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 explore and demonstrate the improved performance of variant callers developed for cancer research compared to single nucleotide polymorphism (SNP) callers in detecting *de novo* somatic mutations.
2. In an *in silico* experiment, we generated Illumina-like sequence reads spiked with simulated mutations at different allelic fractions 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. Based on the *in silico* experiment, variant callers developed for cancer research outperform SNP callers regarding plant mutation recall and precision, especially at low allele frequency. Such variants at low allelic fractions are typically expected for within-individual *de novo* plant mutations, which initially appear in single cells. Reanalysis of published oak data with *Strelka2*, the best-performing caller based on our simulations, identified up to 3.4x more candidate somatic mutations than reported in the original studies.
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. Notwithstanding its apparent simplicity, the understanding of mutation processes is one of the most common conceptual difficulties for biologists (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 currently available in the literature, mutation rates are, however, known to be highly variable across 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). In mammals, the somatic mutation rate varies directly with life span (Cagan *et al.,* 2022). 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 relative contribution of DNA replication and DNA repair errors to mutation rates represents another timely evolutionary question in the field (Gao *et al.,* 2019).

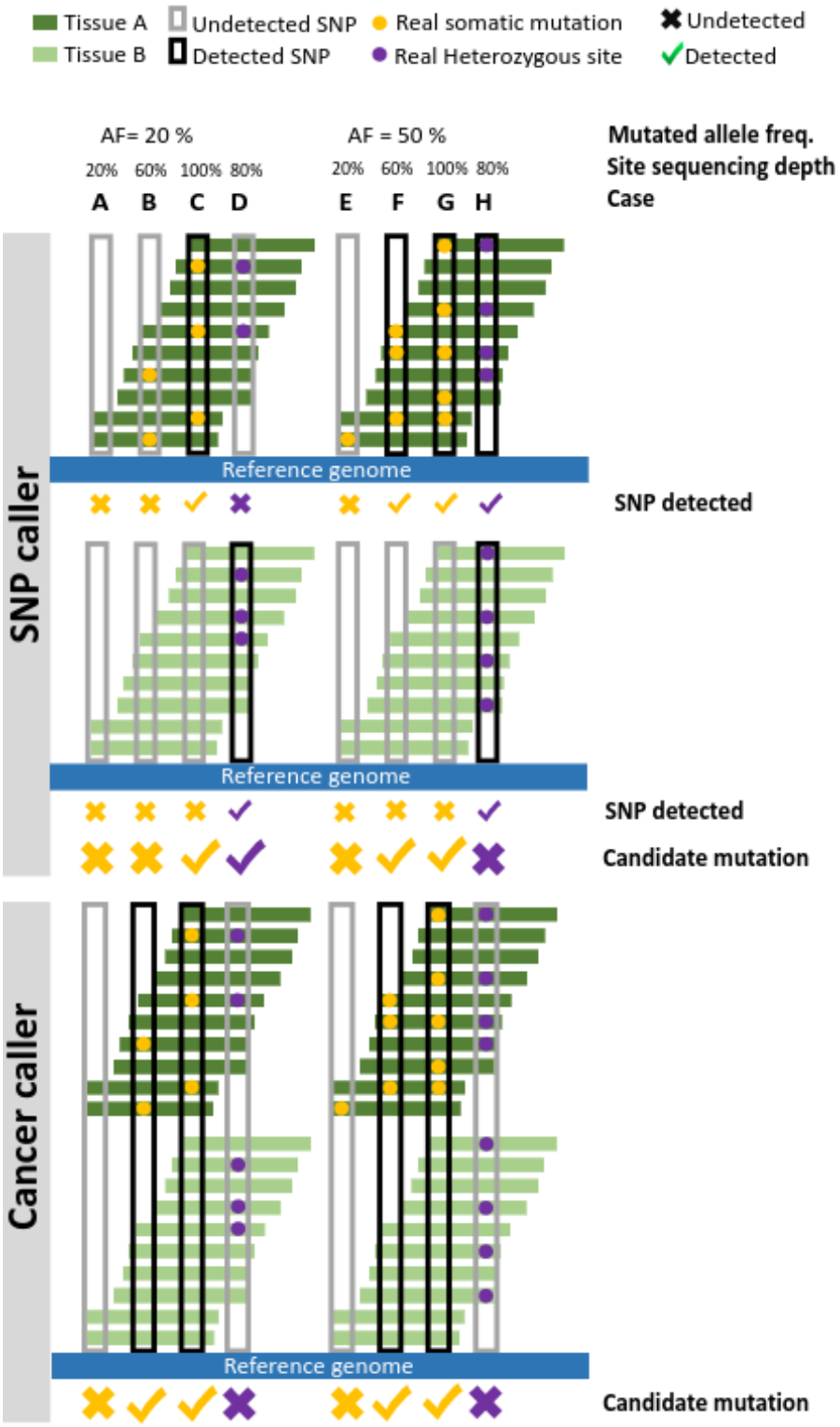
Unlike most animals that transmit to the next generation only mutations present in their germ cells (*i.e.* sperm and eggs), plants are expected to produce heritable somatic mutations as they grow throughout their lives, departing from the so-called Weismann's germ plasm

56 theory (Weismann, 1893; but see also Lanfear, 2018). As a consequence, long-lived species,
57 such as trees, are generally assumed to accumulate more heritable mutations than short-
58 lived species per generation (Hanlon *et al.,* 2019). To generate new knowledge on plant
59 mutation processes (Schoen & Schultz, 2019), several studies examined within-individual
60 variation in long-lived trees, whose individuals can live for more than a thousand years
61 (Schöngart *et al.,* 2017). Two studies used the pedunculate oak (*Quercus robur*), a long-lived
62 diploid and highly heterozygous European tree species, as a plant model to identify somatic
63 mutations. Schmid-Siegert *et al.,* (2017) identified 17 mutations by comparing sequencing
64 data from two branches of a 234-year-old individual. The authors therefore argued that
65 their results are consistent with a low mutation rate in pedunculate oak. Plomion *et al.,*
66 (2018) identified 46 mutations using three branches of a younger century-old individual,
67 which is an almost 10-fold higher rate after taking the tree age difference into account.
68 Plomion *et al.,* (2018) also recovered these new mutations on acorn embryos collected on
69 the same branches as those used for the *de novo* mutation identification, therefore
70 producing empirical support for departure from Weismann's germ plasm theory in oaks. A
71 shared limitation of both studies is that the authors used a single variant caller for *de novo*
72 mutation detection, without investigating beforehand the robustness of the results from the
73 selected variant caller. The absence of a simulation work to identify the best suited
74 detection method prior to the empirical investigations therefore represents a major limit
75 with regards to the accuracy and completeness of the previously reported *de novo*
76 mutations.

77 Variant callers are designed each for a specific purpose, with choices made by the developer
78 on sequence read filtering, and models and thresholds of sensitivity for the output of
79 variants. Broadly, we can distinguish 1. Single nucleotide polymorphism (SNP) callers
80 designed to detect heterozygous sites, i.e. sites with an expected allelic fractions of 0.5 in
81 the analysed sample, and 2. Cancer callers, designed to detect sites that are mutated in a
82 fraction of mutated cells in a sample, i.e. with expected allelic fraction ≤0.5.-When used to
83 detect mutations, SNP callers primarily detect candidate mutations per sample against the
84 reference genome and validate mutation robustness by comparing results between sample
85 pairs (Fig. 1). The per-sample strategy used in SNP callers carries the risk of overlooking
86 low-frequency mutated reads in one or more samples, which risks invalidating the mutation
87 in the other sample. To better address the specificities of detecting low frequency
88 mutations, the development of tools to detect mutations in humans is rapidly expanding in
89 cancer research (Kim et al., 2018; Alioto et al., 2015), where cancer callers identify
90 mutations by comparing two samples, one mutated and one normal sample, against the
91 reference genome (Fig. 1). Detecting mutations in cancers is conceptually similar to
92 detecting somatic mutations in plants, i.e., the aim is to detect mutations that potentially
93 affect only a small fraction of the sequenced tissue. However, cancer research frequently
94 uses very high sequencing depths (100X - 1000X), while the depth available for plants is
95 often considerably lower (e.g., 34X for Hanlon et al., 2019; 40X for Wang et al., 2018; or 70X
96 for Schmid-Siegert et al., 2017), bar a few exceptions (240X for Orr et al., 2019; 250X for
97 Plomion et al., 2018; or 1000X for Watson et al., 2016). Despite the advantage of cancer
98 callers to identify mutations, the specific challenge of detecting low frequency mutations is
99 so far poorly addressed in plants, where SNP callers have been the most frequently used

100 method to detect somatic mutations (Schmid-Siegert et al., 2017; Watson et al., 2016;
101 Hanlon et al., 2019; Orr et al., 2019). To demonstrate the advantages of using cancer callers
102 in somatic mutation detection for basic and applied plant research, we evaluated the
103 performance of the cancer callers compared to SNP callers using biological characteristics
104 and sequencing depth typical of plant studies.

105 Here, we performed both an *in silico* and an empirical data-based evaluation of the
106 performance of variant callers to detect somatic mutations, using simulated reads with
107 known mutations and two large published datasets on the same species (pedunculate oak,
108 *Quercus robur*) that applied different strategies for sequencing depth and mutation
109 detection (Schmid-Siegert *et al.,* 2017; Plomion *et al.,* 2018; see Fig. S1). We particularly
110 explored the cancer callers in a plant research context to answer the following questions:
111 (1) Can cancer research methods, both in terms of protocols (*i.e.* sequencing depth) and
112 tools (*i.e.* callers), improve the detection of plant somatic mutations?; and (2) Can
113 reanalyses of within-individual sequencing data provide new insights regarding plant
114 mutation processes?


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

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

131

132 # Material and methods

133 ## Study design

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

144 ## Generation of mutations

145 To ensure the feasibility of the project and to limit the computational load, a first step is to
146 subsample one or several sequences of user-defined length in the reference genome. The
147 first workflow named *generateMutations* therefore uses a bespoke R script named
148 *sample_genome* to generate these subsets. The workflow then takes advantage of the two
149 scripts included in *simuG* (v1.0.1, Yue & Liti, 2019), *vcf2model.pl,* and *simuG.pl*, respectively,
150 1) to build a model of heterozygous sites distribution for an haploid reference genome

151 based on a user-defined set of known heterozygous sites in *vcf* format and 2) to build the
152 second reference haploid genome comprising a user-defined number of heterozygous sites
153 to accurately represent diploidy. Typically, the user can define a number of heterozygous
154 sites based on the product of nucleotide diversity (π) and genome length (L). The workflow
155 then uses a homemade R script named *generate_mutations* to spike randomly the reference
156 genome with a user-defined number of *de novo* mutations which are drawn in a binomial
157 distribution using a user-defined transition/transversion ratio (**R**). Finally, the workflow
158 takes advantage of *InSilicoSeq* (v1.5.3, Gourlé *et al.,* 2019) defined with the model option
159 *hiseq* to generate datasets of mutated and non-mutated *in silico* Illumina-like sequencing
160 reads with a user-defined sequencing depth (**C**), sequentially using (1) the reference
161 haploid genome; (2) considering reference and alternate allele versions at heterozygous
162 sites, as the workflow was developed for a diploid species; and (3) considering, or ignoring,
163 the simulated de novo mutations which all feature the same user-defined allelic fraction
164 (**AF**).

165 **Detection of mutations**

166 The second workflow named *detectMutations* aims to detect somatic mutations from
167 mapped sequencing reads on a genome reference. Paired-end sequencing reads of every
168 library are quality checked using *FastQC* (v0.11.9) before trimming using *Trimmomatic*
169 (v0.39, Bolger *et al.,* 2014) keeping only paired-end reads without adaptors and a phred
170 score above 15 in a sliding window of 4 bases. Reads are aligned against the reference per
171 chromosome using BWA *mem* with the option to mark shorter splits (v0.7.17, Li & Durbin,

172 2009). Alignments are then compressed using *Samtools view* in CRAM format, sorted by
173 coordinates using *Samtools sort,* and indexed using *Samtools index* (v1.10, Li *et al.,* 2009).
174 Duplicated reads in alignments are marked using *GATK MarkDuplicates* (v4.2.6.1, Auwera *et*
175 *al.,* 2013). Finally, the workflow uses seven variant callers to detect mutations, including
176 variant callers developed originally for SNP calling and variant callers developed initially
177 for cancer research. SNP callers to detect variants included *GATK HaplotypeCaller* with
178 *GATK GenotypeGVCFs* (Auwera *et al.,* 2013) and *freebayes* (v1.3.2, Garrison & Marth, 2012)
179 using and reporting genotype qualities, without priors on allele balance, with a minimum
180 alternate allelic fraction of 0.03, a minimum repeated entropy of 1 and a minimum alternate
181 allele count of 2. Cancer callers developed for mutation detection included *VarScan* (v2.4.3,
182 Koboldt *et al.,* 2009), *Strelka2* (v2.9.10, Kim *et al.,* 2018), *MuSE* (v0.1.1, Fan *et al.,* 2016),
183 *Mutect2* (using a panel of normal and without soft clipped bases; within v4.2.6.1; Benjamin
184 *et al.,* 2019), and *Somatic Sniper* (v1.0.5.0, filtering reads with mapping quality less than 25,
185 filtering mutations with quality less than 15 with prior probability of a mutation of 0.0001;
186 Larson *et al.,* 2012). Then we only focused on the simulated mutations, and therefore
187 excluded from the analyses the known heterozygous sites provided by the user thanks to
188 the vcf file for *GATK*, *freebayes*, *Somatic Sniper*, and *Strelka2* using *BEDTools subtract*
189 (v2.29.2, Quinlan & Hall, 2010) or directly within the variant caller for *Mutect2* and *VarScan*.

190 *In silico* experiment

191 We used the *generateMutations* workflow to generate 1,000 mutations in the oak genome
192 with varying biological and sequencing parameters. To ensure consistency between the *in*

193 *silico* experiment and the reanalysis of empirical data, we used the reference genome
194 "Qrob_PM1N'' of Quercus robur 3P from Bordeaux, ENA accession number PRJEB8388
195 (Plomion *et al.,* 2018), thus assessing the behaviour of variant callers in the same genomic
196 context as used for the empirical work. To reduce the computational load, we only
197 generated mutations on the first megabase of the first chromosome of the oak assembly
198 ("Qrob_Chr01") in order to later focus the detection on this region. To check that the
199 conclusions regarding the callers are independent of the considered genomic region, we
200 ran five independent investigations based on randomly selected genome areas of a
201 megabase in length. Our results were highly congruent over all our investigations
202 (Pearson's correlations across all callers and all simulations, recall: 0.999, precision: 0.947,
203 but see Table S1 for differences among callers). We used known heterozygous sites from
204 the reference genome (Plomion *et al.,* 2018) to simulate back ten thousand heterozygous
205 sites (N = π x L = 10^4, assuming π = 0.01 (Plomion *et al.,* 2018) and L = 1 Mb). We used
206 varying values of transition/transversion ratio (**R** = [2, 2.5, 3]), allelic fraction (**AF** = [0.05,
207 0.1, 0.25, 0.5]), and sequencing depth (**C** = [25, 50, 100, 150, 200]), resulting in 60 simulated
208 datasets of mutated and associated base reads (3**R** x 4**AF** x 5**C**). We then used the
209 *detectMutations* workflow to detect (recall) spiked mutations with every variant caller
210 (*Mutect2, freebayes, GATK, Strelka2, VarScan, Somatic Sniper,* and *MuSe*). Using known
211 spiked mutations, we assessed the number of true positive (TP), false positive (FP), and
212 false negative (FN) for each variant caller to detect mutations and each combination of
213 biological and sequencing parameters. We used the resulting confusion matrix to calculate
214 the recall ($\frac{TP}{TP+FN}$) and the precision rates ($\frac{TP}{TP+FP}$). The recall rate represents the ability of
215 the variant caller to detect all mutations, while the precision rate represents the ability of

216 the variant caller to not confound other sites with mutations. We finally assessed each
217 variant caller to detect mutations using the recall and the precision rates with varying
218 transition/transversion ratio (**R**), allelic fraction (**AF**), and sequencing depth (**C**) to identify
219 the best performing variant caller based on biological and sequencing parameters.

220 Oak data reanalyses

221 We re-analyzed publicly available data on pedunculate oak from two projects conducted in
222 Lausanne, Switzerland (Schmid-Siegert *et al.,* 2017) and Bordeaux, France (Plomion *et al.,*
223 2018) (SRA PRJNA327502 and ENA PRJEB8388, respectively). Sequenced reads of every
224 library were quality checked, trimmed, and mapped with the same strategy than previously
225 described for the simulation work, i.e. using *FastQC*, *Trimmomatic,* and *BWA mem*. We then
226 used the best-performing variant caller at low coverage and allelic fraction based on our *in*
227 *silico* investigation, *Strelka2*, and the variant caller for mutation detection from the original
228 publication to compare the results, *i.e.*, *GATK* with *Best Practices* for the data from Schmid-
229 Siegert *et al.* (2017) and *Mutect2* for the data from Plomion *et al.* (2018). The former
230 comprised 2 libraries of medium sequencing depth (60X) representing one lower and one
231 upper branch. The latter comprised 3 libraries of high sequencing depth (160X)
232 representing 3 branches (lower, mid, and upper). For both data sets, we compared each
233 pair of sample points sequentially as the reference library and the potentially mutated
234 library to distinguish mutations among branches from heterozygous sites and sequencing
235 errors. For the data from Plomion *et al.* 2018, we further filtered out candidate somatic
236 mutations by using a cross-validation procedure to keep a coherent temporal pattern

237 among mutations following the original publication (Plomion *et al.,* 2018). Contrary to a
238 general expectation and a common view in the field (Schmid-Siegert *et al.*, 2017, Orr *et al.*,
239 2019), detected mutations do not always accumulate following the developing plant
240 architecture (Zahradníková *et al.*, 2020; Ren *et al.*, 2021). As a consequence, our cross-
241 validation represents a conservative strategy for the mutation detection, but it should be
242 noted that this strategy could have removed some true somatic mutations. We used these
243 raw datasets to identify the mutations from the original studies after realigning the
244 megabase containing the mutation on the 3P genome using *BLAT* (Kent, 2002). For both
245 datasets, we finally kept candidate mutations with (1) a read depth for both the normal and
246 mutated samples between half and two times the mean sequencing depth (30-120X and 80-
247 320X for Schmid-Siegert *et al.*, (2017) and Plomion *et al.*, (2018) datasets, respectively), (2)
248 an absence of the mutated allele in the normal sample, and (3) a minimum of 10 copies of
249 the mutated allele in the mutated sample. In addition, *Strelka2* calculates an empirical
250 variant score (EVS) based on a supervised random forest classifier trained on data from
251 sequencing runs under various conditions, which provides an overall quality score for each
252 variant (Kim *et al.*, 2018). We took advantage of the EVS to define a conservative set of
253 candidate mutations for both datasets, hereafter referred to as the EVS datasets. Given that
254 the proportion of the genome falling within the sequencing depth boundaries used for the
255 detection (i.e. between 50 and 200% of the mean sequencing depth) varies depending on
256 the dataset, we weighted the observed number of mutations by the proportion of the
257 genome satisfying the sequencing depth criteria to provide a more accurate and
258 comparable estimate of the real total number of mutations. Across both empirical studies,
259 the proportion of the genome with 50-200% sequencing depth varies between 71 and 87%.

260 **Results**

261 To advocate the use of cancer callers in somatic mutation detection for plant research, we
262 simulated sequencing data containing new mutations at a given allelic fraction (i.e. fraction
263 of simulated reads per genomic position carrying the mutated allele), and using varying
264 depths of sequencing (for variable transition/transversion ratios, see Fig S2). We then
265 evaluated the performance of variant callers as a function of allelic fraction and sequencing
266 depth. We found marked differences in: (1) the recall, the ability to recover the simulated
267 mutations; and (2) the precision, the proportion of true simulated mutations among all
268 variants detected. For allelic fractions equal to, or lower than, 0.25, cancer variant callers
269 (*Strelka2, Mutect2, MuSE,* but not *Somatic Sniper*) outperform SNP callers such as *GATK*,
270 *freebayes*, and *VarScan* (Fig. 2 and S3), mainly based on the recall. For allelic fractions over
271 0.25, all variant callers perform similarly well, except for *freebayes*, which identified many
272 false positives. Over the 80 tested parameter combinations, *Strelka2* was the best
273 performing variant caller for various allelic frequencies and sequencing depths (in 57/80
274 simulated datasets , with an average recall of 0.95 for a precision of 0.98, Fig. S2-4 and
275 Table S2 and S3) and an excellent computational efficiency (second fastest caller, Fig. S9).

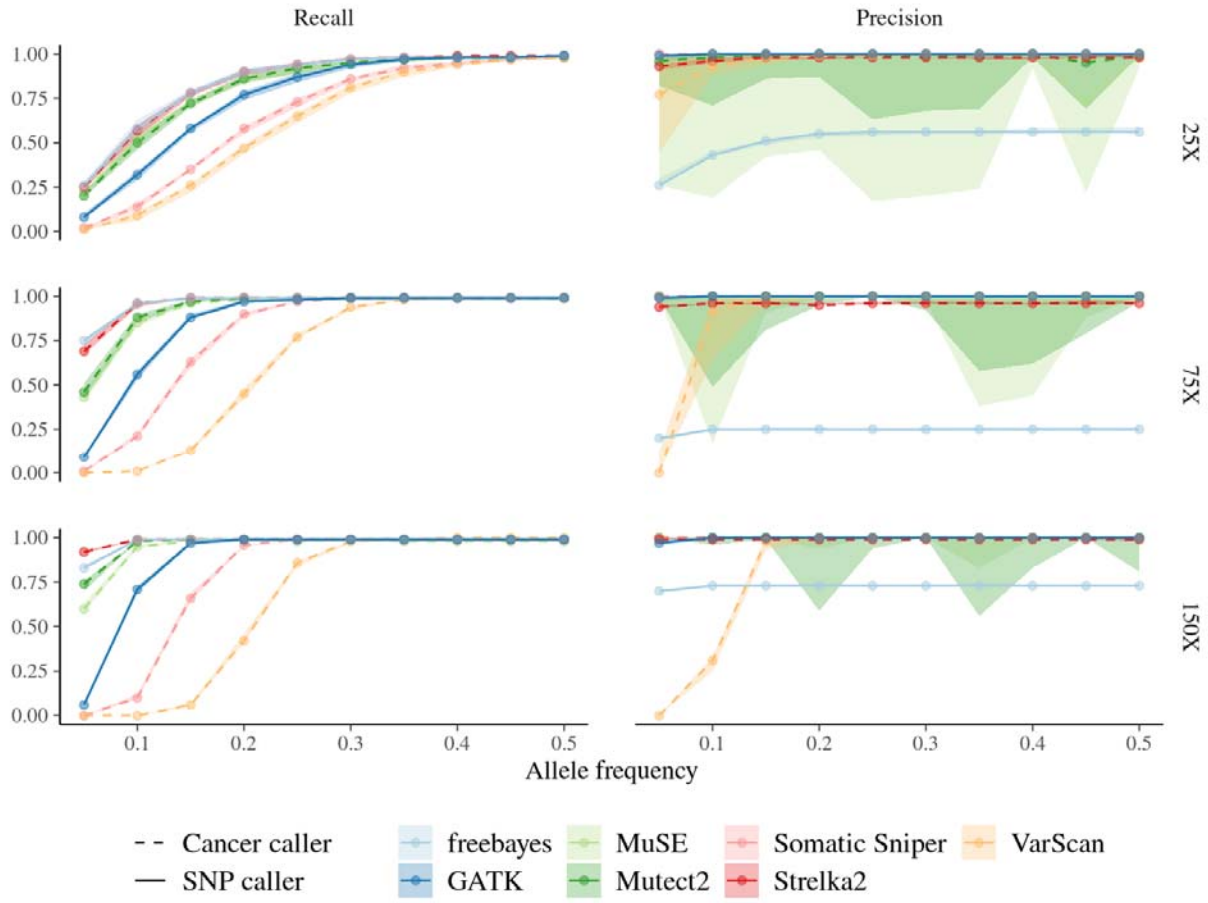


Figure 2: Variant caller performances to identify simulated mutations for varying allelic fractions and sequencing depths (see Fig. S5 for all parameter combinations). The recall is the ability to detect (recover) the simulated mutations. The precision is the proportion of simulated mutations among all variants detected (i.e. including false positives). Each point represents the averaged mutation recall or precision (10 simulations) for increasing allelic fraction 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 distinguishes SNP callers (dashed) and cancer callers (solid).

286 We further investigated the performance of the best performing variant caller based on our
287 *in silico* experiment, *Strelka2*, on two empirical datasets on pedunculate oak (Schmid-
288 Siegert *et al.,* 2017; Plomion *et al.,* 2018) in comparison to the variant callers used in the
289 original publications, *i.e.*, *GATK* and *Mutect2*, respectively. Mapping the raw data of Schmid-
290 Siegert *et al.,* (2017) and Plomion *et al.,* (2018) on the oak genome that we used as a
291 mapping reference for our empirical study, we successfully mapped 14 and 60 of the
292 mutations detected in the original articles, respectively. Across variant callers, we
293 recovered 12 (86%) and 60 (100%) of these original mutations in our total list of candidate
294 somatic mutations (Fig. 3A), strongly supporting the results shown by the two previous
295 studies. However, our analyses were able to detect far more candidate mutations than
296 initially reported. The filtering based on sequencing depth reduced the proportion of the
297 genome covered with adequate sequencing depth to 72% (assuming 30-120X) for Schmid-
298 Siegert *et al.,* (2017) and to 84% (assuming 80-320X) for Plomion *et al.,* (2018). Similarly,
299 the cross-validation between branches following the plant development (see Materials and
300 Methods for details) filtered out 27% of mutations detected for the dataset of Plomion *et al.,*
301 (2018). Using filtering based on sequencing depth and mutated allele copies with cross-
302 validation, *Strelka2* produced a smaller set of candidate mutations than *GATK* but similar to
303 *Mutect2*, with an estimated number of mutation candidates 2- to 4.3-fold higher than that
304 of the original studies (Fig. 3A). Adding *Strelka2* recommended filtering based on empirical
305 variant scores (EVS) yielded the most conservative dataset with a 1.5 to 3.4-fold increase
306 compared to the original number of mutations. Due to lack of access to biological material
307 from the original studies, conclusions were drawn from this list of conservative candidate
308 somatic mutations (but see discussion regarding validation of mutations). The distribution

309 of allelic fractions of detected mutations partly explains differences among detection
310 methods (Fig. 3A), with *Strelka2* and *Mutect2* detecting mutations with lower allelic
311 fractions than the candidate mutations presented in the original publications, especially for
312 the Plomion *et al.,* (2018) study that used higher sequencing depths allowing the
313 identification of these numerous low frequency mutations.

314 Based on the set of conservative mutations detected by *Strelka2* (EVS), we then explored
315 annotations and mutation spectra in both datasets (Fig. 3B-C), which have rarely been
316 explored in model plant species (but see first evidence based on mutation accumulation
317 lines in *Arabidopsis thaliana* in Weng *et al.,* 2019) and never in the wild. The proportions of
318 mutations found in different genomic regions (e.g. genic, intergenic) were highly correlated
319 between both original studies and proportional to the representation of the genomic
320 regions, supporting a random distribution of mutations throughout the genome (Fig. 3B).
321 Mutation spectra of the two studies are significantly correlated (Pearson'r=0.49, p<10-5),
322 with an enrichment in C>T transitions, particularly in some specific genomic contexts (Fig.
323 3C).

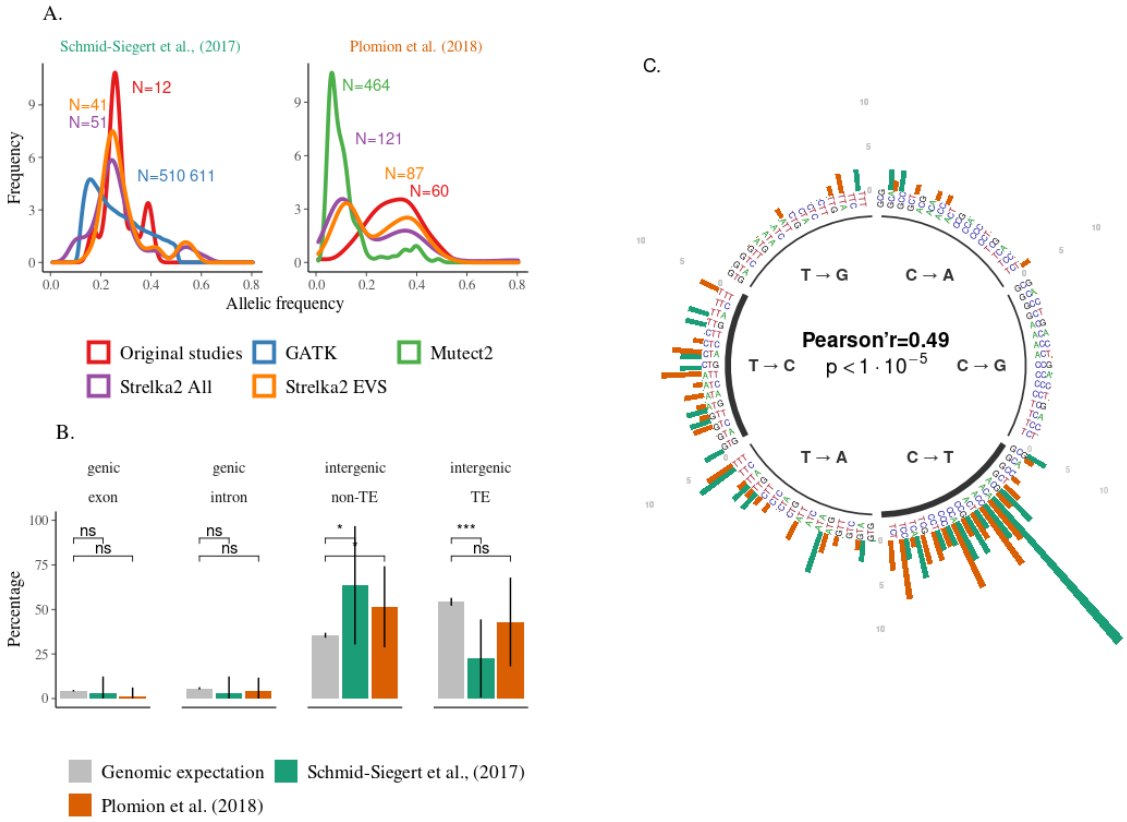


Figure 3: Candidate mutation spectra depending on variant callers and filtering in Schmid-Siegert et al., (2017) and Plomion et al., (2018). ***A.*** *Allelic fraction distribution for every dataset, including the candidate mutations from the original article present in the raw data from the reanalysis (red), the results of GATK with Best Practices (blue), Mutect2 after filtering (green), and Strelka2 after filtering (purple), and the results of Strelka2 using the filtering based on empirical variant scores named EVS (orange). The labels indicate the number of candidate mutations in each dataset. Per caller comparisons are available in Fig. S7.* ***B.*** *Annotation of the mutations detected by Strelka2 across chromosomes using the filtering based on empirical variant scores named EVS for Schmid-Siegert et al., (2017, green) and Plomion et al., (2018, orange) compared to the genomic expectation (grey, see Supplementary Note S1). Error bars represent the standard deviation (SD) of the observed*

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

346 **Discussion**

347 Mutation research in plants still primarily uses SNP callers and methodologies that are not
348 developed for the specificity and complexity of within-individual *de novo* mutation
349 detection. We demonstrated that plant mutation research could benefit from the
350 development of tools and protocols initially designed for human cancer research, which is a
351 rapidly expanding field (Kim *et al.*, 2018). We quantified expected marked differences in the
352 performance of variant callers for mutation detection based on sequencing depth and allelic
353 fraction. We demonstrated that cancer callers performed better than SNP callers for
354 mutation detection at low or intermediate allelic fraction or with low sequencing depth, and
355 similarly well for high allelic fraction. The overall higher detection efficiency demonstrates
356 the benefit of using tools and protocols initially designed for human cancer research. Low

357 allelic fraction mutations, potentially due to the chimeric nature of plant shoot apical
358 meristems structures (Burian, 2021), might be very important due to their great abundance
359 that may balance out their low chance of transmissions. Therefore, plant mutation studies
360 should make greater use of cancer variant callers such as *Strelka2* rather than SNP callers
361 such as *GATK* to detect somatic mutations, in agreement with previous studies on germline
362 mutations detection (Chen *et al.,* 2019), especially for detecting low frequency mutations
363 and when using low sequencing depth. The importance of allele fraction-dependency in
364 variant detection is not restricted to somatic mutations, but also concerns for instance
365 polyploid species, which includes many agriculturally important autopolyploid plant
366 species (e.g. potato, sugarcane). Our simulation framework therefore provides general
367 insights regarding the impact of allele fraction in mutation detection which go beyond
368 somatic mutation detection.

369 Validation of mutations

370 Intuitively, after identifying candidate mutations, one expects their "validation". In our
371 study, it was impossible to validate mutations using genotyping or resequencing because
372 we performed a reanalysis of already available sequence data, without having access to the
373 biological material. Yet, we argue that validation is more complex than generally thought.
374 We specifically discuss (i) allele fraction, (ii) fraction-aware approach, and (iii) offspring
375 validation. High frequency mutations (e.g. AF=0.5) will be easily confirmed by genotyping
376 or Sanger sequencing. But if one considers that low frequency mutations are only present in
377 a fraction of the tissue, SNP genotyping technologies are hindered by the complexity of

378 defining the genotype clustering while Sanger sequencing is hindered by the difficulty of
379 isolating the allelic assay from the background when reading the chromatogram. Therefore,
380 such validation strategies are conservative and favour high frequency mutations. To
381 unambiguously validate low-frequency mutations using intra-individual samples, an allelic
382 fraction-aware approach can be used, such as a hybridisation capture-based sequencing
383 strategy targeting regions of candidate mutations with high-coverage sequencing. However,
384 this would mean that the validation is based on another round of sequencing, bringing back
385 all the biases of the bioinformatics steps. Another strategy could be genotyping of inherited
386 mutations in offspring, working at the family rather than the individual level, which also has
387 its limitations (see Note S2). From our point of view, the current literature on somatic
388 mutations in plants does not mention the limits of validation, focusing on "fully validated"
389 mutations, thus probably favouring high frequency mutations. However, high-frequency
390 mutations are expected to be relatively rare compared to low-frequency mutations, which
391 affect conclusions about the total number of somatic mutations. In the relatively near
392 future, the use of single cell sequencing approaches to reveal intra-individual genetic
393 variation is likely to eliminate the problem at its source, avoiding the use of mixed samples
394 of mutated and non-mutated cells, leading to simple detection and validation.

395 Limits and future directions

396 Our study provided a first simple case of evaluating the performance of callers to detect
397 somatic mutations in plants. Our study was limited to variant callers in the cases of mutated
398 and normal sample pairs using the recommended default parameters. Future studies

399 should consider and quantify the performance of other types of callers, such as subclonal or
400 mosaic variant callers (e.g. *deepSNV* and *MosaicForecast*). Future studies could also consider
401 exploring tuning parameters for the specific task of detecting plant somatic mutations. We
402 hope that the open-source automated, highly reproducible, and scalable workflows
403 constructed for our study, *generateMutations* and *detectMutations*, will allow further
404 exploration of these questions. Our study also overlooked the importance of genome size, as
405 plants can have larger genomes than humans. Trees have a very wide range of genome
406 sizes, from small genomes (0.35 GB) to very large genomes (>20 GB), and our study focused
407 on the relatively small oak genome (0.75 GB). A large genome will obviously be a barrier to
408 high sequencing depth with a fixed budget, but with regard to mutation detection, our study
409 argues for the use of cancer callers in large genomes as they also outperform SNP callers in
410 terms of computational time (Fig. S9). Similarly, our study did not take into account the role
411 of polyploidy as we only focused on a diploid species. We assume that for ploidy greater
412 than two, the performance of cancer callers relative to SNP callers will increase since the
413 cancer caller will be even more efficient at low frequencies. But future studies could also
414 take advantage of our tools to explore the role of polyploidy in caller performance. Another
415 problem that may arise when analysing sample pairs with cancer callers is the rapid
416 increase in pairwise comparisons when using a larger sample size than previous studies
417 (e.g., N=3 in Plomion *et al.* 2018). A simple solution is the use of a single reference sample
418 such as a cambium sample from the base of the tree, which is therefore considered as the
419 closest genome to the seed, to compare it to all samples from branches (Hanlon *et al.,* 2019).

420 ## Conclusion

421 By reanalyzing the raw oak data (Schmid-Siegert *et al.,* 2017; Plomion *et al.,* 2018), we
422 found that the marked differences in the performance of variant callers could account for
423 the discrepancies in genome-wide plant somatic mutation rate estimates. Our reanalysis
424 using the best-performing caller based on our simulations, *Strelka2*, suggests an up to 3.4-
425 fold higher number of mutations than previously reported, a value closer to the
426 expectations based on the theory (Schoen & Schultz, 2019; Burian, 2021). We argue that
427 knowledge and methodological transfers from cancer to plant mutation detection are
428 expected to contribute strongly to the upward trend of this field and to reconcile empirical
429 reports with theoretical expectations.

430 ## Acknowledgments

431 The manuscript benefited from the comments of Nicolas Bierne and five anonymous
432 reviewers.

433 ## Funding

434 This study was funded through an Investissement d'Avenir grant of the ANR: CEBA (ANR-
435 10-LABEX-0025).

436 ## Conflict of interest disclosure

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

439 ## Data, script and code availability

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

445 ## Authors' contributions

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

450 ## References

451 Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck,
452 T.A., Simpson, J.T., Tonon, L., Sertier, A.S., Patch, A.M., Jäger, N., Ginsbach, P., Drews, R.,

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

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

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

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

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

475 Cagan, A., Baez-Ortega, A., Brzozowska, N., Abascal, F., Coorens, T. H. H., Sanders, M. A., …
476 Martincorena, I. (2022). Somatic mutation rates scale with lifespan across mammals.
477 Nature, 604(7906), 517–524. https://doi.org/10.1038/s41586-022-04618-z

478 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,
479 D., Wu, L., Zhou, W.J., Chi, H., Sun, R.X., Dong, M.Q. & He, S.M. (2019). A high-speed search
480 engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked
481 peptides. *Nature Communications*, **10**. Retrieved from http://dx.doi.org/10.1038/s41467-
482 019-11337-z

483 Fan, Y., Xi, L., Hughes, D.S.T., Zhang, J., Zhang, J., Futreal, P.A., Wheeler, D.A. & Wang, W.
484 (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model
485 improves sensitivity and specificity in mutation calling from sequencing data. *Genome*
486 *biology*, **17**, 178. Retrieved from http://dx.doi.org/10.1186/s13059-016-1029-6

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

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

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

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

497 Harris, K. & Pritchard, J.K. (2017). Rapid evolution of the human mutation spectrum. *eLife*,
498 **6**, 1–17.

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

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

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

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

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

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

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

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

516 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. &
517 Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**,
518 2078–2079.

519 Lynch, M., Ackerman, M.S., Gout, J.F., Long, H., Sung, W., Thomas, W.K. & Foster, P.L. (2016).
520 Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, **17**,
521 704–714. Retrieved from http://dx.doi.org/10.1038/nrg.2016.104

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

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

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

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

Ren, Y., He, Z., Liu, P., Traw, B., Sun, S., Tian, D., … Wang, L. (2021). Somatic Mutation Analysis in Salix suchowensis Reveals Early-Segregated Cell Lineages. *Molecular Biology and Evolution*, **38**, 5292–5308. https://doi.org/10.1093/molbev/msab286

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*.

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

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

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

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

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

571 Yue, J.X. & Liti, G. (2019). SimuG: A general-purpose genome simulator. *Bioinformatics*, **35**,
572 4442–4444.

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

576

577 # Supplementary material

578 The following Supporting Information is available for this article:

579 **Note S1.** Genomic expectations for the annotation of the mutations

580 **Note S2.** 'Validation' of mutations in offsprings

581 **Fig. S1.** Study scheme

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

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

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

588 **Fig. S5.** Mutation recall and precision rates for SNP and cancer variant callers by allelic
589 fraction and sequencing depth

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

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

594 **Fig. S8.** Context-dependent mutation spectrum depending on variant callers and filtering in

595 Schmid-Siegert *et al.,* (2017) and Plomion *et al.,* (2018)

596 **Tab. S1.** Mean and standard deviation in performance of variant callers for mutation

597 detection across all simulations

598 **Tab S2.** Mean and standard deviation in performance of variant callers for mutation

599 detection with varying allelic fraction and sequencing depth

600