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.

33 ## Introduction

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

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

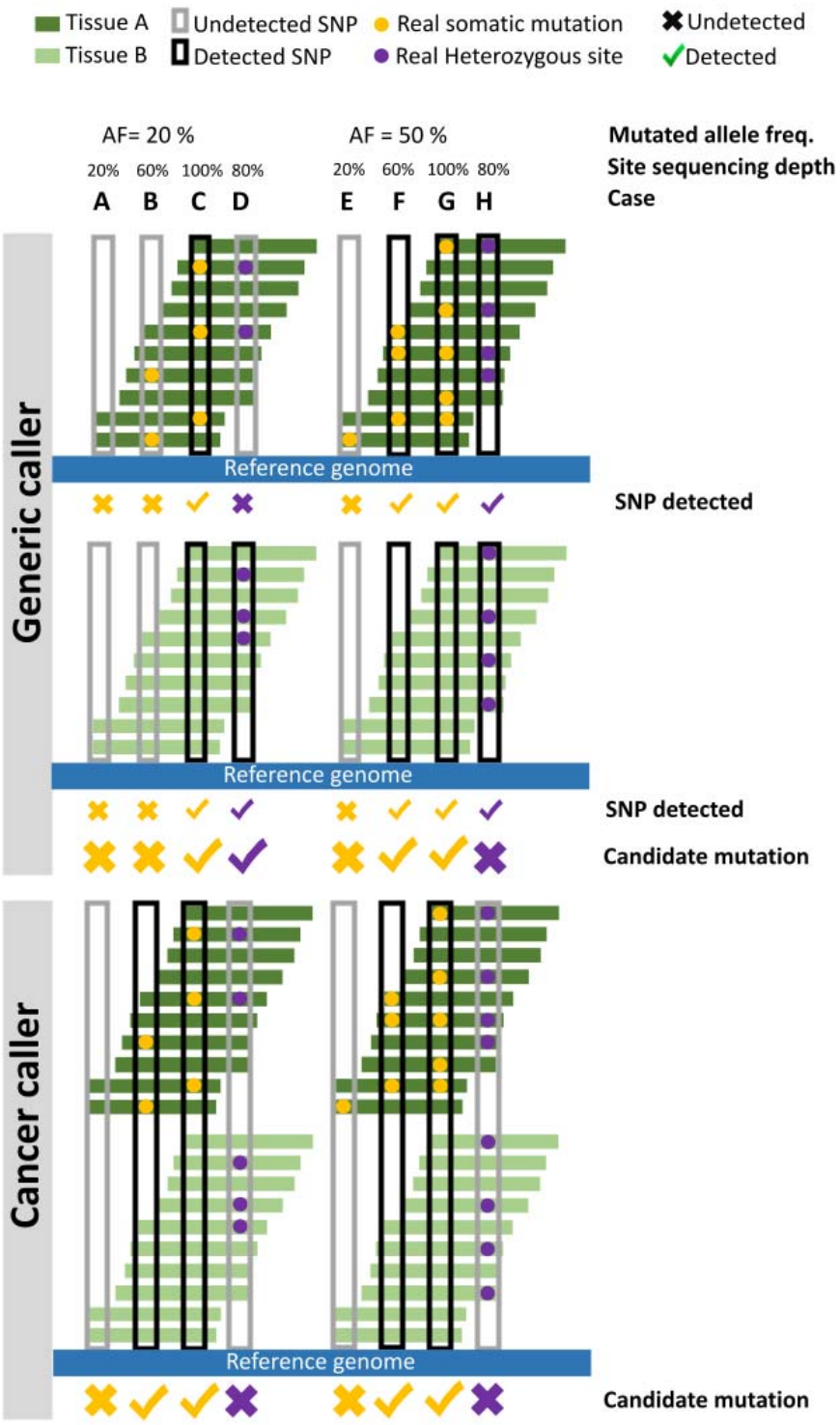
54 per generation (Hanlon *et al.,* 2019). To generate new knowledge on plant mutation
55 processes[9], several studies examined within-individual variation in long-lived trees, whose
56 individuals can live for more than a thousand years (Schöngart *et al.,* 2017). Two studies
57 used the pedunculate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

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

76 challenge is poorly addressed in plants, where mutation detection remains based on generic
77 variant callers, which were initially designed to detect heterozygous sites, which have an
78 expected frequency of 0.5 (Schmid-Siegert *et al.,* 2017; Watson *et al.,* 2016; Hanlon *et al.,*
79 2019; Orr *et al.,* 2019). Generic variant callers primarily detect candidate mutations per
80 sample against the reference genome and validate mutation robustness by comparing
81 results between sample pairs, while cancer callers identify mutations by comparing two
82 samples, one mutated and one normal sample, against the reference genome (Fig. 1). The
83 per-sample strategy used in generic variant callers carries the risk of overlooking low-
84 frequency mutated reads in one or more samples that should invalidate the mutation in the
85 other sample, whereas the consideration of paired samples in cancer variant callers instead
86 better addresses low allelic frequency mutations in one or both samples. Transferring
87 mutation detection tools from cancers to plants requires evaluating their performance in a
88 plant research context. Cancer research frequently uses very high sequencing depths (100X
89 - 1000X), while the depth available for plants is often considerably lower (e.g., 34X for
90 Hanlon *et al.,* 2019; 40X for Wang *et al.,* 2018; or 70X for Schmid-Siegert *et al.,* 2017), bar a
91 few exceptions (240X for Orr *et al.,* 2019; 250X for Plomion *et al.,* 2018; or 1000X for
92 Watson *et al.,* 2016). To improve the detection of mutations for basic and applied plant
93 research, a deep evaluation of the performance of variant callers is needed in relation to the
94 biological features and quality of data typical of plant studies.

95 Here, we performed both an *in silico* and an empirical data-based evaluation of the
96 performance of variant callers to detect somatic mutations, using two large published
97 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?


Tissue A
Tissue B
Undetected SNP
Detected SNP
Real somatic mutation
Real Heterozygous site
Undetected
Detected
AF= 20 %
AF = 50 %
Mutated allele freq.
Site sequencing depth
Case
20% 60% 100% 80%
20% 60% 100% 80%
A B C D
E F G H
Generic caller
Reference genome
SNP detected
Reference genome
SNP detected
Candidate mutation
Cancer caller
Reference genome
Candidate mutation

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

123 # Material and methods

124 ## Study design

125 We developed two workflows: 1) to generate Illumina-like sequencing reads including
126 mutations with varying biological and sequencing parameters; and 2) to detect mutations
127 with multiple variant callers (Fig. S1). We used both *singularity* containers (Kurtzer *et al.,*
128 2017) and the *snakemake* workflow engines (Köster *et al.,* 2012) to build automated, highly
129 reproducible (FAIR), and scalable workflows. We then used both workflows to test the best
130 performing variant caller for mutation detection *in silico* based on biological and
131 sequencing parameters. We finally used the identified variant caller to detect mutations in
132 pedunculate oak, *Quercus robur* L., by re-analysing data from two somatic mutation projects
133 on oaks led by INRA Bordeaux, France (Plomion *et al.,* 2018) and the University of
134 Lausanne, Switzerland (Schmid-Siegert *et al.,* 2017).

135 ## Generation of mutations

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

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

154 Detection of mutations

155 The second workflow named *detectMutations* aims to detect somatic mutations from
156 mapped sequencing reads on a genome reference. Pair-end sequencing reads of every
157 library are quality checked using *FastQC* before trimming using *Trimmomatic* (Bolger *et al.,*
158 2014) keeping only paired-end reads without adaptors and a phred score above 15 in a
159 sliding window of 4 bases. Reads are aligned against the reference per chromosome using
160 BWA *mem* with the option to mark shorter splits (Li & Durbin, 2009). Alignments are then
161 compressed using *Samtools view* in CRAM format, sorted by coordinates using *Samtools*
162 *sort,* and indexed using *Samtools index* (Li *et al.,* 2009). Duplicated reads in alignments are

163 marked using *GATK MarkDuplicates* (Auwera *et al.,* 2013). Finally, the workflow uses seven
164 variant callers to detect mutations, including generic variant callers to detect variants and
165 dedicated variant callers for mutation detection. Generic variant callers to detect variants
166 include *GATK HaplotypeCaller* with *GATK GenotypeGVCFs* (Auwera *et al.,* 2013) and
167 *freebayes* (Garrison & Marth, 2012) using and reporting genotype qualities, without priors
168 on allele balance, with a minimum alternate allele fraction of 0.03, a minimum repeated
169 entropy of 1 and a minimum alternate allele count of 2. Cancer variant callers developed for
170 mutation detection include *VarScan* (Koboldt *et al.,* 2009), *Strelka2* (Kim *et al.,* 2018), *MuSE*
171 (Fan *et al.,* 2016), *Mutect2* (using a panel of normal and without soft clipped bases;
172 Benjamin *et al.,* 2019), and *Somatic Sniper* (filtering reads with mapping quality less than
173 25, filtering mutations with quality less than 15 with prior probability of a mutation of
174 0.0001; Larson *et al.,* 2012). Then we only focused on the simulated mutations, and
175 therefore excluded from the analyses the known heterozygous sites provided by the user
176 thanks to the vcf file for *GATK*, *freebayes*, *Somatic Sniper*, and *Strelka2* using *BEDTools*
177 *subtract* (Quinlan & Hall, 2010) or directly within the variant caller for *Mutect2* and
178 *VarScan*.

179 *In silico* experiment

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

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

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

208 Oak data reanalyses

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

227 noted that this strategy could have removed some true somatic mutations. We used these
228 raw datasets to identify the mutations from the original studies after realigning the
229 megabase containing the mutation on the 3P genome using *BLAT* (Kent, 2002). For both
230 datasets, we finally kept candidate mutations with (1) a read depth for both the normal and
231 mutated samples between half and two times the mean sequencing depth (30-120X and 80-
232 320X for Swiss and French datasets, respectively), (2) an absence of the mutated allele in
233 the normal sample, (3) a minimum of 10 copies of the mutated allele in the mutated sample
234 and (4) an allelic frequency <0.5. In addition, *Strelka2* calculates an empirical variant score
235 (EVS) based on a supervised random forest classifier trained on data from sequencing runs
236 under various conditions, which provides an overall quality score for each variant (Kim *et*
237 *al.*, 2018). We took advantage of the EVS to define a conservative set of candidate mutations
238 for both datasets, hereafter referred to as the EVS datasets. Given that the proportion of the
239 genome falling within the sequencing depth boundaries used for the detection (i.e. between
240 50 and 200% of the mean sequencing depth) varies depending on the dataset, we weighted
241 the observed number of mutations by the proportion of the genome satisfying the
242 sequencing depth criteria to provide a more accurate and comparable estimate of the real
243 total number of mutations. Across both empirical studies, the proportion of the genome
244 with 50-200% sequencing depth varies between 71 and 87%, therefore the impact of the
245 weighting in the estimation of the real total number of mutations is low.

246 ## Results

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

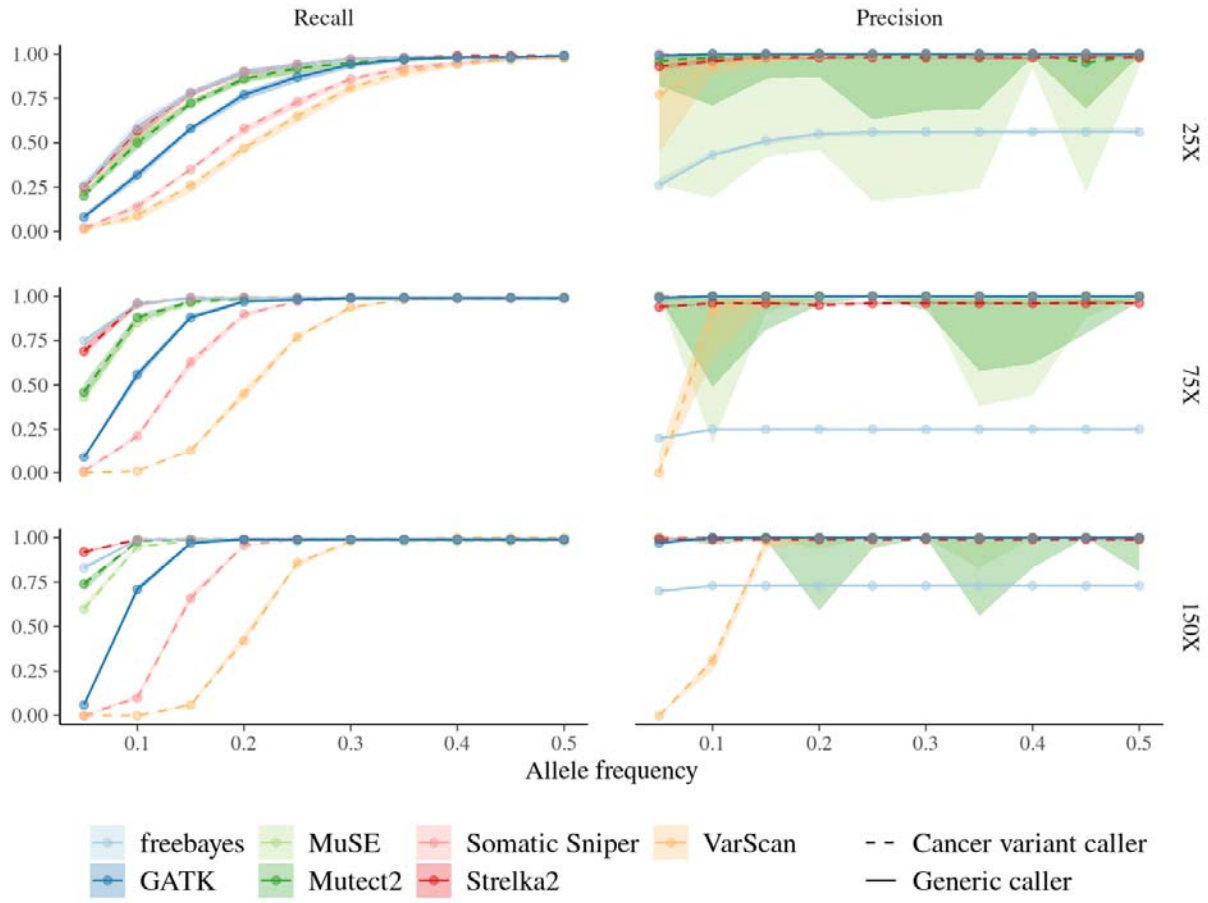
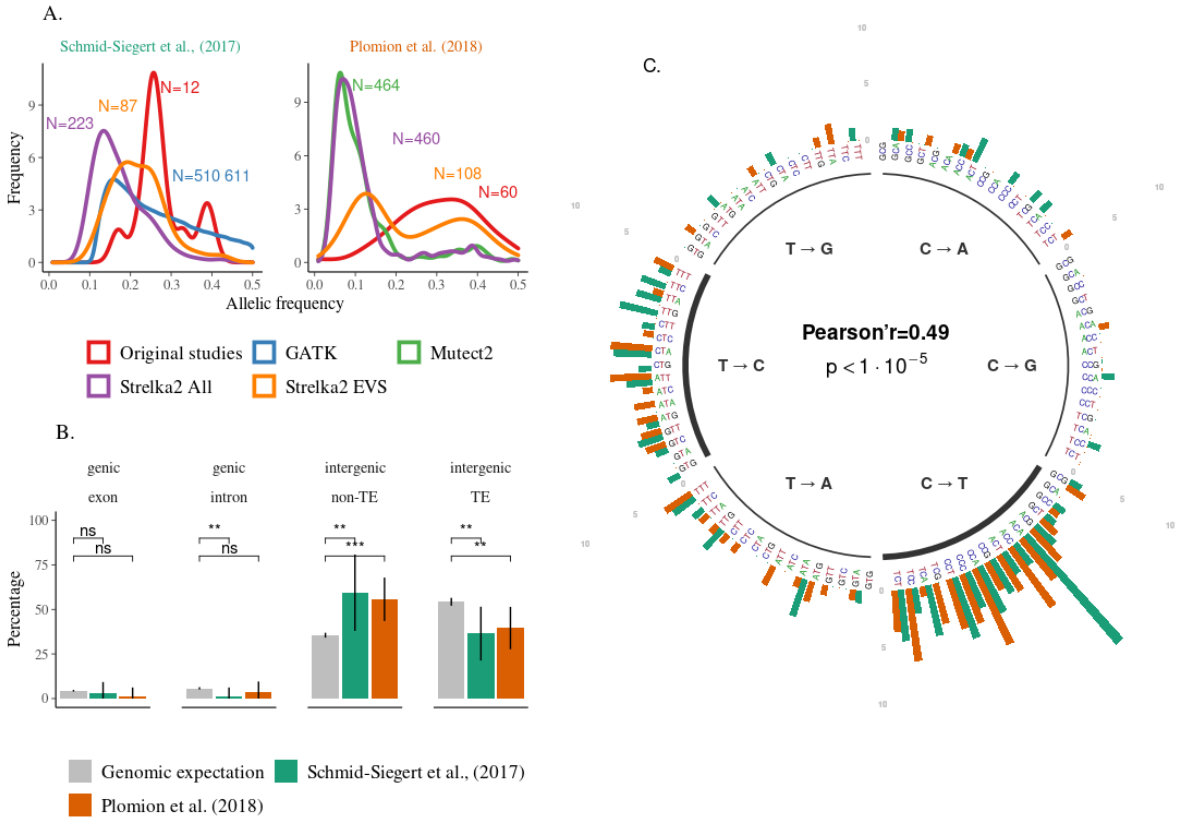


Figure 2: Variant caller performances to identify simulated mutations for varying allelic frequencies and sequencing depths (see Fig. S5 for all parameter combinations). The recall is the ability to detect (recover) the simulated mutations. The precision is the proportion of simulated mutations among 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).

272 We further investigated the performance of the best performing variant caller, *Strelka2*, on
273 two empirical datasets on pedunculate oak (Schmid-Siegert *et al.,* 2017; Plomion *et al.,*
274 2018) in comparison to the variant callers used in the original publications, *i.e.*, *GATK* and
275 *Mutect2*, respectively (see Supplementary Note S2). Mapping the raw data of Schmid-
276 Siegert *et al.,* (2017) and Plomion *et al.,* (2018) on the oak genome that we used as a
277 mapping reference for our empirical study, we successfully mapped 14 and 60 of the
278 mutations detected in the original articles, respectively. Across variant callers, we
279 recovered 12 (86%) and 60 (100%) of these original mutations in our total list of candidate
280 somatic mutations (Fig. 3A), strongly supporting the results shown by the two previous
281 studies. However, our analyses were able to detect far more robust candidate mutations
282 than initially reported. Applying filtering based on sequencing depth and mutated allele
283 copies (see Supplementary Note S2), *Strelka2* produced a smaller set of candidate
284 mutations than *GATK* but similar to *Mutect2*, with an estimated number of mutation
285 candidates 10- to 25-fold higher than that of the original studies (Fig. 3A). Adding *Strelka2*
286 recommended filtering based on empirical variant scores yielded the most conservative
287 dataset with a similar number of mutations between both studies and a 2 to 7-fold increase
288 compared to the original number of mutations. Due to lack of access to biological material
289 from the original studies, conclusions were drawn from this list of conservative candidate
290 somatic mutations (but see supplementary note S4 for a discussion regarding validation of
291 mutations). The distribution of allelic frequencies of detected mutations partly explains
292 differences among detection methods (Fig. 3A), with *Strelka2* and *Mutect2* detecting
293 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*

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

327 Discussion

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

339 transmissions. Therefore, plant mutation studies should make greater use of cancer variant
340 callers such as *Strelka2* rather than generic variant callers such as *GATK* to detect somatic
341 mutations, in agreement with previous studies on germline mutations detection (Chen *et al.*,
342 2019), especially for detecting low frequency mutations and when using low sequencing
343 depth. The importance of allele frequency-dependency in variant detection is not restricted
344 to somatic mutations, but also concerns for instance polyploid species, which includes many
345 agriculturally important autopolyploid plant species (e.g. potato, sugarcane). Our
346 simulation framework therefore provides general insights regarding the impact of allelic
347 dosage in mutation detection which go beyond somatic mutation detection.

348 One problem that may arise when analysing sample pairs with cancer variant callers
349 is the rapid increase in pairwise comparisons when using a larger sample size than
350 previous studies (e.g., N=3 in Plomion *et al.* 2018). A simple solution is the use of a single
351 reference sample such as a cambium sample from the base of the tree, which is therefore
352 considered as the closest genome to the seed, to compare it to all samples from branches
353 (Hanlon *et al.*, 2019). By reanalyzing the raw oak data (Schmid-Siegert *et al.*, 2017; Plomion
354 *et al.*, 2018), we found that the marked differences in the performance of variant callers
355 could account for the discrepancies in genome-wide plant somatic mutation rate estimates.
356 Our reanalysis shows robust evidence for an up to 7-fold higher number of mutations than
357 previously reported, a value closer to the expectations based on the theory (Schoen &
358 Schultz, 2019; Burian, 2021). We argue that knowledge and methodological transfers from
359 cancer to plant mutation detection are expected to contribute strongly to the upward trend
360 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).

375 ## Authors' contributions

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

380 ## References

381 Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck,
382 T.A., Simpson, J.T., Tonon, L., Sertier, A.S., Patch, A.M., Jäger, N., Ginsbach, P., Drews, R.,
383 Paramasivam, N., Kabbe, R., Chotewutmontri, S., Diessl, N., Previti, C., Schmidt, S., Brors, B.,
384 Feuerbach, L., Heinold, M., Gröbner, S., Korshunov, A., Tarpey, P.S., Butler, A.P., Hinton, J.,
385 Jones, D., Menzies, A., Raine, K., Shepherd, R., Stebbings, L., Teague, J.W., Ribeca, P., Giner,
386 F.C., Beltran, S., Raineri, E., Dabad, M., Heath, S.C., Gut, M., Denroche, R.E., Harding, N.J.,
387 Yamaguchi, T.N., Fujimoto, A., Nakagawa, H., Quesada, V., Valdés-Mas, R., Nakken, S., Vodák,
388 D., Bower, L., Lynch, A.G., Anderson, C.L., Waddell, N., Pearson, J.V., Grimmond, S.M., Peto, M.,
389 Spellman, P., He, M., Kandoth, C., Lee, S., Zhang, J., Létourneau, L., Ma, S., Seth, S., Torrents, D.,
390 Xi, L., Wheeler, D.A., López-Otín, C., Campo, E., Campbell, P.J., Boutros, P.C., Puente, X.S.,
391 Gerhard, D.S., Pfister, S.M., McPherson, J.D., Hudson, T.J., Schlesner, M., Lichter, P., Eils, R.,
392 Jones, D.T.W. & Gut, I.G. (2015). A comprehensive assessment of somatic mutation detection
393 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.,

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

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

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

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

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

Schöngart, J., Bräuning, A., Barbosa, A.C.M.C., Lisi, C.S. & Oliveira, J.M. de. (2017). *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.

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

501 # Supplementary material

502 The following Supporting Information is available for this article:

503 **Note S1.** Simulation results

504 **Note S2.** Reanalyses results

505 **Note S3.** Genomic expectations for the annotation of the mutations

506 **Note S4.** 'validation' of mutations

507 **Fig. S1.** Study scheme

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

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

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

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

516 **Fig. S6.** Observed allelic frequencies of candidate mutations depending on variant callers
517 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