somatic mutations, using two large published datasets on the same species (pedunculate oak, *Quercus robur*) that applied different

98 strategies for sequencing depth and mutation detection (Schmid-Siegert *et al.,* 2017;
99 Plomion *et al.,* 2018; see Fig. S1). We particularly explored the recall and precision rates
100 depending on the sequencing depth and allelic frequency of the somatic mutation in tissues
101 to answer the following questions: (1) Can cancer research methods, both in terms of
102 protocols (*i.e.* sequencing depth) and tools (*i.e.* callers), improve the detection of somatic
103 mutations?; and (2) Can reanalyses of within-individual sequencing data provide new
104 insights regarding plant mutation processes?



105

106 *Figure 1. Generic variant callers (top rows) detect candidate mutations per tissue sample*
107 *(dark green and light green) against the reference genome (blue) and validate the robustness*
108 *of mutations by comparing results between sample pairs, while cancer callers identify*
109 *mutations by comparing two samples, one mutated (tissue A, dark green) and one normal*
110 *(tissue B, light green), against the reference genome (blue). At low sequencing depth (**A** and*
111 ***E**), neither the generic nor the cancer variant callers detect a low (**A**) or high (**E**) frequency*
112 *mutation. At intermediate sequencing depths (**B** and **F**), both generic and cancer variant*
113 *callers detect high-frequency mutations (**F**), but cancer variant callers are expected to be*
114 *better at detecting low-frequency mutations than generic variant callers (**B**), which were*
115 *originally designed to detect the expected high-frequency heterozygous sites. At high*
116 *sequencing depths (**C** and **G**), both the generic and cancer variant callers detect high*
117 *frequency (**C**) and low frequency (**G**) mutations. However, with intermediate sequencing depth*
118 *(**D** and **H**), a poorly represented heterozygous site in one tissue may remain undetected in that*
119 *tissue by the generic caller while it may be detected in the second tissu and thus be considered*
120 *a mutation, resulting in a false positive (**D**). By comparing the two samples together, cancer*
121 *callers will avoid this error (**H**).*

122

Material and methods

Study design

We developed two workflows: 1) to generate Illumina-like sequencing reads including mutations with varying biological and sequencing parameters; and 2) to detect mutations with multiple variant callers (Fig. S1). We used both *singularity* containers (Kurtzer *et al.,* 2017) and the *snakemake* workflow engines (Köster *et al.,* 2012) to build automated, highly reproducible (FAIR), and scalable workflows. We then used both workflows to test the best performing variant caller for mutation detection *in silico* based on biological and sequencing parameters. We finally used the identified variant caller to detect mutations in pedunculate oak, *Quercus robur* L., by re-analysing data from two somatic mutation projects on oaks led by INRA Bordeaux, France (Plomion *et al.,* 2018) and the University of Lausanne, Switzerland (Schmid-Siegert *et al.,* 2017).

Generation of mutations

To ensure the feasibility of the project and to limit the computational load, a first step is to subsample one or several sequences of user-defined length in the reference genome. The first workflow named *generateMutations* therefore uses a bespoke R script named *sample_genome* to generate these subsets. The workflow then takes advantage of the two scripts included in *simuG* (Yue & Liti, 2019), *vcf2model.pl,* and *simuG.pl*, respectively, 1) to build a model of heterozygous sites distribution for an haploid reference genome based on

a user-defined set of known heterozygous sites in *vcf* format and 2) to build the second reference haploid genome comprising a user-defined number of heterozygous sites to accurately represent diploidy. Typically, the user can define a number of heterozygous sites based on the product of nucleotide diversity (π) and genome length (L). The workflow uses a homemade R script named *generate_mutations* to spike randomly the reference genome with a user-defined number of mutations which are drawn in a binomial distribution using a user-defined transition/transversion ratio (**R**). Finally, the workflow takes advantage of *InSilicoSeq* (Gourlé *et al.,* 2019) defined with the model option *hiseq* to generate datasets of mutated and non-mutated *in silico* Illumina-like sequencing reads using (1) the original reference haploid genome; (2) the reference haploid genome with heterozygous sites, as the workflow was developed for a diploid species; and (3) the reference genome spiked with mutations following user-defined allelic fraction (**AF**) and depth of sequencing depth (**C**).

Detection of mutations

The second workflow named *detectMutations* aims to detect somatic mutations from mapped sequencing reads on a genome reference. Pair-end sequencing reads of every library are quality checked using *FastQC* before trimming using *Trimmomatic* (Bolger *et al.,* 2014) keeping only paired-end reads without adaptors and a phred score above 15 in a sliding window of 4 bases. Reads are aligned against the reference per chromosome using BWA *mem* with the option to mark shorter splits (Li & Durbin, 2009). Alignments are then compressed using *Samtools view* in CRAM format, sorted by coordinates using *Samtools sort,* and indexed using *Samtools index* (Li *et al.,* 2009). Duplicated reads in alignments are

marked using *GATK MarkDuplicates* (Auwera *et al.,* 2013). Finally, the workflow uses seven variant callers to detect mutations, including generic variant callers to detect variants and dedicated variant callers for mutation detection. Generic variant callers to detect variants include *GATK HaplotypeCaller* with *GATK GenotypeGVCFs* (Auwera *et al.,* 2013) and *freebayes* (Garrison & Marth, 2012) using and reporting genotype qualities, without priors on allele balance, with a minimum alternate allele fraction of 0.03, a minimum repeated entropy of 1 and a minimum alternate allele count of 2. Cancer variant callers developed for mutation detection include *VarScan* (Koboldt *et al.,* 2009), *Strelka2* (Kim *et al.,* 2018), *MuSE* (Fan *et al.,* 2016), *Mutect2* (using a panel of normal and without soft clipped bases; Benjamin *et al.,* 2019), and *Somatic Sniper* (filtering reads with mapping quality less than 25, filtering mutations with quality less than 15 with prior probability of a mutation of 0.0001; Larson *et al.,* 2012). Then we only focused on the simulated mutations, and therefore excluded from the analyses the known heterozygous sites provided by the user thanks to the vcf file for *GATK*, *freebayes*, *Somatic Sniper*, and *Strelka2* using *BEDTools subtract* (Quinlan & Hall, 2010) or directly within the variant caller for *Mutect2* and *VarScan*.

In silico experiment

We used the *generateMutations* workflow to generate 1000 mutations in the oak genome with varying biological and sequencing parameters. To ensure consistency between the *in silico* experiment and the reanalysis of empirical data, we used the reference genome "Qrob_PM1N'' of Quercus robur 3P from Bordeaux, ENA accession number PRJEB8388

(Plomion *et al.,* 2018), thus assessing the behaviour of variant callers in the same genomic context as used for the empirical work. To reduce the computational load, we only generated mutations on the first megabase of the first chromosome of the oak assembly ("Qrob_Chr01") in order to later focus the detection on this region. To check that the conclusions regarding the callers are independent of the considered genomic region, we ran five independent investigations based on randomly selected genome areas of a megabase in length. Our results were highly congruent over all our investigations (Pearson's correlations, recall: 0.999, precision: 0.947). We used known heterozygous sites from the reference genome (Plomion *et al.,* 2018) to simulate back one thousand heterozygous sites (π = 0.01 , L = 1 Mb , N = π x L = 10^4). We used varying values of transition/transversion ratio (**R** = [2, 2.5, 3]), allelic fraction (**AF** = [0.05, 0.1, 0.25, 0.5]), and sequencing depth (**C** = [25, 50, 100, 150, 200]), resulting in 60 simulated datasets of mutated and associated base reads (3**R** x 4**AF** x 5**C**). We then used the *detectMutations* workflow to detect (recall) spiked mutations with every variant caller (*Mutect2, freebayes, GATK, Strelka2, VarScan, Somatic Sniper,* and *MuSe*). Using known spiked mutations, we assessed the number of true positive (TP), false positive (FP), and false negative (FN) for each variant caller to detect mutations and each combination of biological and sequencing parameters. We used the resulting confusion matrix to calculate the recall ($\frac{TP}{TP+FN}$) and the precision rates ($\frac{TP}{TP+FP}$). The recall rate represents the ability of the variant caller to detect all mutations, while the precision rate represents the ability of the variant caller to not confound other sites with mutations. We finally assessed each variant caller to detect mutations using the recall and the precision rates with varying transition/transversion

ratio (**R**), allelic fraction (**AF**), and sequencing depth (**C**) to identify the best performing variant caller based on biological and sequencing parameters.

Oak data reanalyses

We re-analyzed publicly available oak data from two projects led by Bordeaux, France (Plomion *et al.,* 2018) and Lausanne, Switzerland (Schmid-Siegert *et al.,* 2017) (SRA PRJNA327502 and ENA PRJEB8388, respectively). We then used the best-performing variant caller based on our *in silico* investigation, *Strelka2*, and the variant caller for mutation detection from the original publication to compare the results, *i.e.*, *GATK* with *Best Practices* for Swiss data (Schmid-Siegert *et al.,* 2017) and *Mutect2* for French data (Plomion *et al.,* 2018). The Swiss data comprised 2 libraries of medium sequencing depth (60X) representing one lower and one upper branch. The French data comprised 3 libraries of high sequencing depth (160X) representing 3 branches (lower, mid, and upper). For both Swiss and French data, we compared each pair of sample points sequentially as the reference library and the potentially mutated library to distinguish mutations among branches from heterozygous sites and sequencing errors. For the French data, we further filtered out candidate somatic mutations by using a cross-validation procedure to keep a coherent temporal pattern among mutations following the original publication (Plomion *et al.,* 2018). Contrary to a general expectation and a common view in the field (Schmid-Siegert *et al.*, 2017, Orr *et al.*, 2019), detected mutations do not always accumulate following the developing plant architecture (Zahradníková *et al.*, 2020). As a consequence, our cross-validation represents a conservative strategy for the mutation detection, but it should be

noted that this strategy could have removed some true somatic mutations. We used these raw datasets to identify the mutations from the original studies after realigning the megabase containing the mutation on the 3P genome using *BLAT* (Kent, 2002). For both datasets, we finally kept candidate mutations with (1) a read depth for both the normal and mutated samples between half and two times the mean sequencing depth (30-120X and 80-320X for Swiss and French datasets, respectively), (2) an absence of the mutated allele in the normal sample, (3) a minimum of 10 copies of the mutated allele in the mutated sample and (4) an allelic frequency <0.5. In addition, *Strelka2* calculates an empirical variant score (EVS) based on a supervised random forest classifier trained on data from sequencing runs under various conditions, which provides an overall quality score for each variant (Kim *et al.*, 2018). We took advantage of the EVS to define a conservative set of candidate mutations for both datasets, hereafter referred to as the EVS datasets. Given that the proportion of the genome falling within the sequencing depth boundaries used for the detection (i.e. between 50 and 200% of the mean sequencing depth) varies depending on the dataset, we weighted the observed number of mutations by the proportion of the genome satisfying the sequencing depth criteria to provide a more accurate and comparable estimate of the real total number of mutations. Across both empirical studies, the proportion of the genome with 50-200% sequencing depth varies between 71 and 87%, therefore the impact of the weighting in the estimation of the real total number of mutations is low.

Results

To compare the performance of different variant callers to detect mutations, we simulated sequencing data containing new mutations at a given allele frequency (fraction of simulated reads per genomic position carrying the mutated allele), and using varying depths of sequencing (for variable transition/transversion ratios, see Supplementary Note S1). We then evaluated the performance of variant callers as a function of allele frequency and sequencing depth. We found marked differences in: (1) the recall, the ability to recover the simulated mutations; and (2) the precision, the proportion of truly simulated mutations among all variants detected. For allele frequencies equal to, or lower than, 0.25, cancer-specific variant callers (*Strelka2, Mutect2, MuSE,* but not *Somatic Sniper*) outperform generic variant callers such as *GATK*, *freebayes*, and *VarScan* (Fig. 2). For allele frequencies over 0.25, all variant callers perform similarly well, except for *freebayes*, which identified many false positives. Over the 80 tested parameter combinations, *Strelka2* was the best performing variant caller for various allelic frequencies and sequencing depths (in 57/80 simulated datasets , with an average recall of 0.95 for a precision of 0.98, Fig. S2-4 and Table S1 and S2) and the second fastest caller (Fig. S9).

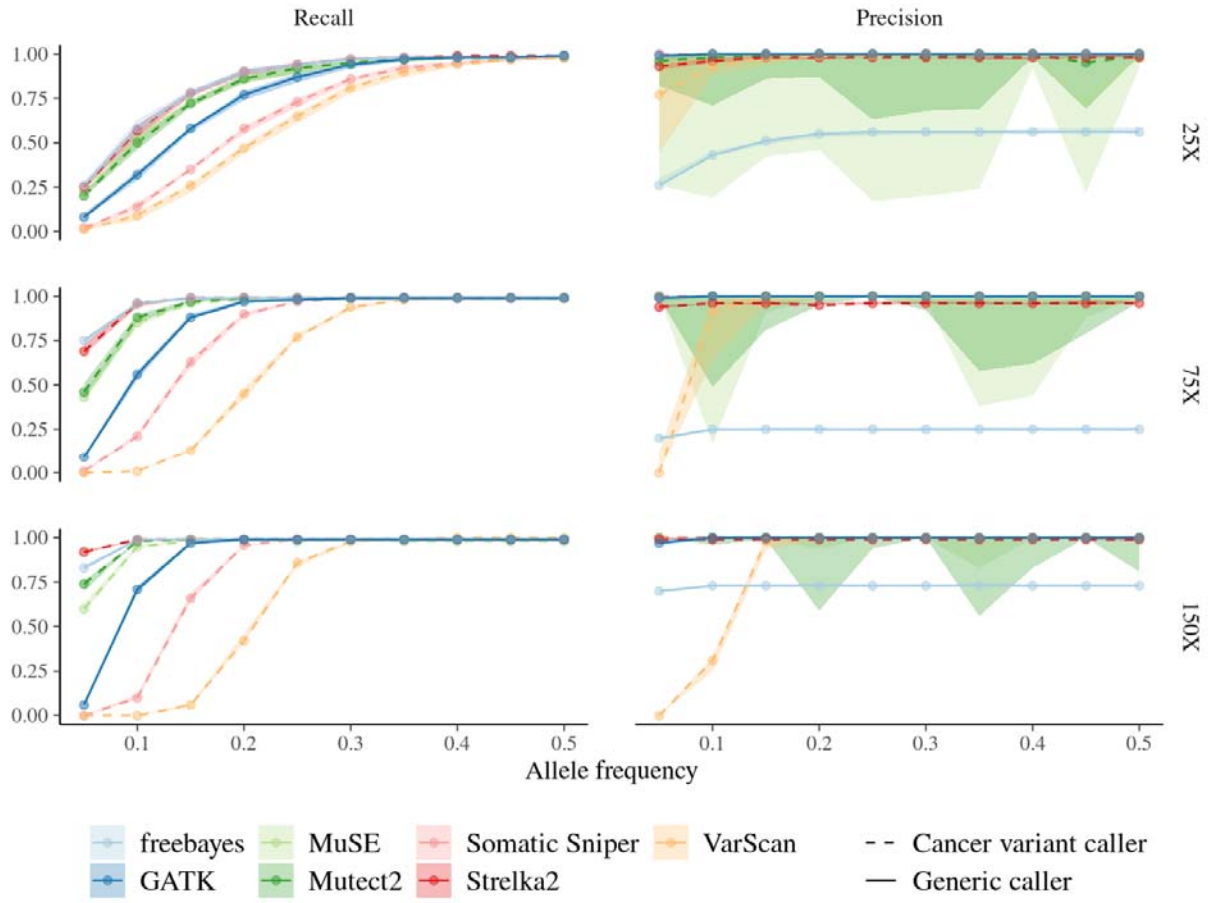
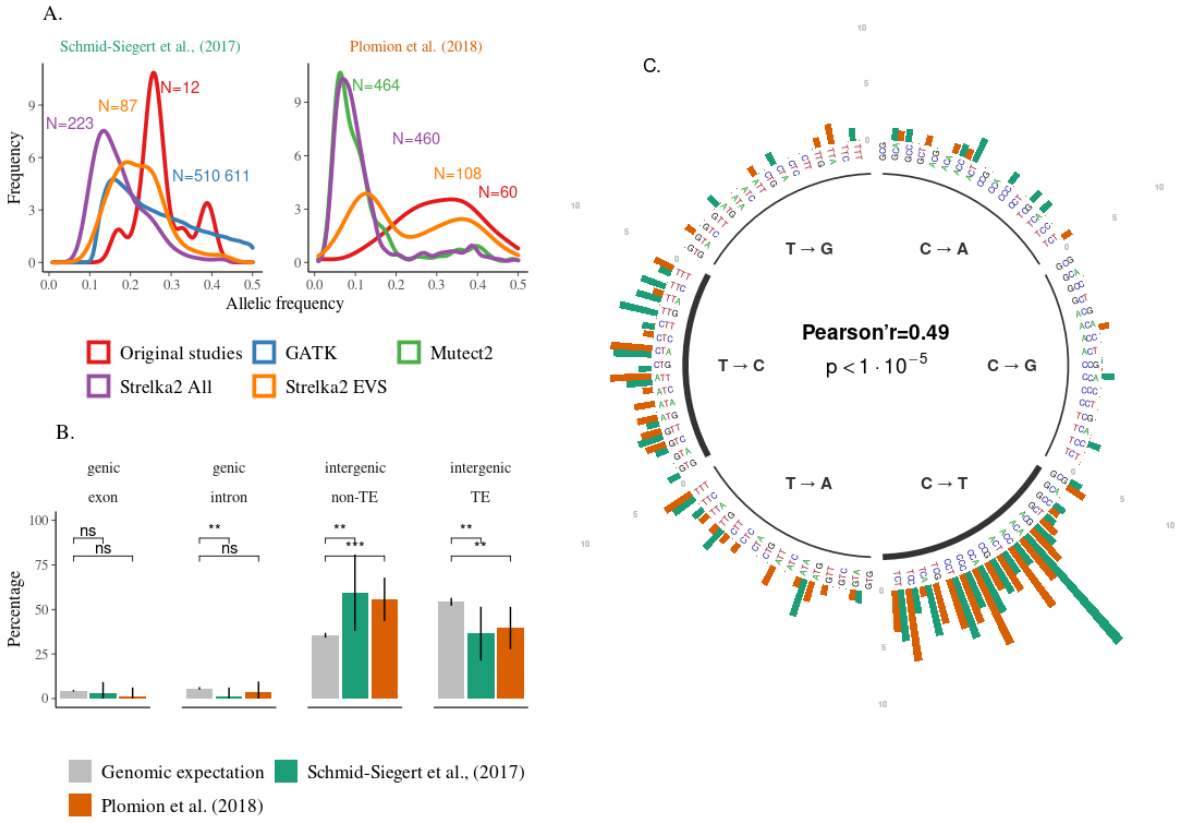


Figure 2: Variant caller performances to identify simulated mutations for varying allelic frequencies and sequencing depths (see Fig. S5 for all parameter combinations). The recall is the ability to detect (recover) the simulated mutations. The precision is the proportion of simulated mutations among all variants detected (i.e. including false positives). Each point represents the averaged mutation recall or precision (10 simulations) for increasing allelic frequency and sequencing depth. The shaded area represents the variation of recall and precision rates over the 10 replicates computed for all callers, but only visible for the precision of Muse, Mutect2, and VarScan. Linetype opposes generic callers (dashed) against cancer variant callers developed for cancer research (solid).

We further investigated the performance of the best performing variant caller, *Strelka2*, on two empirical datasets on pedunculate oak (Schmid-Siegert *et al.,* 2017; Plomion *et al.,* 2018) in comparison to the variant callers used in the original publications, *i.e.*, *GATK* and *Mutect2*, respectively (see Supplementary Note S2). Mapping the raw data of Schmid-Siegert *et al.,* (2017) and Plomion *et al.,* (2018) on the oak genome that we used as a mapping reference for our empirical study, we successfully mapped 14 and 60 of the mutations detected in the original articles, respectively. Across variant callers, we recovered 12 (86%) and 60 (100%) of these original mutations in our total list of candidate somatic mutations (Fig. 3A), strongly supporting the results shown by the two previous studies. However, our analyses were able to detect far more robust candidate mutations than initially reported. Applying filtering based on sequencing depth and mutated allele copies (see Supplementary Note S2), *Strelka2* produced a smaller set of candidate mutations than *GATK* but similar to *Mutect2*, with an estimated number of mutation candidates 10- to 25-fold higher than that of the original studies (Fig. 3A). Adding *Strelka2* recommended filtering based on empirical variant scores yielded the most conservative dataset with a similar number of mutations between both studies and a 2 to 7-fold increase compared to the original number of mutations. Due to lack of access to biological material from the original studies, conclusions were drawn from this list of conservative candidate somatic mutations (but see supplementary note S4 for a discussion regarding validation of mutations). The distribution of allelic frequencies of detected mutations partly explains differences among detection methods (Fig. 3A), with *Strelka2* and *Mutect2* detecting mutations with lower allelic frequencies than the candidate mutations presented in the

294 original publications, especially for the Plomion *et al.,* (2018) study that used higher
295 sequencing depths.

296 Based on the set of conservative mutations detected by *Strelka2* (EVS), we then explored
297 annotations and mutation spectra in both datasets (Fig. 3B-C), which have rarely been
298 explored in model plant species (but see first evidence based on mutation accumulation
299 lines in Arabidopsis in Weng *et al.,* 2019) and never in the wild. The proportions of
300 mutations found in different genomic regions (e.g. genic, intergenic) were highly correlated
301 between both original studies and proportional to the representation of the genomic
302 regions, supporting a random distribution of mutations throughout the genome (Fig. 3B).
303 Mutation spectra of the two studies are significantly correlated (Pearson'r=0.49, $p<7.4\ 10$-
304 5), with an enrichment in C>T transitions, particularly in some specific genomic contexts
305 (Fig. 3C).



306

307 *Figure 3: Candidate mutation spectra depending on variant callers and filtering in Schmid-*
308 *Siegert et al., (2017) and Plomion et al., (2018).* ***A.*** *Allelic frequency distribution for every*
309 *dataset, including the candidate mutations from the original article present in the raw data*
310 *from the reanalysis (red), the results of GATK with Best Practices (blue), Mutect2 after*
311 *filtering (green), and Strelka2 after filtering (purple), and the results of Strelka2 using the*
312 *filtering based on empirical variant scores named EVS (orange). The labels indicate the*
313 *number of candidate mutations in each dataset. Per caller comparisons are available in Fig.*
314 *S7.* ***B.*** *Annotation of the mutations detected by Strelka2 across chromosomes using the*
315 *filtering based on empirical variant scores named EVS for Schmid-Siegert et al., (2017, green)*
316 *and Plomion et al., (2018, orange) compared to the genomic expectation (grey, see*
317 *Supplementary Note S3). Error bars represent the standard deviation (SD) of the observed*

*percentages across chromosomes, and the annotation above the columns indicates the significance of the Student's t-test two-sided comparing the mean percentage of mutations to the mean genomic expectation, with ns, **, and *** corresponding to non-significant, p<0.01, and p<0.001 differences, respectively.* ***C.*** *Context-dependent mutation spectra depending on mutation types for the results of Strelka2 using the filtering based on empirical variant scores named EVS. Mutation types have been summarised into six main classes with thicker lines for transversion compared to transition, and then differentiated depending on their 5' and 3' genomic contexts, see Fig. S8-9. Pearson's correlation r measures the two-sided correlation of the mutation spectra between Schmid-Siegert et al., (2017) and Plomion et al., (2018).*

Discussion

Mutation research in plants still primarily uses generic variant callers and methodologies that are not developed for the specificity and complexity of within-individual *de novo* mutation detection. We examined if plant mutation research could benefit from the development of tools and protocols initially designed for human cancer research, which is a rapidly expanding field (Kim *et al.*, 2018). We found marked differences in the performance of variant callers for mutation detection based on sequencing depth and allelic frequency. We found that cancer variant callers performed better than generic variant callers for mutation detection at low or intermediate allelic frequency or with low sequencing depth, and similarly well for high allelic frequency. Low allelic frequency mutations, potentially due to the chimeric nature of plant shoot apical meristems structures (Burian, 2021), might be very important due to their great abundance that may balance out their low chance of

transmissions. Therefore, plant mutation studies should make greater use of cancer variant callers such as *Strelka2* rather than generic variant callers such as *GATK* to detect somatic mutations, in agreement with previous studies on germline mutations detection (Chen *et al.*, 2019), especially for detecting low frequency mutations and when using low sequencing depth. The importance of allele frequency-dependency in variant detection is not restricted to somatic mutations, but also concerns for instance polyploid species, which includes many agriculturally important autopolyploid plant species (e.g. potato, sugarcane). Our simulation framework therefore provides general insights regarding the impact of allelic dosage in mutation detection which go beyond somatic mutation detection.

One problem that may arise when analysing sample pairs with cancer variant callers is the rapid increase in pairwise comparisons when using a larger sample size than previous studies (e.g., N=3 in Plomion *et al.* 2018). A simple solution is the use of a single reference sample such as a cambium sample from the base of the tree, which is therefore considered as the closest genome to the seed, to compare it to all samples from branches (Hanlon *et al.*, 2019). By reanalyzing the raw oak data (Schmid-Siegert *et al.*, 2017; Plomion *et al.*, 2018), we found that the marked differences in the performance of variant callers could account for the discrepancies in genome-wide plant somatic mutation rate estimates. Our reanalysis shows robust evidence for an up to 7-fold higher number of mutations than previously reported, a value closer to the expectations based on the theory (Schoen & Schultz, 2019; Burian, 2021). We argue that knowledge and methodological transfers from cancer to plant mutation detection are expected to contribute strongly to the upward trend of this field and to reconcile empirical reports with theoretical expectations.

Acknowledgments

The manuscript benefited from the comments of three anonymous reviewers.

Funding

This study was funded through an Investissement d'Avenir grant of the ANR: CEBA (ANR-10-LABEX-0025).

Conflict of interest disclosure

The authors declare they have no conflict of interest relating to the content of this article. MH is a recommender for PCI Evol Biol.

Data, script and code availability

Reanalyzed reads and corresponding genomes were extracted from GenBank under accession BioProject PRJNA327502 and from European Nucleotide Archive under project accession code PRJEB8388. `generateMutations` and `detectMutations` pipelines are available on GitHub (https://github.com/sylvainschmitt/generateMutations and https://github.com/sylvainschmitt/detectMutations).

Authors' contributions

All authors conceived the ideas; SS developed the pipelines, conducted the virtual experiment and the data reanalyses; SS analysed outputs and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck, T.A., Simpson, J.T., Tonon, L., Sertier, A.S., Patch, A.M., Jäger, N., Ginsbach, P., Drews, R., Paramasivam, N., Kabbe, R., Chotewutmontri, S., Diessl, N., Previti, C., Schmidt, S., Brors, B., Feuerbach, L., Heinold, M., Gröbner, S., Korshunov, A., Tarpey, P.S., Butler, A.P., Hinton, J., Jones, D., Menzies, A., Raine, K., Shepherd, R., Stebbings, L., Teague, J.W., Ribeca, P., Giner, F.C., Beltran, S., Raineri, E., Dabad, M., Heath, S.C., Gut, M., Denroche, R.E., Harding, N.J., Yamaguchi, T.N., Fujimoto, A., Nakagawa, H., Quesada, V., Valdés-Mas, R., Nakken, S., Vodák, D., Bower, L., Lynch, A.G., Anderson, C.L., Waddell, N., Pearson, J.V., Grimmond, S.M., Peto, M., Spellman, P., He, M., Kandoth, C., Lee, S., Zhang, J., Létourneau, L., Ma, S., Seth, S., Torrents, D., Xi, L., Wheeler, D.A., López-Otín, C., Campo, E., Campbell, P.J., Boutros, P.C., Puente, X.S., Gerhard, D.S., Pfister, S.M., McPherson, J.D., Hudson, T.J., Schlesner, M., Lichter, P., Eils, R., Jones, D.T.W. & Gut, I.G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, **6**.

394 Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Angel, G. del, Levy-Moonshine, A., Jordan,
395 T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S. &
396 DePristo, M.A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome
397 Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, **43**, 483–492.
398 Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1110s43

399 Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C. & Lichtenstein, L. (2019). Calling
400 Somatic SNVs and Indels with Mutect2.

401 Bolger, A.M., Lohse, M. & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina
402 sequence data. *Bioinformatics*, **30**, 2114–2120.

403 Burian, A. (2021). Does Shoot Apical Meristem Function as the Germline in Safeguarding
404 Against Excess of Mutations? *Frontiers in Plant Science*, **12**, 1–9.

405 Chen, Z.L., Meng, J.M., Cao, Y., Yin, J.L., Fang, R.Q., Fan, S.B., Liu, C., Zeng, W.F., Ding, Y.H., Tan,
406 D., Wu, L., Zhou, W.J., Chi, H., Sun, R.X., Dong, M.Q. & He, S.M. (2019). A high-speed search
407 engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked
408 peptides. *Nature Communications*, **10**. Retrieved from http://dx.doi.org/10.1038/s41467-
409 019-11337-z

410 Fan, Y., Xi, L., Hughes, D.S.T., Zhang, J., Zhang, J., Futreal, P.A., Wheeler, D.A. & Wang, W.
411 (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model

412 improves sensitivity and specificity in mutation calling from sequencing data. *Genome*
413 *biology*, **17**, 178. Retrieved from http://dx.doi.org/10.1186/s13059-016-1029-6

414 Gao, Z., Moorjani, P., Sasani, T.A., Pedersen, B.S., Quinlan, A.R., Jorde, L.B., Amster, G. &
415 Przeworski, M. (2019). Overlooked roles of DNA damage and maternal age in generating
416 human germline mutations. *Proceedings of the National Academy of Sciences of the United*
417 *States of America*, **116**, 9491–9500.

418 Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read
419 sequencing. 1–9. Retrieved from http://arxiv.org/abs/1207.3907

420 Gourlé, H., Karlsson-Lindsjö, O., Hayer, J. & Bongcam-Rudloff, E. (2019). Simulating Illumina
421 metagenomic data with InSilicoSeq. *Bioinformatics*, **35**, 521–522.

422 Hanlon, V.C.T., Otto, S.P. & Aitken, S.N. (2019). Somatic mutations substantially increase the
423 per-generation mutation rate in the conifer Picea sitchensis. *Evolution Letters*, **3**, 348–358.

424 Harris, K. & Pritchard, J.K. (2017). Rapid evolution of the human mutation spectrum. *eLife*,
425 **6**, 1–17.

426 Kent, W.J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, **12**, 656–664.

427 Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y.,
428 Beyter, D., Krusche, P. & Saunders, C.T. (2018). Strelka2: fast and accurate calling of
429 germline and somatic variants. *Nature Methods*, **15**, 591–594.

430 Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M.,
431 Wilson, R.K. & Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing
432 of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.

433 Köster, J. & Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine.
434 *Bioinformatics*, **28**, 2520–2522.

435 Kurtzer, G.M., Sochat, V. & Bauer, M.W. (2017). Singularity: Scientific containers for mobility
436 of compute. *PLoS ONE*, **12**, 1–20.

437 Lanfear, R. (2018). Do plants have a segregated germline? *PLoS Biology*, **16**, 1–13.

438 Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis,
439 E.R., Wilson, R.K. & Ding, L. (2012). Somaticsniper: Identification of somatic point mutations
440 in whole genome sequencing data. *Bioinformatics*, **28**, 311–317.

441 Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
442 transform. *Bioinformatics*, **25**, 1754–1760.

443 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. &
444 Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**,
445 2078–2079.

446 Lynch, M., Ackerman, M.S., Gout, J.F., Long, H., Sung, W., Thomas, W.K. & Foster, P.L. (2016).
447 Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, **17**,
448 704–714. Retrieved from http://dx.doi.org/10.1038/nrg.2016.104

449 Martincorena, I. & Campbell, P.J. (2015). Somatic mutation in cancer and normal cells.
450 *Science*, **349**, 1483–1489.

451 Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y. & Vijg, J. (2017). Differences between
452 germline and somatic mutation rates in humans and mice. *Nature Communications*, **8**, 1–8.
453 Retrieved from http://dx.doi.org/10.1038/ncomms15183

454 Orr, A.J., Padovan, A., Kainer, D., Külheim, C., Bromham, L., Bustos-Segura, C., Foley, W., Haff,
455 T., Hsieh, J.F., Morales-Suarez, A., Cartwright, R.A. & Lanfear, R. (2019). A phylogenomic
456 approach reveals a low somatic mutation rate in a long-lived plant. *bioRxiv*.

457 Plomion, C., Aury, J.M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., Faye, S., Francillonne,
458 N., Labadie, K., Le Provost, G., Lesur, I., Bartholomé, J., Faivre-Rampant, P., Kohler, A., Leplé,
459 J.C., Chantret, N., Chen, J., Diévart, A., Alaeitabar, T., Barbe, V., Belser, C., Bergès, H., Bodénès,
460 C., Bogeat-Triboulot, M.B., Bouffaud, M.L., Brachi, B., Chancerel, E., Cohen, D., Couloux, A., Da
461 Silva, C., Dossat, C., Ehrenmann, F., Gaspin, C., Grima-Pettenati, J., Guichoux, E., Hecker, A.,

462 Herrmann, S., Hugueney, P., Hummel, I., Klopp, C., Lalanne, C., Lascoux, M., Lasserre, E.,
463 Lemainque, A., Desprez-Loustau, M.L., Luyten, I., Madoui, M.A., Mangenot, S., Marchal, C.,
464 Maumus, F., Mercier, J., Michotey, C., Panaud, O., Picault, N., Rouhier, N., Rué, O., Rustenholz,
465 C., Salin, F., Soler, M., Tarkka, M., Velt, A., Zanne, A.E., Martin, F., Wincker, P., Quesneville, H.,
466 Kremer, A. & Salse, J. (2018). Oak genome reveals facets of long lifespan. *Nature Plants*, **4**,
467 440–452. Retrieved from http://dx.doi.org/10.1038/s41477-018-0172-3

468 Prevost, L., Knight, J., Smith, M., & Lurain, U. M. (2013). Student writing reveals their
469 heterogeneous thinking about the origin of genetic variation in populations.

470 Quinlan, A.R. & Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing
471 genomic features. *Bioinformatics*, **26**, 841–842.

472 Schmid-Siegert, E., Sarkar, N., Iseli, C., Calderon, S., Gouhier-Darimont, C., Chrast, J., Cattaneo,
473 P., Schütz, F., Farinelli, L., Pagni, M., Schneider, M., Voumard, J., Jaboyedoff, M., Fankhauser,
474 C., Hardtke, C.S., Keller, L., Pannell, J.R., Reymond, A., Robinson-Rechavi, M., Xenarios, I. &
475 Reymond, P. (2017). Low number of fixed somatic mutations in a long-lived oak tree.
476 *Nature Plants*, **3**, 926–929. Retrieved from http://dx.doi.org/10.1038/s41477-017-0066-9

477 Schoen, D.J. & Schultz, S.T. (2019). Somatic Mutation and Evolution in Plants. *Annual Review*
478 *of Ecology, Evolution, and Systematics*, **50**, 49–73.

479 Schöngart, J., Bräuning, A., Barbosa, A.C.M.C., Lisi, C.S. & Oliveira, J.M. de. (2017).
480 *Dendroecology Tree-Ring Analyses Applied to Ecological Studies*.

481 Smith, M. K., & Knight, J. K. (2012). Using the Genetics Concept Assessment to document
482 persistent conceptual difficulties in undergraduate genetics courses. Genetics, 191(1), 21–
483 32. https://doi.org/10.1534/genetics.111.137810

484 Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., Zhang, X., Zhao, L., Zhang, Y., Jia, Y., Qin, C., Yu,
485 L., Huang, J., Yang, S., Hurst, L.D. & Tian, D. (2019). The architecture of intra-organism
486 mutation rate variation in plants. *PLoS Biology*, **17**, 1–29.

487 Watson, J.M., Platzer, A., Kazda, A., Akimcheva, S., Valuchova, S., Nizhynska, V., Nordborg, M.
488 & Riha, K. (2016). Germline replications and somatic mutation accumulation are
489 independent of vegetative life span in Arabidopsis. *Proceedings of the National Academy of*
490 *Sciences of the United States of America*, **113**, 12226–12231.

491 Weismann, A. (1893). *The germ-plasm: a theory of heredity*. Scribner's.

492 Weng, M.L., Becker, C., Hildebrandt, J., Neumann, M., Rutter, M.T., Shaw, R.G., Weigel, D. &
493 Fenster, C.B. (2019). Fine-grained analysis of spontaneous mutation spectrum and
494 frequency in arabidopsis thaliana. *Genetics*, **211**, 703–714.

495 Yue, J.X. & Liti, G. (2019). SimuG: A general-purpose genome simulator. *Bioinformatics*, **35**,
496 4442–4444.

497 Zahradníková, E., Ficek, A., Brejová, B., Vinař, T., & Mičieta, K. (2020). Mosaicism in old trees
498 and its patterns. *Trees - Structure and Function,* **34**, 357–370.
499 https://doi.org/10.1007/s00468-019-01921-7

500

Supplementary material

The following Supporting Information is available for this article:

Note S1. Simulation results

Note S2. Reanalyses results

Note S3. Genomic expectations for the annotation of the mutations

Note S4. 'validation' of mutations

Fig. S1. Study scheme

Fig. S2. Variation in the performance of variant callers for mutation detection with varying biological and sequencing parameters

Fig. S3. Variation in the performance of variant callers for mutation detection with varying biological and sequencing parameters

Fig. S4. Best performing variant callers for mutation detection depending on allelic fraction (allelic frequency) and coverage (sequencing depth)

Fig. S5. Mutation recall and precision rates for generic and mutation-specific variant callers by allelic fraction and sequencing depth

Fig. S6. Observed allelic frequencies of candidate mutations depending on variant callers and filtering in Schmid-Siegert *et al.,* (2017) and Plomion *et al.,* (2018)

518 **Fig. S7.** Percentage of nucleotide change types of candidate mutations depending on variant
519 callers and filtering in Schmid-Siegert *et al.,* (2017) and Plomion *et al.,* (2018)

520 **Fig. S8.** Context-dependent mutation spectrum depending on variant callers and filtering in
521 Schmid-Siegert *et al.,* (2017) and Plomion *et al.,* (2018)

522 **Tab. S1.** Mean and standard deviation in performance of variant callers for mutation
523 detection across all simulations

524 **Tab S2.** Mean and standard deviation in performance of variant callers for mutation
525 detection with varying allelic frequency and sequencing depth

526