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

Sylvain Schmitt[1,*], Thibault Leroy[2,3], Myriam Heuertz[4], Niklas Tysklind[5]

[1]CNRS, UMR EcoFoG (Agroparistech, Cirad, INRAE, Université des Antilles, Université de la Guyane), Campus Agronomique, 97310 Kourou, French Guiana

[2]Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, A-1030 Vienna, Austria

[3]IRHS-UMR1345, Université d'Angers, INRAE, Institut Agro, SFR 4207 QuaSaV, 49071 Beaucouzé, France

[4]Université Bordeaux, INRAE, BIOGECO, 69 route d'Arcachon, CS 80227, 33612 Cestas Cedex, France

[5]INRAE, UMR EcoFoG (Agroparistech, CNRS, Cirad, Université des Antilles, Université de la Guyane), Campus Agronomique, 97310 Kourou, French Guiana

[*]Correspondence: Sylvain Schmitt, UMR EcoFoG, Campus Agronomique, 97310 Kourou, French Guiana. Email: sylvain.m.schmitt@gmail.com. ORCID: 0000-0001-7759-7106

Abstract

Mutation, the source of genetic diversity, is the raw material of evolution; however, it remains an understudied process in plants. Using simulations, we demonstrate that generic variant callers, commonly used to detect mutations in plants, are outperformed by methods developed for cancer research. Reanalysis of published data identified up to 7x more somatic mutations than initially reported, advocating the use of cancer research callers to boost mutation research in plants.

24 ## Main text

25 DNA sequence mutation is the raw material for evolutionary change, but, despite its crucial
26 role, many fundamental questions around the mutation process are still open. Mutation
27 rates are highly variable along the tree of life, differing by several orders of magnitude
28 among species and kingdoms, and are considered as an evolvable trait *per se*[1]. Mutations
29 are assumed to be random, but the rate at which different nucleotides mutate strongly
30 depends on the surrounding nucleotides[2], hereafter referred to as a mutation spectrum.
31 The mutation spectra themselves are now believed to evolve over time[3], even at relatively
32 short evolutionary timescales[4]. The drivers of new mutations, previously thought to be
33 simply due to DNA replication errors, are now also debated[5].

34 Unlike most animals that transmit to the next generation only mutations present in their
35 germ cells (sperm and eggs), plants are expected to produce heritable somatic mutations as
36 they grow throughout their lives, departing from the so-called Weismann's germ plasm
37 theory[6,7]. As a consequence, long-lived species, such as trees, are generally assumed to
38 accumulate more heritable mutations than herbs per generation[8]. To generate new
39 knowledge on plant mutation processes[9], several studies examined within-individual
40 variation in long-lived trees, whose individuals can live for more than a thousand years[10].
41 Two studies used the pedunculate oak (*Quercus robur*), a long-lived European tree species,
42 as a plant model to identify somatic mutations. Schmid-Siegert et al.[11] identified 17
43 mutations by comparing sequencing data from two branches of a 234-year-old individual.
44 The authors therefore argued that their results are consistent with a low mutation rate in

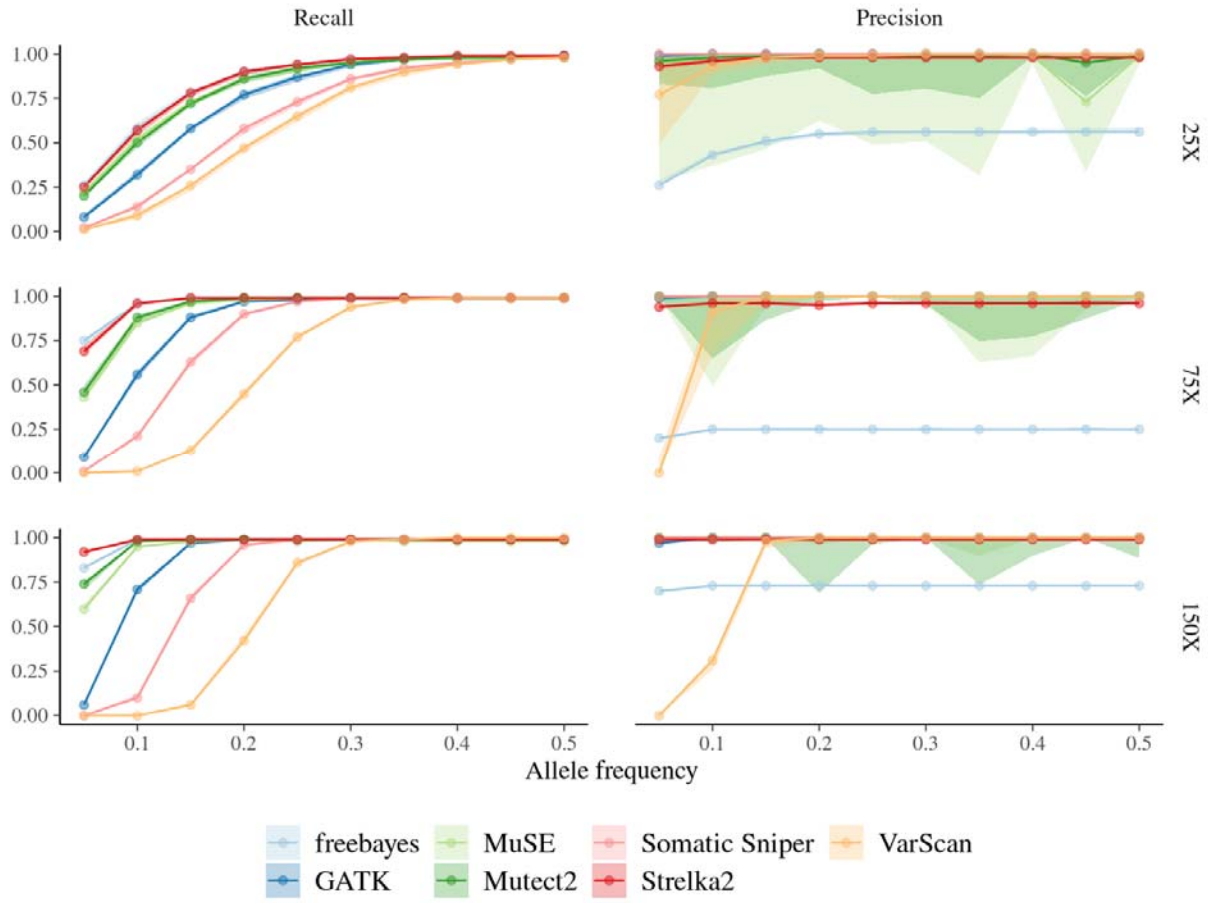
45 pedunculate oak. Plomion et al.[12] identified 46 mutations using three branches of a younger
46 (century-old) individual, which is an almost 10-fold higher rate after taking the tree age
47 difference into account. A shared limitation of both studies is that the authors did not
48 identify the detection method best suited for the features of somatic mutations.

49 The development of tools to detect mutations in humans is rapidly expanding in cancer
50 research[13,14]. Detecting mutations in cancers is conceptually similar to detecting somatic
51 mutations in plants, *i.e.*, the aim is to detect mutations that potentially affect only a small
52 fraction of the sequenced tissue. This specific challenge is poorly addressed in plants, where
53 mutation detection remains based on generic variant callers, which were initially designed
54 to detect heterozygous sites[11-12,15-17]. Transferring mutation detection tools from cancers to
55 plants requires evaluating their performance in a plant research context. Cancer research
56 frequently uses very high sequencing depths (100X - 1000X), while the depth available for
57 plants is often considerably lower (e.g., 34X[16], 40X[18] or 70X[11]), bar a few exceptions
58 (240X[17], 250X[12] or 1000X[15]). To improve the detection of mutations for basic and applied
59 plant research, a deep evaluation of the performance of variant callers is needed in relation
60 to the biological features and quality of data typical of plant studies.

61 We performed both an *in silico* and an empirical data-based evaluation of the performance
62 of variant callers to detect somatic mutations, using two large published datasets on the
63 same species (pedunculate oak, *Quercus robur*) that applied different strategies for
64 sequencing depth and mutation detection[11,12] (see Fig. S1). We particularly explored the
65 role of the sequencing depth and allelic frequency of the somatic mutation in plant tissues

66 to answer the following questions: (1) Can cancer-derived methods, both in terms of
67 protocols (i.e. sequencing depth) and tools (i.e. callers), improve the detection of somatic
68 mutations in plants?; and (2) Can reanalyses of within-individual sequencing data provide
69 new insights regarding plant mutation processes?

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



84

85 **Fig. 1 | Variant caller performances to identify simulated mutations for varying allelic**
86 **frequencies and sequencing depths (see Fig. S5 for all parameter combinations).** The
87 recall is the ability to detect (recover) the simulated mutations. The precision is the
88 proportion of simulated mutations among all variants detected (i.e. including false
89 positives). Each point represents the averaged mutation recall or precision (10 simulations)
90 for increasing allelic frequency and sequencing depth. The shaded area represents the 90th
91 quantiles of recall and precision rates over the 10 replicates computed for all callers, but
92 only visible for the precision of *Muse*, *Mutect2*, and *VarScan*.

93 We further investigated the performance of the best performing variant caller, *Strelka2*, on
94 two empirical datasets on pedunculate oak[11,12] in comparison to the variant callers used in
95 the original publications, *i.e,. GATK* and *Mutect2*, respectively (see Supplementary Note S2).
96 Mapping the raw data of Schmid-Siegert et al.[11] and Plomion et al.[12] on the oak genome that
97 we used as a mapping reference for our empirical study, we successfully mapped 14 and 60
98 of the mutations detected in the original articles, respectively. Across variant callers, we
99 recovered 12 (86%) and 60 (100%) of these original mutations in our total list of candidate
100 somatic mutations (Fig. 2A), strongly supporting the results shown by the two previous
101 studies. However, our analyses were able to detect far more candidate mutations than
102 initially reported. Applying filtering based on sequencing depth and mutated allele copies
103 (see Supplementary Note S2), *Strelka2* produced a more reliable set of candidate mutations
104 than *GATK* but similar to *Mutect2*, with an estimated number of mutation candidates 10- to
105 25-fold higher than that of the original studies (Fig. 2A). Adding *Strelka2* automatic filtering
106 yielded the most robust and conservative dataset with a similar number of mutations
107 between both studies and a 2- to 7-fold increase compared to the original number of
108 mutations. The distribution of allelic frequencies of detected mutations partly explains
109 differences among detection methods (Fig. 2A), with *Strelka2* and *Mutect2* detecting
110 mutations with lower allelic frequencies than the candidate mutations presented in the
111 original publications, especially for the Plomion et al.[12] study that used higher sequencing
112 depths.

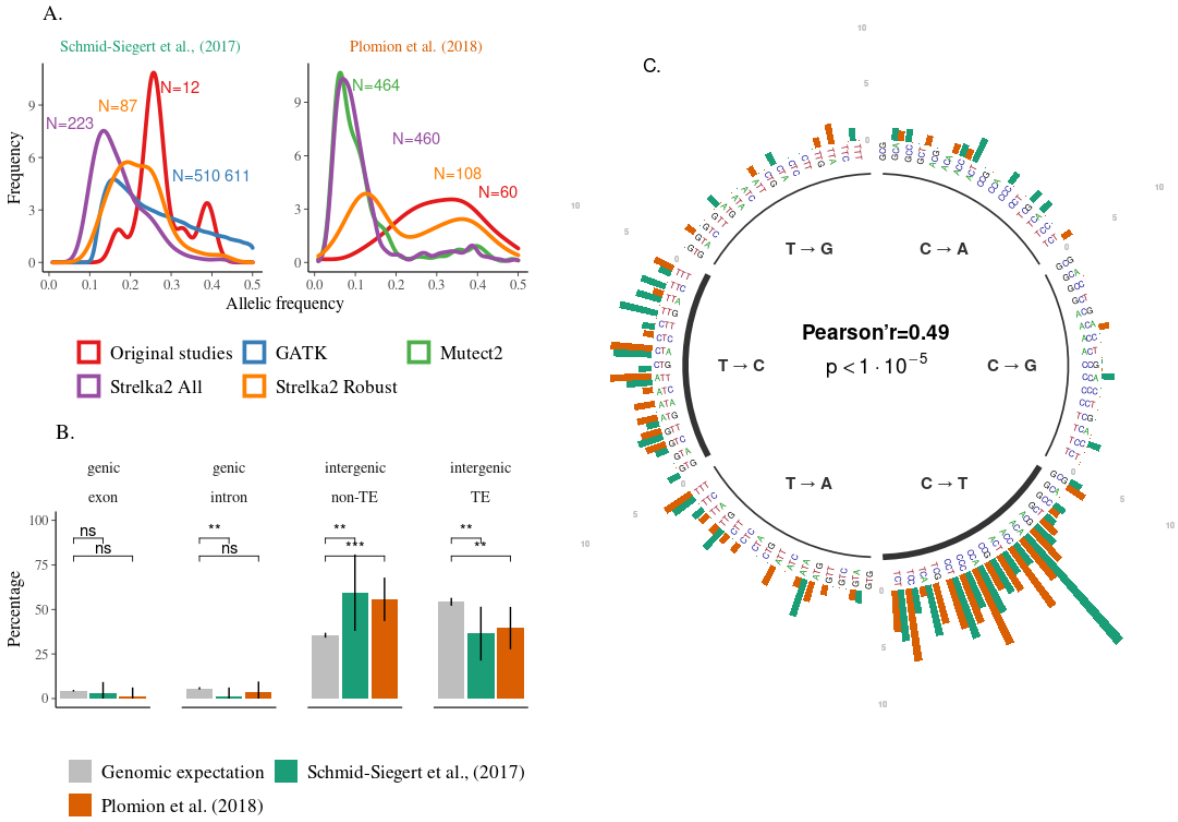


Fig. 2 | Candidate mutation spectra depending on variant callers and filtering in Schmid-Siegert et al.[11] and Plomion et al.[12]. A. Allelic frequency distribution for every dataset, including the candidate mutations from the original article present in the raw data from the reanalysis (red), the results of *GATK* with Best Practices (blue), *Mutect2* (green), and *Strelka2* after filtering (purple), and the results of *Strelka2* using the automatic filtering named robust (orange). The labels indicate the number of candidate mutations in each dataset. Per caller comparisons are available in Fig. S6. **B.** Annotation of the mutations detected by *Strelka2* across chromosomes using the automatic filtering named robust for Schmid-Siegert *et al.*[11] (green) and Plomion *et al.*[12] (orange) compared to the genomic expectation (grey, see Supplementary Note S3). Error bars represent the standard deviation of the observed percentages across chromosomes, and the annotation above the columns

125 indicates the significance of the Student's t-test comparing the mean percentage of
126 mutations to the mean genomic expectation, with *ns*, **, and *** corresponding to non-
127 significant, *p<0.01,* and *p<0.001* differences, respectively. **C.** Context-dependent mutation
128 spectra depending on mutation types for the results of *Strelka2* using the automatic filtering
129 named robust. Mutation types have been summarised into six main classes with thicker
130 lines for transversion compared to transition, and then differentiated depending on their 5'
131 and 3' genomic contexts, see Fig. S7-8. Pearson's correlation r measures the correlation of
132 the mutation spectra between Schmid-Siegert *et al.*[11] and Plomion *et al.*[12].

133 Based on the set of robust mutations detected by *Strelka2*, we then explored annotations
134 and mutation spectra in both datasets (Fig. 2B-C), which have rarely been explored in
135 model plant species (but see first evidence based on mutation accumulation lines in
136 Arabidopsis[19]) and never in the wild. The proportions of mutations found in different
137 genomic regions (e.g. genic, intergenic) were highly correlated between both original
138 studies and proportional to the representation of the genomic regions, supporting a
139 random distribution of mutations throughout the genome (Fig. 2B). Mutation spectra of the
140 two studies are significantly correlated (Pearson'r=0.49, p<7.4 10-5), with an enrichment in
141 C>T transitions, particularly in some specific genomic contexts (Fig. 2C).

142 In conclusion, mutation research in plants still primarily uses generic variant callers and
143 methodologies that are not developed for the specificity and complexity of mutation
144 detection; however, plant research could benefit from the development of tools and
145 protocols initially designed for human cancer research, which is a rapidly expanding field[13].

146 Here, we found marked differences in the performance of variant callers for mutation
147 detection based on sequencing depth and allelic frequency. We found that cancer variant
148 callers performed better than generic variant callers for mutation detection at low or
149 intermediate allelic frequency or with low sequencing depth, and similarly well for high
150 allelic frequency. Therefore, plant mutation studies should make greater use of cancer
151 variant callers such as *Strelka2* rather than generic variant callers such as *GATK* to detect
152 mutations, especially for low frequency mutations and when using low sequencing depth.
153 By reanalyzing the raw oak data[11,12], we found that the marked differences in the
154 performance of variant callers account for the discrepancies in genome-wide plant somatic
155 mutation rate estimates. Our reanalysis shows robust evidence for an up to 7-fold higher
156 number of mutations than previously reported, a value closer to the expectations based on
157 the theory[9,20]. We argue that knowledge and methodological transfers from cancer to plant
158 mutation detection are expected to contribute strongly to the upward trend of this field
159 and to reconcile empirical reports with theoretical expectations.

160 Material and Methods

161 **Study design.** We developed two workflows: 1) to generate Illumina-like sequencing reads
162 including mutations with varying biological and sequencing parameters; and 2) to detect
163 mutations with multiple variant callers (Fig. S1). We used both *singularity* containers[21] and
164 the *snakemake* workflow engines[22] to build automated, highly reproducible (FAIR), and
165 scalable workflows. We then used both workflows to test the best performing variant caller
166 for mutation detection *in silico* based on biological and sequencing parameters. The

167 workflows are publicly available at https://github.com/sylvainschmitt. We finally used the
168 identified variant caller to detect mutations in pedunculate oak, *Quercus robur* L., by re-
169 analysing data from two somatic mutation projects on oaks led by INRA Bordeaux, France[12]
170 and the University of Lausanne, Switzerland[11].

171 **Generation of mutations.** The first workflow named *generateMutations* takes advantage of
172 the two scripts included in *simuG*[23], *vcf2model.pl,* and *simuG.pl*, respectively, to build a
173 model of heterozygous sites distribution in the haploid reference genome based on a user-
174 defined base of known heterozygous sites in *vcf* format and to build the second reference
175 haploid genome comprising a user-defined number of heterozygous sites to accurately
176 represent diploidy. Typically, the user can define a number of heterozygous sites based on
177 the product of nucleotide diversity (π) and genome length (L). The workflow uses a
178 homemade R script named *generate_mutations* to spike randomly the reference genome
179 with a user-defined number of mutations which are drawn in a binomial distribution using
180 a user-defined transition/transversion ratio (**R**). Finally, the workflow takes advantage of
181 *InSilicoSeq*[24] defined with the model option hiseq to generate datasets of mutated and non-
182 mutated *in silico* Illumina-like sequencing reads using (1) the original reference haploid
183 genome, (2) the reference haploid genome with heterozygous sites (pedunculate oak is a
184 diploid species), and (3) the reference genome spiked with mutations following user-
185 defined allelic fraction (**AF**) and depth of sequencing depth (**C**).

186 **Detection of mutations.** The second workflow named *detectMutations* aims to detect
187 somatic mutations from mapped sequencing reads on a genome reference. Pair-end

188 sequencing reads of every library are quality checked using *FastQC* before trimming using
189 *Trimmomatic*[25] keeping only pair-end reads without adaptors and a phred score above 15
190 in a sliding window of 4 bases. Reads are aligned against the reference per chromosome
191 using BWA *mem* with the option to mark shorter splits[26]. Alignments are then compressed
192 using *Samtools view* in CRAM format, sorted by coordinates using *Samtools sort,* and
193 indexed using *Samtools index*[27]. Duplicated reads in alignments are marked using *GATK*
194 *MarkDuplicates*[28]. Finally, the workflow uses seven variant callers to detect mutations,
195 including generic variant callers to detect variants and dedicated variant callers for
196 mutation detection. Generic variant callers to detect variants include the combination of
197 *GATK HaplotypeCaller* and *GATK GenotypeGVCFs*[28] and *freebayes*[29] using and reporting
198 genotype qualities, without priors on allele balance, with a minimum alternate fraction of
199 0.03, a minimum repeated entropy of 1 and a minimum alternate count of 2. Cancer variant
200 callers developed for mutation detection include *VarScan*[30], *Strelka2*[13], *MuSE*[31], *Mutect2*[32]
201 (using a panel of normal and without soft clipped bases), and *Somatic Sniper*[33] (filtering
202 reads with mapping quality less than 25, filtering mutations with quality less than 15 with
203 prior probability of a mutation of 0.0001). Then we only focused on the simulated
204 mutations, and therefore excluded from the analyses the known heterozygous sites
205 provided by the user thanks to the vcf file for *GATK*, *freebayes*, *Somatic Sniper*, and *Strelka2*
206 using *BEDTools substract*[34] or directly within the variant caller for *Mutect2* and *VarScan*.

207 ***In silico* experiment.** We used the *generateMutations* workflow to generate mutations in
208 the oak genome with varying biological and sequencing parameters. To ensure consistency
209 between the *in silico* experiment and the reanalyses, we used the reference genome

210 "Qrob_PM1N" of Quercus robur 3P from Bordeaux[12], thus assessing the behavior of variant
211 callers in the same genomic context. To reduce the computational load, we only generated
212 mutations on the first megabase of the first chromosome of the oak assembly
213 ("Qrob_Chr01") in order to later focus the detection on this region. We used known
214 heterozygous sites from the reference genome[12] to simulate back one thousand
215 heterozygous sites (π = 0.01 , L = 1 Mb , N = π x L = 10^4). We used varying values of
216 transition/transversion ratio (**R** = [2, 2.5, 3]), allelic fraction (**AF** = [0.05, 0.1, 0.25, 0.5]), and
217 sequencing depth (**C** = [25, 50, 100, 150, 200]), resulting in 60 simulated datasets of
218 mutated and associated base reads (3**R** x 4**AF** x 5**C**). We then used the *detectMutations*
219 workflow to detect back spiked mutations with every variant caller (*Mutect2, freebayes,*
220 *GATK, Strelka2, VarScan, Somatic Sniper,* and *MuSe*). Using known spiked mutations, we
221 assessed the number of true positive (TP), false positive (FP), and false negative (FN) for
222 each variant caller to detect mutations and each combination of biological and sequencing
223 parameters. We used the resulting confusion matrix to calculate the recall ($\frac{TP}{TP+FN}$) and the
224 precision rates ($\frac{TP}{TP+FP}$). The recall rate represents the ability of the variant caller to detect
225 all mutations, while the precision rate represents the ability of the variant caller to not
226 confound other sites with mutations. We finally assessed each variant caller to detect
227 mutations using the recall and the precision rates with varying transition/transversion
228 ratio (**R**), allelic fraction (**AF**), and sequencing depth (**C**) to identify the best performing
229 variant caller based on biological and sequencing parameters.

230 **Oak data reanalyses.** We re-analyzed publicly available oak data from two projects led by
231 Bordeaux, France[12] and Lausanne, Switzerland[11] (SRA PRJNA327502 and ENA PRJEB8388,

232 respectively). We then used the best-performing variant caller based on our *in silico*
233 investigation, *Strelka2*, and the variant caller for mutation detection from the original
234 publication to compare the results *i.e. GATK* with *Best Practices* for Swiss data[11] and
235 *Mutect2* for French data[12]. The Swiss data comprised 2 libraries of medium sequencing
236 depth (60X) representing one lower and one upper branch. The French data comprised 3
237 libraries of high sequencing depth (160X) representing 3 branches (lower, mid, and upper).
238 For both Swiss and French data, we compared each pair of sample points sequentially as
239 the normal library and the mutated library to decipher mutations among branches from
240 heterozygous sites and sequencing errors. For the French data, we further filtered out
241 candidate somatic mutations by using a cross-validation procedure to keep a coherent
242 temporal pattern among mutations following the original publication[12]. We used these raw
243 datasets to identify the mutations from the original studies after realigning them on the 3P
244 genome using *BLAT*[35]. For both datasets, we finally kept candidate mutations with (1) a
245 read depth for both the normal and mutated samples between half and two times the mean
246 sequencing depth (30-120X and 80-320X for Swiss and French datasets, respectively), (2)
247 an absence of the mutated allele in the normal sample, (3) a minimum of 10 copies of the
248 mutated allele in the mutated sample and (4) an allelic frequency <0.5. In addition, we took
249 advantage of the Empirical Somatic Score (ESS) calculated by *Strelka2*[13] to build a robust
250 dataset of candidate mutations for both datasets. Finally, we measured the proportion of
251 the genome following the sequencing depth requirement, *i.e.* half and two times the mean
252 sequencing depth, for both studies to estimate the total number of mutations from the
253 observed number of candidates.

Acknowledgements

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

Authors' contributions

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

Data Accessibility

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

References

Main text

1. Lynch, M., Ackerman, M. S., Gout, J. F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, *17*(11), 704–714. https://doi.org/10.1038/nrg.2016.104
2. Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, *349*(6255), 1483–1489. https://doi.org/10.1126/science.aab4082
3. 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*(May), 1–8. https://doi.org/10.1038/ncomms15183
4. Harris, K., & Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *eLife*, *6*, 1–17. https://doi.org/10.7554/eLife.24284
5. Gao, Z., Moorjani, P., Sasani, T. A., Pedersen, B. S., Quinlan, A. R., Jorde, L. B., … Przeworski, M. (2019). Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(19), 9491–9500. https://doi.org/10.1073/pnas.1901259116
6. Weismann, A. (1893). *The germ-plasm: a theory of heredity*. Scribner's.
7. Lanfear, R. (2018). Do plants have a segregated germline? *PLoS Biology*, *16*(5), 1–13. https://doi.org/10.1371/journal.pbio.2005439
8. Hanlon, V. C. T., Otto, S. P., & Aitken, S. N. (2019). Somatic mutations substantially increase the per-generation mutation rate in the conifer Picea sitchensis . *Evolution Letters*, *3*(4), 348–358. https://doi.org/10.1002/evl3.121

9. Schoen, D. J., & Schultz, S. T. (2019). Somatic Mutation and Evolution in Plants. *Annual Review of Ecology, Evolution, and Systematics*, *50*, 49–73. https://doi.org/10.1146/annurev-ecolsys-110218-024955
10. Schöngart, J., Bräuning, A., Barbosa, A. C. M. C., Lisi, C. S., & de Oliveira, J. M. (2017). *Dendroecology Tree-Ring Analyses Applied to Ecological Studies*. *Springer Nature*.
11. Schmid-Siegert, E., Sarkar, N., Iseli, C., Calderon, S., Gouhier-Darimont, C., Chrast, J., … Reymond, P. (2017). Low number of fixed somatic mutations in a long-lived oak tree. *Nature Plants*, *3*(12), 926–929. https://doi.org/10.1038/s41477-017-0066-9
12. Plomion, C., Aury, J. M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., … Salse, J. (2018). Oak genome reveals facets of long lifespan. *Nature Plants*, *4*(7), 440–452. https://doi.org/10.1038/s41477-018-0172-3
13. Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., … Saunders, C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, *15*(8), 591–594. https://doi.org/10.1038/s41592-018-0051-x
14. Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., … Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms10001
15. Watson, J. M., Platzer, A., Kazda, A., Akimcheva, S., Valuchova, S., Nizhynska, V., … Riha, K. (2016). Germline replications and somatic mutation accumulation are independent of vegetative life span in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(43), 12226–12231. https://doi.org/10.1073/pnas.1609686113

16. Hanlon, V. C. T., Otto, S. P., & Aitken, S. N. (2019). Somatic mutations substantially increase the per-generation mutation rate in the conifer Picea sitchensis . *Evolution Letters*, *3*(4), 348–358. https://doi.org/10.1002/evl3.121
17. Orr, A. J., Padovan, A., Kainer, D., Külheim, C., Bromham, L., Bustos-Segura, C., … Lanfear, R. (2019). A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. *bioRxiv*. https://doi.org/10.1101/727982
18. Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., … Tian, D. (2019). The architecture of intra-organism mutation rate variation in plants. *PLoS Biology*, *17*(4), 1–29. https://doi.org/10.1371/journal.pbio.3000191
19. Weng, M. L., Becker, C., Hildebrandt, J., Neumann, M., Rutter, M. T., Shaw, R. G., … Fenster, C. B. (2019). Fine-grained analysis of spontaneous mutation spectrum and frequency in arabidopsis thaliana. *Genetics*, *211*(2), 703–714. https://doi.org/10.1534/genetics.118.301721
20. Burian, A. (2021). Does Shoot Apical Meristem Function as the Germline in Safeguarding Against Excess of Mutations? *Frontiers in Plant Science*, *12*(August), 1–9. https://doi.org/10.3389/fpls.2021.707740
21. Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLoS ONE*, *12*(5), 1–20. https://doi.org/10.1371/journal.pone.0177459
22. Köster, J., & Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

23. Yue, J. X., & Liti, G. (2019). SimuG: A general-purpose genome simulator. Bioinformatics, 35(21), 4442–4444. https://doi.org/10.1093/bioinformatics/btz424
24. Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., & Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics, 35(3), 521–522. https://doi.org/10.1093/bioinformatics/bty630
25. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
26. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324
27. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
28. Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, *43*(1), 483–492. https://doi.org/10.1002/0471250953.bi1110s43
29. Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing, 1–9. Retrieved from http://arxiv.org/abs/1207.3907
30. Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., … Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing of individual

and pooled samples. *Bioinformatics*, *25*(17), 2283–2285. https://doi.org/10.1093/bioinformatics/btp373

31. Fan, Y., Xi, L., Hughes, D. S. T., Zhang, J., Zhang, J., Futreal, P. A., … Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*, *17*(1), 178. https://doi.org/10.1186/s13059-016-1029-6
32. Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., & Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. https://doi.org/10.1101/861054
33. Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., … Ding, L. (2012). Somaticsniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, *28*(3), 311–317. https://doi.org/10.1093/bioinformatics/btr665
34. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033
35. Kent, W. J. (2002). BLAT---The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202
36.