Title: Low number of fixed somatic mutations in a long-lived oak tree.

Authors: 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 CS, Keller L, Pannell JR, Reymond A, Robinson-Rechavi M, Xenarios I, Reymond P

Journal: Nature plants

Year: 2017 Dec

Issue: 3

Volume: 12

Pages: 926-929

DOI: 10.1038/s41477-017-0066-9

Low number of fixed somatic mutations in a long-lived oak tree

Emanuel Schmid-Siegert[1,†], Namrata Sarkar[2,3,4,†], Christian Iseli[1,†], Sandra Calderon[1], Caroline Gouhier-Darimont[5], Jacqueline Chrast[2], Pietro Cattaneo[5], Frédéric Schütz[2], Laurent Farinelli[6], Marco Pagni[1], Michel Schneider[7], Jérémie Voumard[8], Michel Jaboyedoff[8], Christian Fankhauser[2*], Christian S. Hardtke[5*], Laurent Keller[3*], John R. Pannell[3*], Alexandre Reymond[2*], Marc Robinson-Rechavi[3,4*], Ioannis Xenarios[1,2,7*] and Philippe Reymond[5*]

[1]Vital-IT Competence Center, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [2]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland, [3]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland [4]Evolutionary Bioinformatics Group, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [5]Department of Plant Molecular Biology, University of Lausanne, 1015 Lausanne, Switzerland, [6]Fasteris SA, 1228 Plan-les-Ouates, Switzerland, [7]Swiss-Prot group, Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland, [8]Risk Analysis Group, Institute of Earth Sciences, University of Lausanne, 1015 Lausanne, Switzerland, [†]These authors contributed equally.

[*]Correspondence: christian.fankhauser@unil.ch, christian.hardtke@unil.ch, laurent.keller@unil.ch, john.pannell@unil.ch, alexandre.reymond@unil.ch, marc.robinson-rechavi@unil.ch, ioannis.xenarios@sib.swiss, philippe.reymond@unil.ch

Because plants do not possess a defined germline, deleterious somatic mutations can be passed to gametes, and a large number of cell divisions separating zygote from gamete formation may lead to many mutations in long-lived plants. We sequenced the genome of two terminal branches of a 234-year-old oak tree and found several fixed somatic single-nucleotide variants (SNVs) whose sequential appearance in the tree could be traced along nested sectors of younger branches. Our data suggest that stem cells of shoot meristems in trees are robustly protected from the accumulation of mutations.

The accumulation of deleterious mutations is a fundamental aspect of plant ageing and evolution. Because the pedigree of cell division that generates somatic tissues is poorly understood, the number of cell divisions that separate zygote from gamete formation is difficult to estimate; this number is expected to be particularly large in trees and could plausibly lead to a large number of DNA replication errors[1-3]. Apical meristem, which contain stem cells, arises from the embryo. These cells divide and produce progenitor cells that undergo division, elongation and differentiation to form the main stem. Tree architecture is determined by axillary meristems, which form in leaf axils, and are responsible for the emergence of side branches. They are separated from the apical meristem by elongating internodes. In oak, early and repeated growth cessation of terminal meristems leads to a branching pattern originating from such axillary meristems. In turn, axillary meristems grow out and produce secondary axillary meristems. This process is reiterated indeterminately to produce highly ramified trees of large stature, resulting in thousands of terminal ramets[4]. The cumulative number of cell divisions separating meristems may lead to somatic mutations caused by replication errors and exposure to the environment. Although mechanisms like DNA repair, programmed cell death or arrest of cell division can prevent mutation load, some mutations may be fixed in stem cell populations, colonizing whole meristems and derived tissues. To detect such fixed mutations, we sequenced genomic DNA from the terminal branches of an iconic old oak tree (*Quercus robur*) on the University of Lausanne campus, known as the 'Napoleon Oak' by the academic community. The tree was 22 years old when, on May 12, 1800, Napoleon Bonaparte and his troops crossed what is now the Lausanne University campus, on their way to conquer Italy. At the time of sample harvest for our study, the dividing apical meristems of this magnificent tree (Figure 1, Supplementary Figure 1) had been exposed for 234 years to potential mutagens, such as UV light and radioactive radiation.

To identify fixed somatic variants (i.e., those present in an entire sector of the Napoleon Oak) and to reconstruct their origin and distribution among branches, we collected 26 leaf samples from different locations on the tree. We first sequenced the genome from leaves sampled on terminal ramets of one lower and one upper branch of the tree. We then used a combination of short-read Illumina and single-molecule real-time (SMRT, Pacific Biosciences) sequencing to generate a *de novo* assembly of the oak genome. After removing contigs < 1000bp, we established a draft sequence of ca. 720 megabases (Mb) at a coverage of ca. 70X, with 85,557 scaffolds and a N50 length of 17,014. Our sequence is thus in broad agreement with the published estimated genome size of 740 Mbp[5]. The oak genome is predicted to encode 49,444 predicted protein-coding loci (Supplementary Table 1).

We used two approaches to identify SNVs (single-nucleotide variants) between the sequenced genomes of the two terminal branches. First, we aligned Illumina paired-reads on the repeat-masked genome in combination with the GATK[6] variant caller. This allowed us to establish a list of 3,488 potential SNV candidates with high confidence scores. From this list, 1,536 SNVs were experimentally tested by PCR-seq, of which only seven could be confirmed (see Methods). Second, we used fetchGWI[7] to map read pairs to the non-masked genome. We were able to call 5,330 potential SNVs from the mapped reads using a simple read pileup process. Further analysis identified 82 putatively variable positions, including the seven already identified using the repeat-masked genome analysis described above (see Methods). Ten of the remaining 75 candidates from the second approach were confirmed by PCR-seq, increasing the total number of confirmed SNVs separating the two genomes to 17 (Figure 1, Supplementary Table 2); these were further confirmed by Sanger sequencing. Based on a conservative estimate, we are likely to have missed no more than 17 further such sites among candidate SNVs (see Supplementary Methods). Furthermore, analyses of the false-negative rate suggest that we have missed between 4 and 13 additional SNVs (see Supplementary Methods). We thus estimate a grand total of between 38 and 47 SNVs between the two analyzed genomes, giving a fixed mutation rate of between 4.2 and 5.2 x 10^{-8} substitutions/site/generation.

As expected, all 17 confirmed SNVs were heterozygous. Indeed, because the level of heterozygosity of the Napoleon Oak genome is 0.7%, the probability of finding a single SNV at sites that were initially homozygous in both samples is only 0.12. Intriguingly, two SNVs were found on the same contig, separated by only 12 bp (Supplementary Figure 2 and Supplementary Table 2). Sixteen SNVs occurred in introns or non-coding sequences that are probably neutral. The remaining SNV (SNV1), which occurred in a large sector of the tree, generates an arginine-to-glycine conversion in a putative E3-ubiquitin ligase (Supplementary Table 2). The functional impact of exchanging a positively charged arginine with a non-charged and smaller glycine residue is unknown and deserves further analysis.

Having confidently established 17 SNVs, we then assessed their occurrence throughout the tree. We used Sanger sequencing to genotype the remaining 24 terminal branches sampled from other parts of the tree and checked for the presence of each SNV. SNVs were found in different sectors of the tree in a nested hierarchy that clearly indicates the accumulation of mutations along branches during development (Figure 1, Supplementary Figure 3). These results both provide independent confirmation of the originally identified SNVs, and demonstrate their gradual, nested appearance and fixation in developmentally connected branches during growth. Thus, while the exact ontogeny of the Napoleon Oak may be difficult to reconstruct, our SNV analysis generated a nested set of lineages supported by derived mutations, analogous to a phylogenetic tree.

The fixed mutation rate in annual plants has been estimated to range from 5 x 10^{-9} to 30 x 10^{-9} substitutions/site/generation, based on mutations accumulated during divergence between monocots and dicots[8]. Values for mutation accumulation lines of *Arabidopsis thaliana* maintained in the laboratory range between 7.0 and 7.4 x 10^{-9}, which corresponds to ~1 mutation/genome/generation[9,10]. *Arabidopsis* is an annual plant that reaches approximately 30 cm in height before producing seeds. In contrast, the physical distance traced along branches between the terminal branches we sequenced for the Napoleon Oak is about 40 m (Figure 1). Thus, the lineages in oak were separated by a considerably larger physical distance than in Arabidopsis (40 m instead of 30 cm), implying a higher number of mitoses between them, although the exact number is difficult to estimate. If we hypothesize that the number of fixed mutations per generation is correlated with the number of mitotic divisions from zygote to gametes of the next generation[1,11], the much greater size of the oak tree should drastically impact the total numbers of SNVs accumulating along its branches. In addition, contrary to *Arabidopsis* whose life cycle is only 2-3 months, the apical mersitems of the Napoleon Oak were exposed to mutagenic UV light for 234 years; it is thus not altogether surprising that the majority of SNVs were likely due to UV-induced mutations (see Supplementary Discussion). If we take into account these two factors, we expect the per-generation mutation rate in oak to be approximately two orders of magnitude larger than in Arabidopsis, a value considerably higher than the observed < 10-fold difference (see above). The surprisingly low frequency of fixed mutations suggests that a mechanism is in place to prevent their accumulation in the tree.

Classical studies of shoot apical meristem organization have found that the most distal zone has a significantly lower rate of cell division than more basal regions of the apex, and might therefore be relatively protected from replication errors[12,13]. In a recent study that followed the fate of dividing cells in the apical meristems of *Arabidopsis* and tomato, Burian et al.[14] showed that an unexpectedly low number of divisions separate apical from axillary meristems. In these herbaceous plants, axillary meristems are separated from apical meristem stem cells by seven to nine cell divisions, with internode growth occurring through the division of cells behind the meristem. The number of cell divisions between early embryonic stem cells and terminal meristems thus depends more on the number of branching events than on absolute plant size. Burian et al.[14] postulated that if the same growth pattern described above for *Arabidopsis* and tomato applies to trees, the number of fixed somatic mutations might be much lower than is commonly thought, and they should be found in relatively small sectors as nested sets of mutations. Napoleon Oak's apical meristems are of similar diameters to those of tomato[14] (Supplementary Figure 4) and show similar ontogeny. It thus seems reasonable to suppose that the growth pattern described in *Arabidopsis* and tomato is quite general in flowering plants and might also apply to long-lived trees. The low number of SNVs and their nested appearance in sectors of the Napoelon Oak are thus consistent with hypotheses proposed in Burian et al.[14].

Mutations accumulate with age, irrespective of plant stature, and long-term exposure to UV radiation contributes to such changes. As noted above, the type of observed SNVs were mostly G:C→A:T transitions, indicative of UV-induced mutagenesis (see Supplementary Discussion). Oaks protect their meristems in buds under multi-layered leaf-like structures (Supplementary Figure 4), potentially reducing the incidence of UV mutagenesis. The relatively low number of fixed mutations identified in our study may thus be explained by the protective nature of oak bud morphology as well as by the pattern of cell division predicted by Burian et al.[14]. Our results also suggest that mutations due to replication errors in long-lived plants may be less important than environmentally induced mutations. In this context, it is noteworthy that there was no evidence for an expansion of DNA-repair genes in the oak genome compared to *Arabidopsis* (Supplementary Table 3).

To our knowledge, only two examples of functional mosaicism have been reported in trees, a low incidence that might be attributable to the low number of fixed mutations that we report here. Although most non-neutral mutations should be maladaptive, eucalyptus trees have been observed with a few branches that are biochemically distinct from the rest of the canopy and have become resistant to Christmas beetle defoliation[15,16]. Functionally relevant somatic mutations, such as SNV1 in our study, may thus occasionally contribute to adaptive evolution if transferred to the fruits, but will more typically increase the genetic load of a population, with implications for inbreeding depression and mating-system evolution (Supplementary Discussion).

Our data give an unprecedented view of the limited role played by fixed somatic mutations in a long-lived organism, and support the notion that stem cells in trees, although vulnerable to environment-induced and replication-induced mutations, are probably quite well protected from them. Consistent with this finding, a recent study in *Arabidopsis* has shown that the number of cell divisions from germination to gametogenesis is independent of life span and vegetative growth[17]. Additional studies on different tree species and older specimens are needed to test the generality of our study. This work also illustrates the potential for analyses of multiple genomes from single individuals, which throw exciting new light on the rate, distribution and potential impact of fixed somatic mutations in both plant and animal tissues[18,19].

Methods

Materials and genome sequencing. Leaves were collected in April 2012 from the terminal part of a lower (sample 0) and an upper branch (sample 66) of the Napoleon Oak (*Q. robur*) on the Lausanne University Campus (Switzerland, 46°31'18.9"N 6°34'44.5"E). The age of the tree was estimated by a tree ring analysis from a sample taken at the basis of the trunk (Laboratoire Romand de Dendrochronologie, 1510 Moudon, Switzerland). DNA from the two samples was extracted and the genome sequenced. Paired-end sequencing libraries with insert size of 400 bp were constructed for each DNA sample according to the manufacturer's instructions. Then, 100 bp paired-reads were generated on Illumina HiSEq 2000 at Fasteris (www.fasteris.com). In addition, 3 kb mate-pair libraries from sample 0 were constructed and sequenced with single-molecule real-time (SMRT) technology according to the manufacturer's instructions (Pacific Biosciences). Short reads were combined with PacBio reads to assemble a reference genome (Supplementary Methods).

SNV identification. We used two different methods to identify SNVs (see flowchart, Supplementary Figure 5). In the first one, Illumina reads (278,547,120 and 278,651,792 for sample 0 and 66, respectively) were aligned to the masked (RepeatMasker, v4.05) *de novo* assembly with Bowtie2 (v2.2.2, https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.2.2) using default parameters. GATK[6] v2.5.2 was used for local realignment and variant calling using standard hard filtering parameters according to GATK Best Practices recommendations[20]. Prior to variant calling, each sample was screened for duplicates using PICARD tools (http://broadinstitute.github.io/picard/ v2.9.0, MarkDuplicates). Variants with confidence score ≥50 were retained further. We identified 1,832,554 heterozygous sites common to both samples, as well as 314,865 putative differences between sample 0 and 66 (165,489 sites predicted to be homozygous on sample 0 and heterozygous on sample 66 and 149,376 homozygous on sample 66 and heterozygous on sample 0). The distribution of the confidence scores of the 1,832,554 heterozygous sites common to both samples was a superposition of a Gaussian distribution, peaking at 910, possibly representing true positives, and of an exponential distribution, possibly representing the decreasing number of false positives with regard to increasing confidence score. Importantly, the distribution of scores of the sites with putative differences between samples was an exponential distribution of very low values, similar to the potential false positives of shared heterozygote sites. We thus hypothesized that sites that are truly different between samples 0 and 66 were unlikely to be present at sites with a confidence score below 300. From 3,488 putative SNVs with a confidence score ≥300 on the heterozygous sites and ≥200 on homozygous sites, we selected 1,536 SNVs for validation by PCR-seq (Supplementary Methods). We identified only 7 true SNVs that were further confirmed by Sanger sequencing. This low rate is consistent with the expectation from the distribution of GATK scores for these sites.

In the second method, Illumina reads of samples 0 and 66 were mapped against the non-masked oak genome assembly. The genome was 719,779,348 bp long, but 69,130,634 (9.52%) of those nucleotides were gaps and were discarded, leaving an actual search space of 650,648,714 bp. Of the latter, 458,143,725 nucleotides with a read coverage ≥8 in both samples were analysed further (Supplementary Figure 6). The mapping process was performed at the read pair level by the genome mapping tool, fetchGWI[7], followed by a detailed sequence alignment tool, align0[21]. Potential SNVs were called from the mapped reads by a simple read pileup process followed by detection of positions where the pileup shows variations with respect to the reference genome; this produced a list of 5,330 positions. Those positions were browsed through a local adaptation of the samtools pileup browser[22] to evaluate the quality of the mapping in the surrounding region and to discriminate between well-assembled high-quality regions with two alleles per sample, or low complexity and possibly badly assembled repeated regions. Criteria for selection were ≥8 reads in each orientation (see above); 100% homozygosity site for one sample and at least 30% minor allele frequency for the other sample with variants in both orientations; and coherent sequence ±50 bp from variant site. This manual process led to the selection of 82 putative variable positions, including the seven already identified. Upon experimental validation, 10 of the remaining 75 candidates were confirmed by PCR-seq and Sanger sequencing. The Food and Drug Administration (FDA) has evaluated this approach in an effort to assess, compare, and improve techniques used in DNA testing on human genome variation analysis (https://precision.fda.gov/challenges/consistency). Within this frame, our method reached a F-score (F-score evaluates precision and recall) over 95% comparable to other identifiers like BWA coupled with GATK.

SNV Genotyping. Leaf DNA from different locations on the tree was prepared and amplified using primers located 100-150 bp away from the 17 confirmed SNVs (Supplementary Table 4). Amplicons were then subjected to Sanger sequencing.

3D Modeling of the Oak. We used LiDAR (Light Detection and Ranging) technology to scan the oak with a 3D laser scanner (Leica). Terrestrial LiDAR scans were taken around the oak every 60°. The 6 scans were cleaned from background objects and aligned in order to generate a 1.2 million 3D-points cloud (Polywork, www.innovmetric.com). Mesh from the 3D-points cloud was colorized to produce the final 3D oak model.

Data availability. All Illumina reads and SMRT sequences have been deposited in GenBank under accession BioProject PRJNA327502.

References

1. Scofield, D.G. & Schultz, S.T. *Proc. Royal Soc. London B: Biol. Sci.* **273,** 275–282 (2006).
2. Ally., D., Ritland, K. & Otto, S.P. *PLoS Biol.* **8,** e1000454 (2010).
3. Bobiwash, K., Schultz, S.T. & Schoen, D.J. *Heredity* **111,** 338–344 (2013).
4. Millet, J. *L'architecture des arbres des régions tempérées: son histoire, ses concepts, ses usages* (MultiMondes, 2012).
5. Plomion, C. *et al. Mol. Ecol. Resour.* **16,** 254-265 (2016).
6. McKenna, A. *et al. Genome Res.* **20,** 1297-303 (2010).
7. Iseli, C., Ambrosini, G., Bucher, P. & Jongeneel, C.V. I *PLoS One* **2,** e579 (2007).
8. Wolfe, K.H., Li, W.H. & Sharp, P.M. *Proc. Natl. Acad. Sci. USA* **84,** 9054–9058 (1987).

9. Ossowski, S. *et al. Science* **327,** 92–94 (2010).
10. Yang, S. *et al. Nature* **523,** 463-467 (2015).
11. Scofield, D.G. *Am. J. Bot.* **93,** 1740-1747 (2006).
12. Romberger, J.A., Hejnowicz, Z. & Hill, J.F. Plant structure: function and development. (Springer-Verlag, 1993).
13. Kwiatkowska, D. *J. Exp. Bot.* **59,** 187-201 (2008).
14. Burian, A., Barbier de Reuille, P. & Kuhlemeier, C. *Curr. Biol.* **26,** 1385-1394 (2016).
15. Edwards, P.B., Wanjura, W.J., Brown, W.V. & Dearn, J.M. *Nature* **347,** 434 (1990).
16. Padovan, A., Lanfear, R., Keszei, A., Foley, W.J. & Külheim, C. *BMC Plant Biol.* **13,** 29 (2013).
17. Watson, J.M. *et al. Proc. Natl. Acad. Sci.* USA **113,** 12226–12231 (2016).
18. Behjati, S. *et al. Nature* **513,** 422–425 (2014).
19. Lodato, M.A. *et al. Science* **350,** 94–98 (2015).
20. Van der Auwera, G.A. *et al. Curr. Protocols Bioinfo.* **43,** 11.10.1-11.10.33 (2013).
21. Myers, E.W. & Miller, W. *Comput. Appl. Biosci.* **4,** 11-17 (1988).
22. Li, H. *et al. Bioinformatics* **25,** 2078-9 (2009).

Acknowledgements

This work was funded by the University of Lausanne through a supportive grant from the University rectorate and by the Swiss National Science Foundation (Agora Grant CRAGI3_145652). The Pacific Biosciences RS II sequencing was performed at the Lausanne Genomic Technologies Facility (GTF). The purchase of the GTF's RS II instrument was financed in part by the *Loterie Romande* through the *Fondation pour la Recherche en Médecine Génétique.* We thank Keith Harshman, Johann Weber, and Mélanie Dupasquier from the GTF for sequencing. We thank Cris Kuhlemeier for sharing unpublished results, Jean Tercier for tree-ring analysis, Transistor communication for graphical production of the 3D oak, Woodtli+Leuba SA for sample collection, Nicolas Guex for advice on SNV calling and Jean-Jacques Strahm and Marco Bonetti for providing oak images.

Author contributions

L. F. sequenced the genome. E.S.-S., S.C., M.P. assembled and annotated the genome. N.S., E.S.-S., C.I. identified SNVs. C.G.-D., J.C. extracted DNA and confirmed SNVs. E.S.-S., M.R.-R. analyzed genome duplication. P.C. produced cross-sections of oak apical meristems.

M. S. established a list of DNA repair genes. F. S. provided statistical help with the analyses. J.V., M.J. produced a 3D model of the oak tree. C.H., C.F., L.K., I.X., M.R.-R., J.P., A.R., P.R. conceived the project and wrote the manuscript.

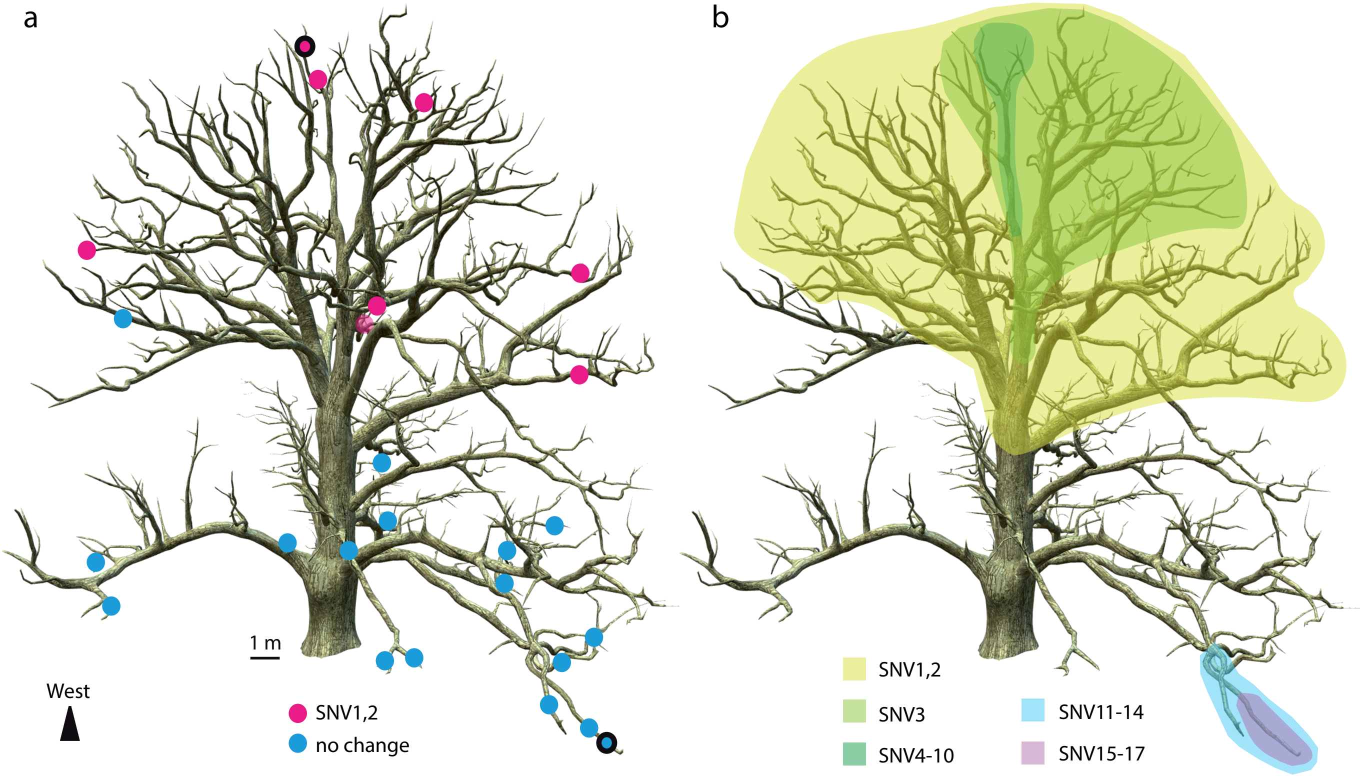
Additional information

Supplementary information is available for this paper.

Correspondence and requests for materials should be addressed to P. R.

Figure Legends

Figure 1 | Distribution of somatic mutations in the Napoleon Oak. a, The genome of two leaf samples (outlined dots) was sequenced to identify single-nucleotide variants (SNV). 17 SNVs were confirmed and analysed in 26 other leaf samples to map their origin. A reconstructed image of the Napoleon Oak shows similar location of two SNVs (magenta dots) on the tree. Blue dots represent genotypes that are non-mutant for these SNVs. Three non-mutant samples are not visible on this projection. Location of other SNVs can be found in Supplementary Figure 3. **b,** Location of all identified SNVs. Sectors of the tree containing each group of SNVs are represented by different colours.

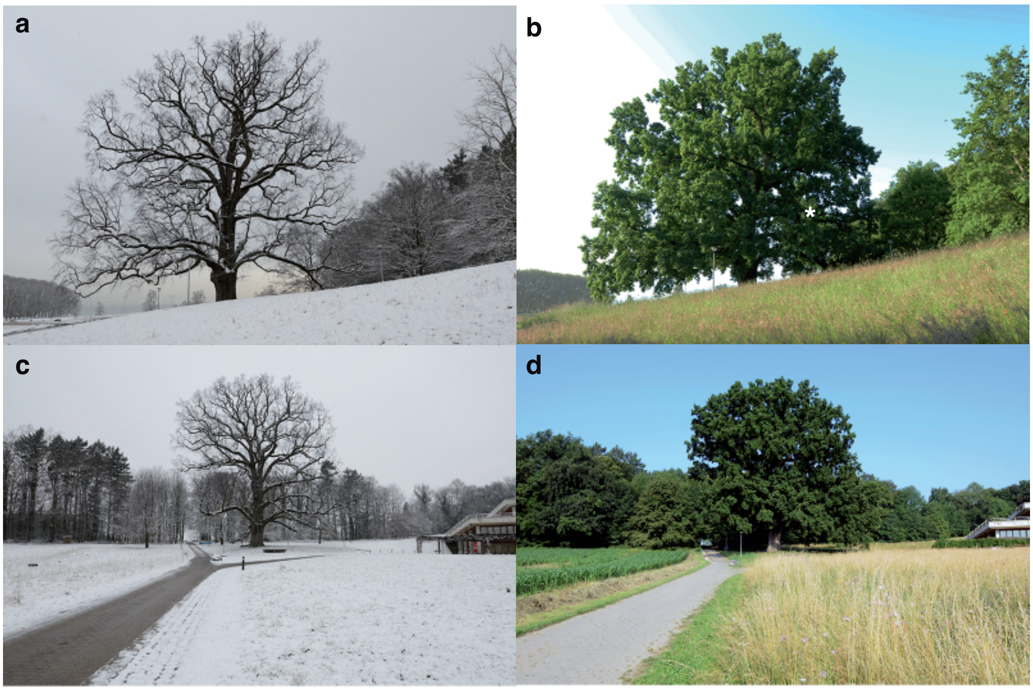

a
1 m
West
SNV1,2
no change
b
SNV1,2
SNV3
SNV4-10
SNV11-14
SNV15-17

Low number of fixed somatic mutations in a long-lived oak tree

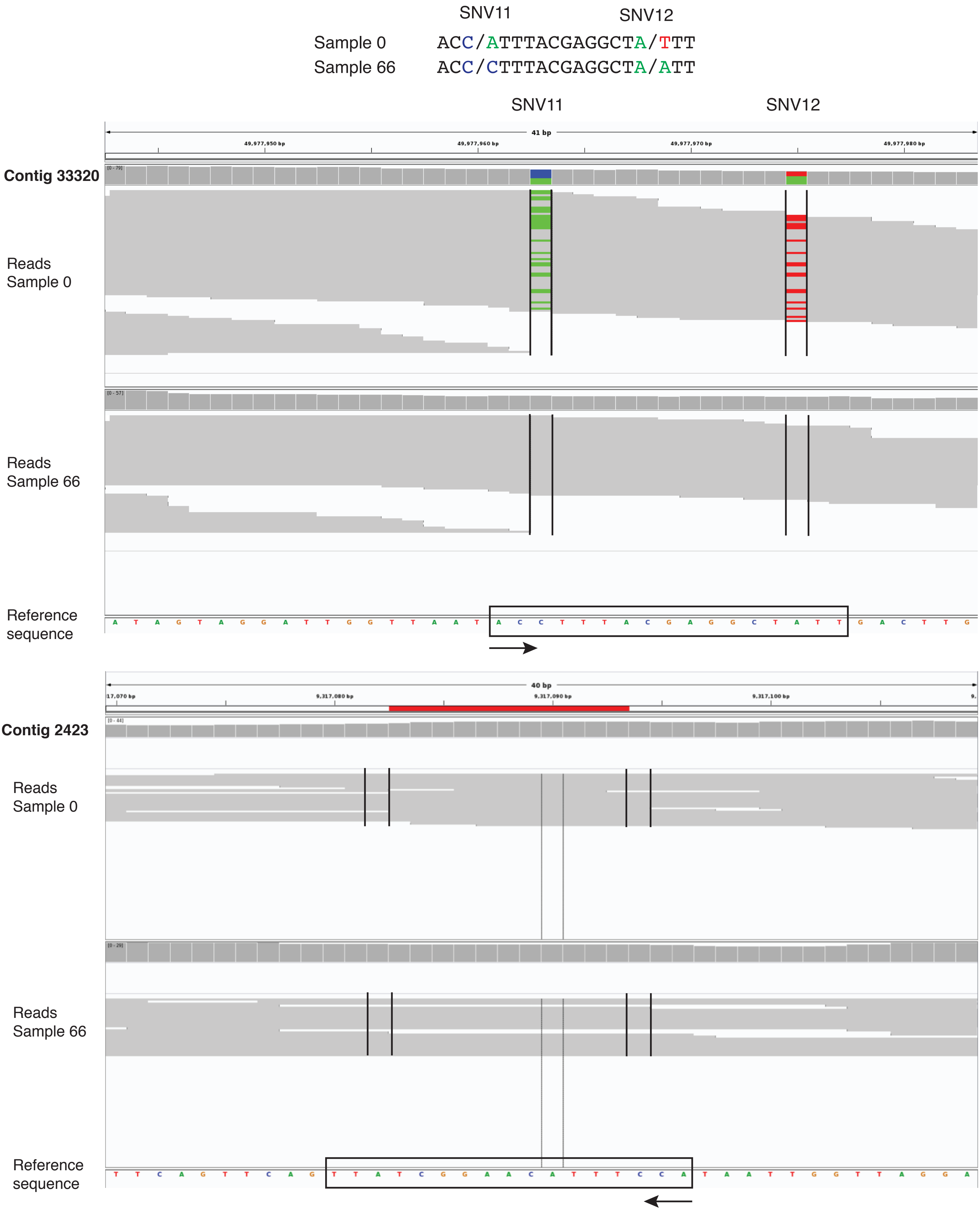
Emanuel Schmid-Siegert[1,†], Namrata Sarkar[2,3,4,†], Christian Iseli[1,†], Sandra Calderon[1], Caroline Gouhier-Darimont[5], Jacqueline Chrast[2], Pietro Cattaneo[5], Frédéric Schütz[2], Laurent Farinelli[6], Marco Pagni[1], Michel Schneider[7], Jérémie Voumard[8], Michel Jaboyedoff[8], Christian Fankhauser[2*], Christian S. Hardtke[5*], Laurent Keller[3*], John R. Pannell[3*], Alexandre Reymond[2*], Marc Robinson-Rechavi[3,4*], Ioannis Xenarios[1,2,7*] and Philippe Reymond[5*]

[1]Vital-IT Competence Center, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [2]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland, [3]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland [4]Evolutionary Bioinformatics Group, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [5]Department of Plant Molecular Biology, University of Lausanne, 1015 Lausanne, Switzerland, [6]Fasteris SA, 1228 Plan-les-Ouates, Switzerland, [7]Swiss-Prot group, Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland, [8]Risk Analysis Group, Institute of Earth Sciences, University of Lausanne, 1015 Lausanne, Switzerland, [†]These authors contributed equally.

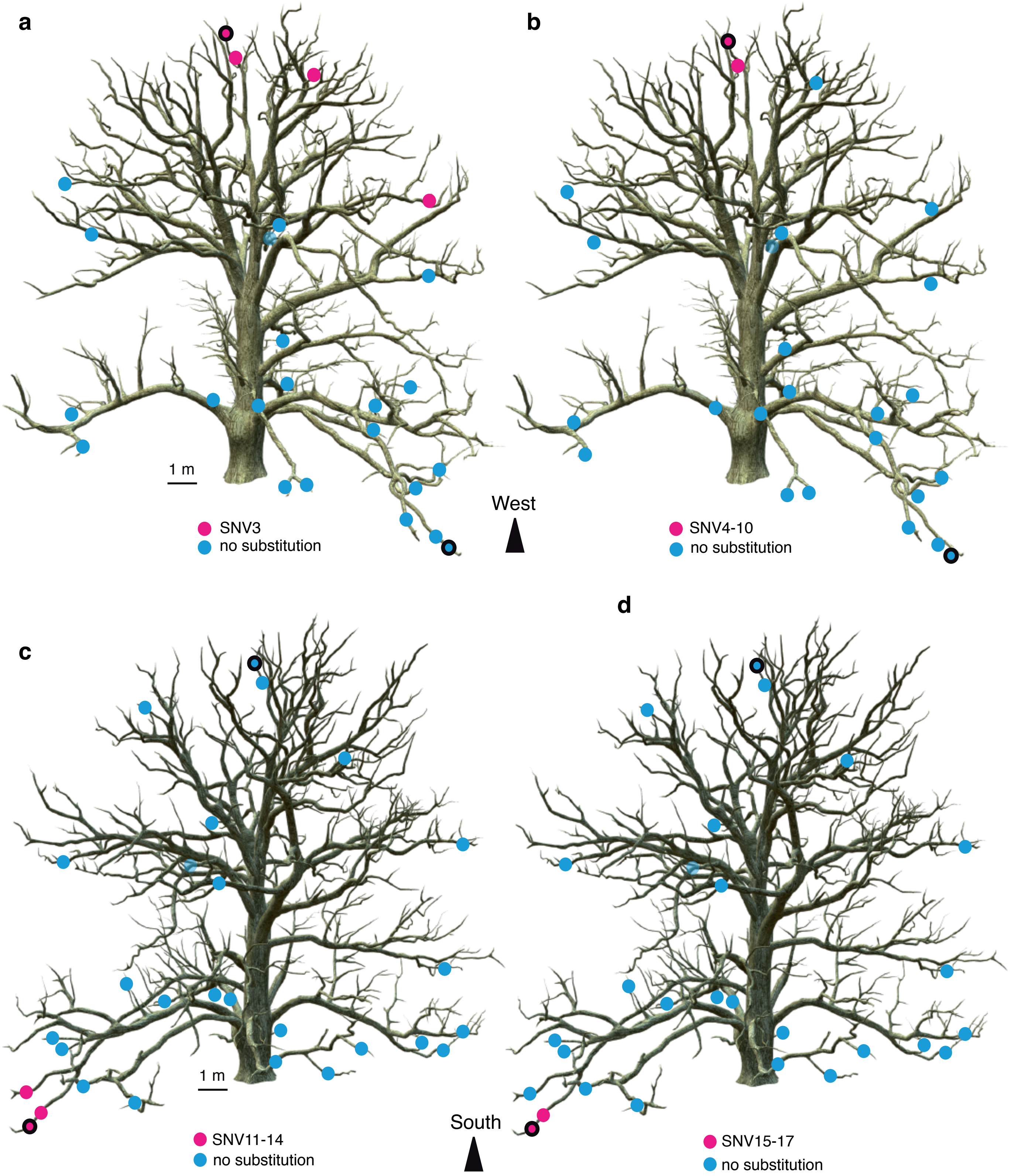
[*]Correspondence: christian.fankhauser@unil.ch, christian.hardtke@unil.ch, laurent.keller@unil.ch, john.pannell@unil.ch, alexandre.reymond@unil.ch, marc.robinson-rechavi@unil.ch, ioannis.xenarios@sib.swiss, philippe.reymond@unil.ch



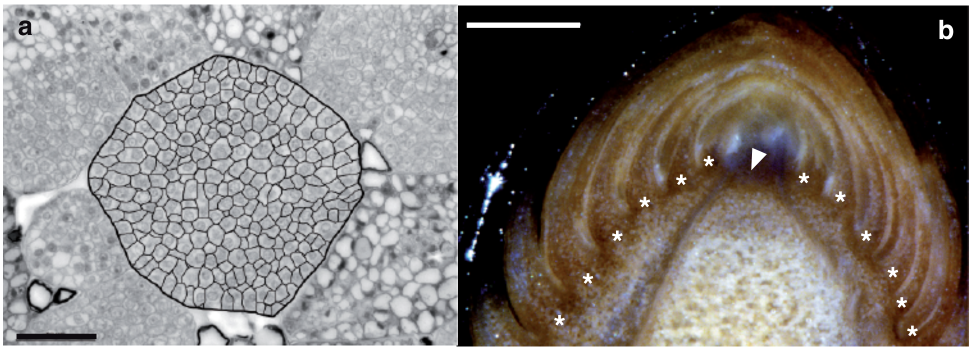
Supplementary Figure 1 | Napoleon Oak. Photographs of the Napoleon Oak on the Lausanne University campus taken in winter and summer. **a, b,** South view. **c, d,** North-West view.



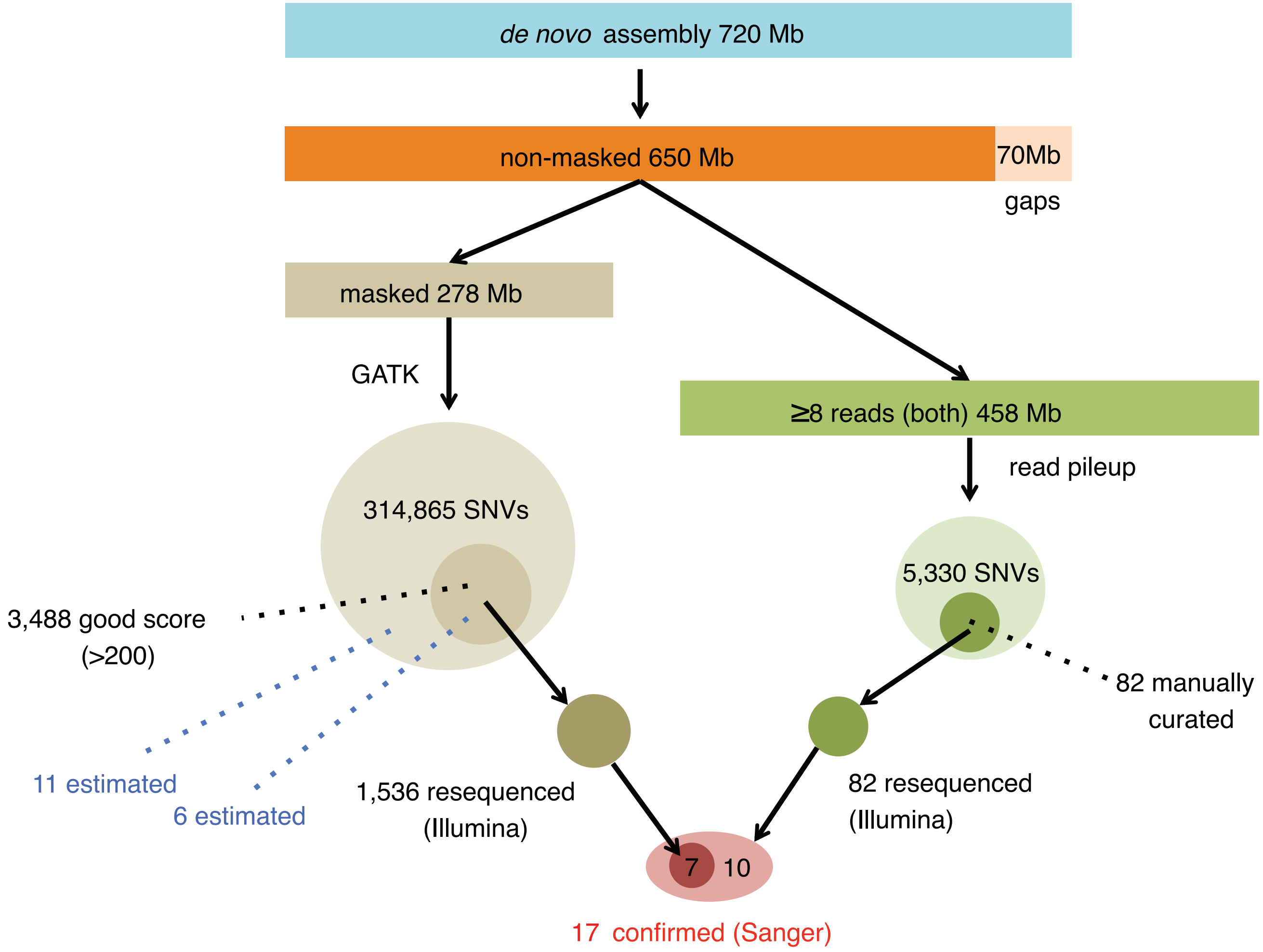
Supplementary Figure 2 | Read alignment for SNV11 and SNV12. The region on Contig 33320 where two consecutive SNVs were identified is shown with read alignment for both genome samples (top). Positions in reads that differ from the reference sequence are colored according to the base identity. A region on Contig 2423 with high similarity is shown but does not contain SNVs (bottom). Sequence orientation is indicated by arrows.



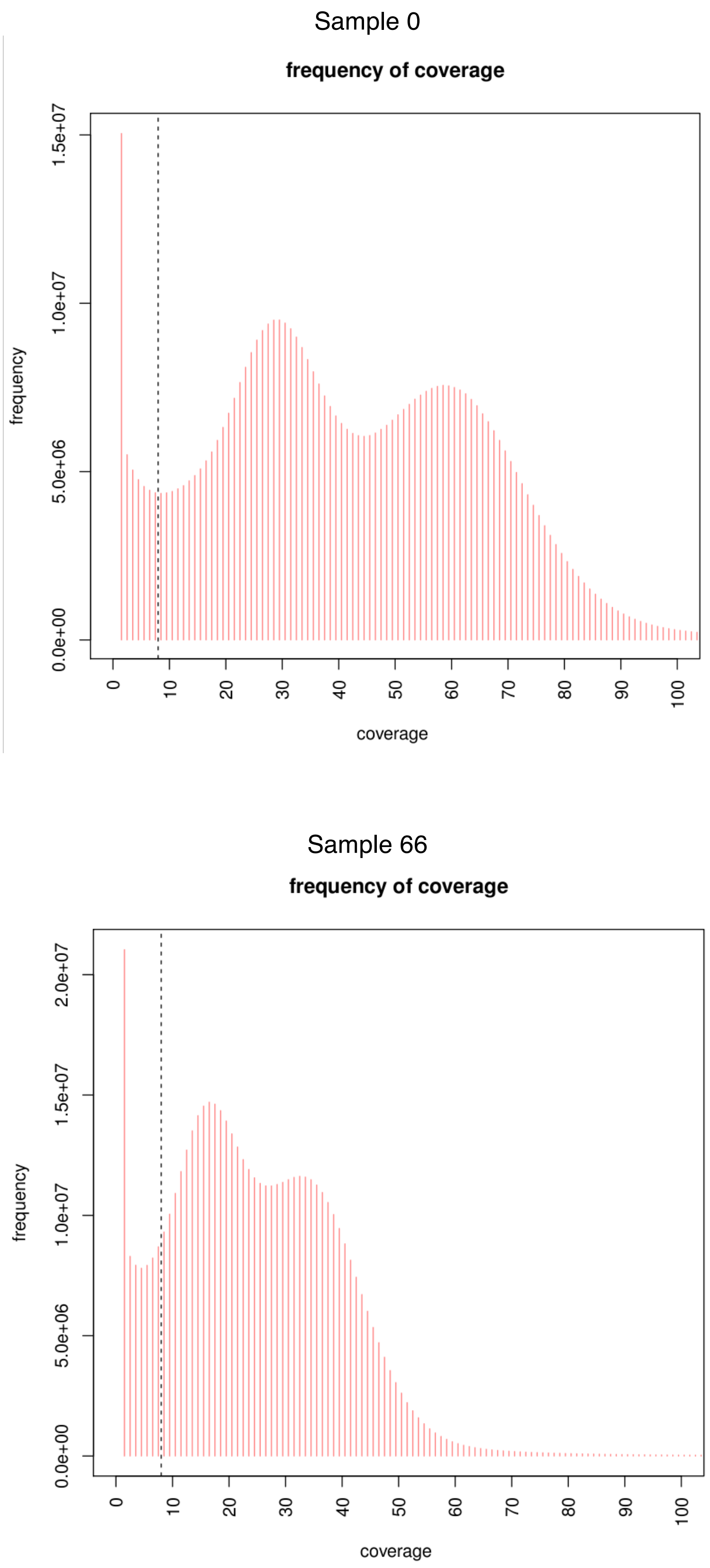
Supplementary Figure 3 | Distribution of somatic mutations in the Napoleon Oak. The genome of two leaf samples (outlined dots) was sequenced to identify single-nucleotide variants (SNV). 17 SNVs were confirmed and analysed in 26 other leaf samples to map their origin. **a-d,** Reconstructed images of the Napoleon Oak show the location of different SNVs (magenta dots) on the tree. Blue dots represent genotypes that are non-mutant for these SNVs.



Supplementary Figure 4 | Napoleon Oak apical meristem. a, Cross-section of an apical meristem. Meristematic cells are delineated. Surrounding cells belong to leaf-like structures surrounding the meristem. Scale bar, 50 µm. **b**, Longitudinal section of an apical bud. Apical meristem (arrowhead) is surrounded by leaf-like structures (stars). Scale bar, 500 µm.

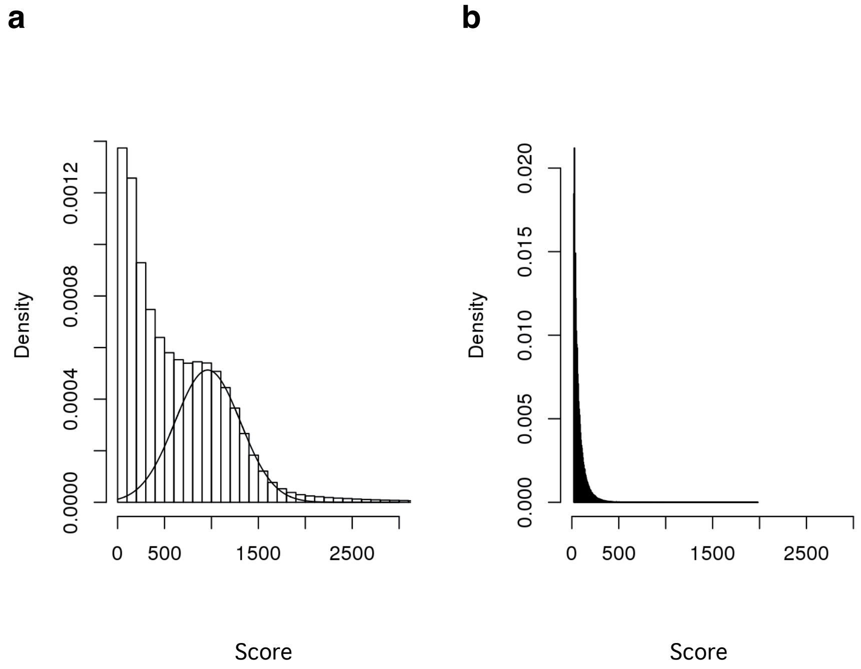


Supplementary Figure 5 | Flowchart of SNV identification methods.


Sample 0

Sample 66

Supplementary Figure 6 | Read coverage. The coverage distribution of Illumina reads used for SNV calling is shown for Samples 0 and 66. The dashed line represents the 8 X cutoff used for the analysis.



Supplementary Figure 7 | Distribution of variants. a, Distribution of the confidence scores of the 1'832'554 heterozygous sites common to both samples 0 and 66. **b,** Distribution of the confidence scores of 165'489 heterozygous sites in sample 66 that are homozygous in sample 0.

!

!

!

!

!

!

!

!

!

!

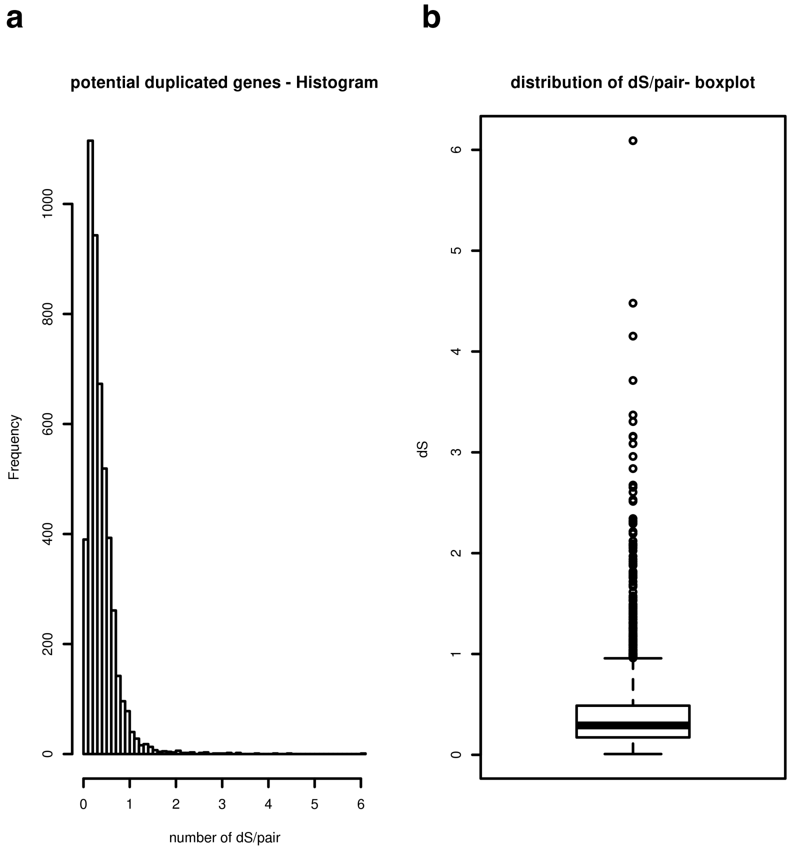
!

!

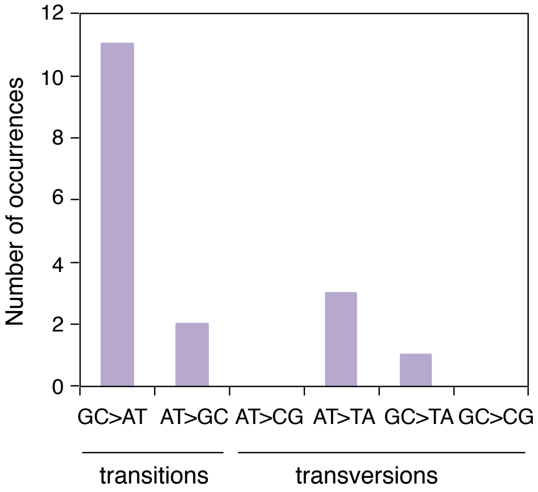
!

!

!



Supplementary Figure 8 | Analysis of oak genome duplication. a, Frequency plot and **b**, box plot showing the distribution of synonymous distances (dS) on a stringent set of 4,777 paralog pairs. This analysis was done with a threshold BLAST E-value <1e-10 and by removing multigene families of more than 20 members.



Supplementary Figure 9 | Spectrum of somatic mutations between two Napoleon Oak genomes. The type of substitution for 17 confirmed oak SNVs is shown.

Supplementary Table 1: *Quercus robur* genome statistics.

Genome	
Total genome length (bp)	719,779,348
Number of scaffolds	85,557
Maximum scaffold length (bp)	317,245
NG50 based on 740 Mbp (bp)	17,014
Gaps (%)	9.52
Masked (%)	39.84
Genes	
Average length (bp)	2,360
Maximum length (bp)	47,221
Average intron length (bp)	740
Average exon length (bp)	232
Proteome	
Total predicted proteins	49,444
Full proteins	44,096
Partial proteins	5,348
Nb proteins with orthologous in *Glycine max*	39,656
Nb orthologous in *Glycine max* + functional annotation	16,323
Nb orthologous in *Glycine max* + function via ATH	23,333

Supplementary Table 2. SNVs in the Napoleon Oak.

SNV	Contig	Lower branch genome (0)	Upper branch genome (66)	Context	Position
SNV1	Contig12293_20040	TCTGA	TCT/CGA		Exon (R→G)
SNV2	Contig8610_5366	AACAG	AAC/TAG	CpNG	Intron
SNV3	Contig17717_5512	ACCAT	ACC/TAT	dipyrimidine	Non coding
SNV4	Contig19224_2528	TACAT	TAC/TAT		Non coding
SNV5	Contig3344_66711	AACGC	AAC/TGC	CpG	Non coding
SNV6	Contig420_15205	CTTGA	CTT/AGA		Non coding
SNV7	Contig46021_5283	TCCTA	TCC/TTA	dipyrimidine	Non coding
SNV8	Contig4756_544	AAGGT	AAG/AGT	dipyrimidine	Intron
SNV9	Contig61424_5311	ATTTG	ATT/ATG		Non coding
SNV10	Contig79811_6871	AACAA	AAC/TAA		Non coding
SNV11,12	Contig33320_1101-13	ACC/ATTTACGAGGCTA/TTT	ACCTTTACGAGGCTATT		Non coding
SNV13	Contig1217_8596	TCG/AGG	TCGGG	dipyrimidine	Non coding
SNV14	Contig15467_11236	AGG/AAT	AGGAT	dipyrimidine	Intron
SNV15	Contig4515_9475	GTC/TGT	GTCGT	CpG, dipyrimidine	Intron
SNV16	Contig28929_3009	TTT/CGG	TTTGG		Non coding
SNV17	Contig32076_9167	TGG/AGC	TGGGC	dipyrimidine	Non coding

Mutated bases are shown in red. Homozygous sites generate heterozygote sites.

Supplementary Table 3. Orthology and duplication of DNA repair genes in oak and peach trees.

Branch of the phylogeny	DNA repair genes			All genes		
	duplicated	total	% total	duplicated	total	% total
Quercus robur	0	228	0.0	258	10,199	3.5
Prunus persica	5	228	2.2	860	16,004	5.4
Quercus - Prunus common ancestor	1	228	0.4	523	8,474	6.2

All gene counts are for *Arabidopsis thaliana* orthologs; the number of "all genes" varies according to the number of orthologs detected for each set (i.e., from *Quercus robur*, from *Prunus persica*, or shared). All orthology detection and lineage-specific duplication calls are from OMA (see Experimental Procedures)

Supplementary Table 4. List of primers used for genotyping and Sanger sequencing

SNV	Contig	Forward primer (5'-3')	Reverse primer (5'-3')
SNV1	Contig12293_20040	CCCTTGCCTGTAAGGAATCA	TGCTATGCTTGGAAAAACCA
SNV2	Contig8610_5366	GGCTGAACAAAGTTGAGTGGA	TGTAAGCCCTCATCCCATGT
SNV3	Contig17717_5512	CAACGAACTCACAGGACGTG	AGCTTTGTCATCAGCCTTCAG
SNV4	Contig19224_2528	CTTTTTACAATGCCCCCAGA	AAATGCAAGACATCGCTCCT
SNV5	Contig3344_66711	AGAAAATGTGGACGCTGACC	GCCGTATTGTTGTTGGGAAC
SNV6	Contig420_15205	CGAGCATTGATCGAATACCA	TGTGGCCATCCAAGATTAAA
SNV7	Contig46021_5283	AACTGTCGAGCATTGGGTTT	GGATTGCCAAAAGGAGGAAT
SNV8	Contig4756_544	GGCAGGCAGAGACACAAACT	GGAGAGTGGTGGGAATTTGA
SNV9	Contig61424_5311	GCATCGACCAACTGGTTTTT	CAGTTGCCCTCCATTTGATT
SNV10	Contig79811_6871	CCCAAAAAGTTCCAGCTCAG	ATGACGACTAAGGGCGTGTT
SNV11	Contig33320_1101	GATTGGATGTGGGATCCTTG	GGCAATTTCACTACCCTTGG
SNV12	Contig33320_1113	GATTGGATGTGGGATCCTTG	GGCAATTTCACTACCCTTGG
SNV13	Contig1217_8596	CGACAGATGCTGCTATCGAG	AACGATGAAGATCAGGAAGCA
SNV14	Contig15467_11236	TCTGTGATCCACGTGTTGGT	GGCGCCTAAACAAGTCTCAG
SNV15	Contig4515_9475	TTGGCCTATATTTGAAACCAAT	AGTCGGCAAATCCAAAATTC
SNV16	Contig28929_3009	AGCACCCGATAAGCTCAAAA	GTCTTCAGCTCTGCCACCTC
SNV17	Contig32076_9167	TTCATTGCAATTTCCACAGG	TCATCATCCAAGCCTGACG

Supplementary Methods

Genome assembly. For sample 0, a paired-end library generated 2 x 151,194,704 reads (coverage 40X) and a mate-pair library generated 2 x 107,264,298 reads (coverage 29X). For sample 66, a paired-end library generated 2 x 158,505,474 reads (coverage 42X) and a mate-pair library generated 2 x 124,076,608 reads (coverage 33X). These reads were filtered and trimmed prior assembly using Trimmomatic (v0.3; leading:3, trailing:3, slidingwindow:4:15, minlen:36, custom adapter library)[23] and assembled using SOAPdenovo2 (v2.04.240, kmer 49)[24]. In a second step the assembly was scaffolded with mate-pairs using the same program. The assembly was further scaffolded with long single-molecule PacBio reads (22 SMRT cells, XL-C2 and P4-C2 chemistry, coverage 19X) and the program AHA (http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/; SMRTPipe 2.0.1 manually driven, settings (5,2,50,70), no gap-filling). Assembled sequences <1000 bp were removed to facilitate further analysis. The genome was extended with all paired-end libraries and SSPACE[25] (v2.0, -x = 1,z = 0,-k = 5,-a = 0.7,-n = 15,-T = 20,-p = 0,-o = 20,-t = 0,-m = 32,-r = 0.9) and gaps were filled using Gapfiller (v1.10, all paired-end libraries)[26].

We screened the paired-end libraries for potential non-oak sequences using metaphlan (v1.7.7)[27]. Based on metaphlan results, reference genomes were obtained for the non-oak genomes and the oak scaffolds were filtered against these using blast (ncbi-blast v2.28, >90% sequence identity and E-value <1e-5). The genome was next scaffolded again using the PacBio reads and PBJelly (v14.1.14)[28]. If not further specified, programs were used with their standard settings.

Gene prediction and annotation. Repetitive elements were analysed by first generating a specific repeat model using RepeatModeler (http:/www.repeatmasker.org; v1.0.7, -engine wublast). Repetitive regions in the genome were subsequently masked with the obtained model using RepeatMasker (http:/www.repeatmasker.org; v4.0.3). Genes were predicted by generating a *Q. robur* specific gene prediction model for Augustus (v3.0.1)[29], as described in Tran et al.[30]. Instead of RNAseq reads, we used the UniProtKB reference proteome of *Glycine max* mapped with the splice aware mapper exonerate (V2.2.0, model protein2genome, geneseed 250 –minintron 20, --maxintron 20000)[31]. Using this model we predicted genes and subsequently their encoded proteins for the hard-masked version of the genome (settings: no hints, no UTR predicted, no alternative transcripts). Non-coding elements were annotated using RFAM (v1.5; infernal 1.0.2; blast 2.2.26; hmmer 3.1b1)[32] in the genome with coding regions masked but repetitive elements unmasked. The predicted proteome was annotated based on homology using the FASTA toolkit (http://www.ebi.ac.uk/Tools/sss/fasta/; v36.3.5e) as following: proteins from the *Glycine max* proteome were first mapped with ggsearch (-b 1 -d 0 -E 1e-5 -m 8 -T 10); proteins that did not map were mapped in a next step with glsearch (-b 1 -d 0 -E 1e-5 -m 8 -T 10) and finally the rest with ssearch (-b 1 -d 0 -E 1e-5 -m 8 -T 10). The functional protein annotation was overtaken from *Glycine max*. For proteins with unknown function in *Glycine max,* we extended the annotation using the OMA database (www.omabrowser.org) and orthologous proteins from *Arabidopsis*. PFAM[33] was used additionally to obtain functional domain annotations for the proteome and the concatenated proteome annotation was transferred onto the oak genome.

PCR-seq. A modification of the published RT-PCR-seq method[34] was used. Briefly, pairs of primers for 50-150 bp amplicons containing the targeted sequence were designed using Primer3. Touchdown PCR amplification was performed in a final volume of 12.5 ml with JumpStart REDTaq ReadyMix (Sigma-Aldrich), a primer concentration of 0.4 mM and 2 ng of gDNA per reaction in 384-well plates. Equal volumes of PCR products were pooled for each DNA template (sample 0 and 66). One ml of each pool was then purified with the QIAquick PCR Purification Kit (Qiagen) following the manufacturer's instructions. The

KAPA LTP Library Preparation Kit (Kapa Biosystems) was used, starting with 500 ng of purified PCR products, to create a library compatible with an Illumina sequencing platform. Clean-ups between enzymatic steps were performed with Nucleospin PCR Clean-up columns (Macherey-Nagel). After ligation of pentabase adapters, libraries were run on a 2 % agarose gel and extracted using the MinElute Gel Extraction kit (Qiagen). Libraries were sequenced on HiSeq 2000 after six cycles of amplification (Lausanne Genomic Technologies Facility). Amplicon reads were aligned, with no mismatches allowed, to a compendium of the expected amplimers that bore the reference allele, the alternate allele identified in the heterozygote sample, as well as the remaining two nucleotides at the variable position; this allowed an unbiased estimation of the error rate generated by the sequencing itself. As this method might have missed *bona fide* changes between the two sampled branches that present other heterozygous sites close by, we also aligned amplicon sequencing reads directly to the reference genome, with mismatches allowed.

Estimation of the possible missed SNVs. About half of sites that were heterozygous in only one sample and had a confidence score ≥200 were assessed experimentally by PCR-seq (1,536 out of 3,488 sites). Given the confidence scores of the tested sites, we estimated that we missed fewer than 6 SNVs in the sites not evaluated by PCR-seq. We then evaluated the number of true positives missed within candidates with confidence scores <200. We fitted a mixture of two distributions, including a normal distribution that should fit the correct calls, modelled on the 1,832,554 sites that were predicted to be heterozygous in both samples (Supplementary Figure 7a). Applying this distribution to the data for sites that are homozygous on sample 0 and heterozygous on sample 66 (case 1, Supplementary Figure 7b), we find that the distribution of correct calls is insignificant compared to the rest. In details, when fitting this normal distribution to the data, the expected number of correct calls with a score < 200 is 5.24. Extrapolating this calculation for sites that are heterozygous on sample 0 and homozygous on sample 66 (case 2), we estimate that we have missed fewer than 11 true SNVs for both cases. We thus estimate a total of 17 missed SNVs (6 with a score ≥200 and 11 with a score <200). Note that we did not assess the presence of larger somatic changes such as copy number variants, small indels, and transposition events.

Estimation of the false negative rate. A few recent studies have tried to estimate the false negative rate of SNV calling for large genomes assembled with short read sequences[10,35]. The main method used in those studies was to introduce simulated SNVs into the data, and check how well they were recovered. To this end, we introduced 500 SNVs in each of the sequenced oak genome (sample 0 and 66). The BAM file from the original SNV call (fetchWGI) with mapped reads from sample 0 and 66 was used for this analysis. The information track from the coverage analysis identified regions in the genome which contained >= 8x coverage for both samples, suitable for SNV calling with our method (bedtools intersect v2.26). Regions that were unambiguously homozygous in both samples were identified by a pile-up using samtools (v1.3, -u -BQ0 -d10000000 -v). This restricted genome space with >= 8x coverage and 100% homozygous reference for both samples was split into single nucleotide annotation using bedops[36] (v2.4.28, --chop 1) and 500 random positions were extracted in each sample using Sample[37] (v1.0.3). To each of the 1000 positions we added a random SNV frequency between 30% and 100% following a gamma-distribution with similar characteristics than the original called SNVs (using R, fitdistrplus[38], v1.0-9).

Two BAM files were created containing reads from sample 0 with 500 simulated SNVs and reads from sample 66 with the other 500 simulated SNVs, using BAMSURGEON[39] (v1.0,addsnv,-d 0.7 –mindepth 8). This successfully generated a “true set” of SNVs for 466 and 460 sites, respectively, as evaluated with BAMSURGEON (makevcf), which discarded some sites due to technical issues within the inserted region. Next, SNVs were called between sample 0 + 466 SNVs and sample 66 and, similarly, between sample 66 + 460 SNVs and sample 0, using the same strategy as for the original SNV analysis. The overlap between called SNVs and the true set was evaluated using bedtools (intersect). Of 466 SNVs simulated in sample 0, 421 (90.3%) were recovered, whereas of 460 SNVs simulated in sample 66, 331 (72.0%) were recovered.

Whole-genome duplication. Simple clustering based on homology, (i.e., clustering the predicted proteins by identity, CD-HIT, min 90% similarity), retrieved 1,098 proteins that have a >90% identity to another protein, which is not suggestive of recent whole genome duplication. Whole genome duplication should lead to an excess of relatively old paralogs, whereas small-scale duplicates are expected to be enriched in very recent paralogs. This can be estimated from the distribution of synonymous distances (dS)[40,41]. We computed the dS on a stringent set of 4,777 paralog pairs with BLAST E-value <1e-10, removing large multigene families (more than 20 members). The distribution of dS values is clearly unimodal, with an excess of low dS values (i.e., young paralogs, Supplementary Figure 8). This also does not support a recent whole genome duplication in the oak lineage.

To address the possibility of a more ancient duplication event, we compared our oak genome reference with itself using "BLAST all versus all" as suggested in Panchy et al.[42], (i.e., similarity ≥30%, match length ≥150AA and E-value ≤1e-5). Following this procedure we have 49,444 proteins, of which 3,650 are duplicated (7.4%), 2,070 are triplicated (4.1%) and 23.7% are present in more copies with diminishing frequency. In summary, a total of 17,474 oak proteins out of 49,444 appear to be duplicated (35%), which is less than that reported for closely related species (e.g. *Medicago sativa* has about 50,000 genes of which >75% are duplicated, according to Panchy et al.[42]). We then assessed whether the similarity identified above was local, properties of similar domains, or extended along the entire protein, indicative of duplicated proteins. We found only 973 oak proteins that have duplications extending over their entire lengths. In summary, it is possible that the oak genome underwent duplication, as suggested by Panchy et al.[42], but this event appears to be rather old, as we have very few (<3%) duplicated genes with very high similarity (>90%) and no second peak in the dS distribution (Supplementary Figure 8). It seems unlikely that such a duplication event should compromise the identification of *bona fide* variants. Note that if the duplication would have hindered the capacity to detect these variants, they would not be found in nested sectors of the tree but rather in all 26 samples assessed.

Analysis of DNA repair genes. Orthologs between *Arabidopsis*, *Prunus persica* (peach) and *Q. robur* were called using the OMA database[43]. One-to-many orthologs, e.g., between *Arabidopsis* and *Q. robur*, represent duplication in the oak lineage since the divergence from *Arabidopsis*; they are also known as in-paralogs of oak. We classified these in-paralogs according to whether the duplication was shared by *P. persica* and *Q. robur* (i.e., one copy in *Arabidopsis* relative to several copies in both the peach and oak genomes), or whether it was peach- or oak-specific (i.e., one copy in *Arabidopsis* and peach, relative to several copies in oak). The number of duplicates was reported as the number of genes that could be called duplicate (i.e., the number of orthologs between each tree genome and *Arabidopsis*, Supplementary Table 5). We then manually compiled a list of *Arabidopsis* genes involved in DNA repair from SwissProt/UniProtKB annotations (Supplementary Table 6). We then counted specifically the number of duplicates for genes involved in DNA repair and reported this as the number of orthologs associated with this function (Supplementary Table 3 and 7).

Supplementary Discussion

We found that G:C→A:T transitions were the most frequent class of SNVs observed in the Napoleon Oak (Supplementary Figure 9). Ultraviolet (UV) light causes G:C→A:T transitions at dipyrimidine sites in plants[44]. Among the 11 G:C→A:T transitions that we observed, seven were in a dipyrimidine context (Supplementary Table 2). In addition, spontaneous deamination of methylated cytosine leads to thymine change at CpG or CpNG sites [22]. However, there were only three G:C→A:T transitions in such a context (Supplementary Table 2). It thus seems plausible that UV light may have caused most of the G:C→A:T transitions we observed, although other factors, such as cytosine deamination and replication errors, may account for other SNVs. Although the oak lineages sampled have not been separated by any meiosis events, which in yeast was found to elevate the generational mutation rate[45], they have been exposed to the natural environment, which in *Arabidopsis* is known to significantly enhance mutation rate when compared to a controlled lab environment[46]. However, a study of mutation accumulation lines in Arabidopsis showed that after 30 generations the majority of somatic mutations were UV-induced G:C→A:T transitions, suggesting that the contribution of meiosis-induced changes in plants is limited[9].

Our results throw new light on explanations proposed for differences in the distribution of mating systems between short- and long-lived plants. While many annuals and short-lived plants have undergone evolutionary transitions from outcrossing to selfing[47], often involving a loss of self-incompatibility systems[48], long-lived woody species are more likely to be fully outcrossing[49], including oaks[50]. Theoretical analysis indicates that a high somatic mutation rate could account for this difference, because somatic mutations would contribute to the genetic load of the population and thus to inbreeding depression, disfavouring self-fertilization[1]. Inbreeding depression is indeed higher in long-lived woody species than annuals[51], and the observation of higher inbreeding depression caused by within-branch than between-branch selfing points to the accumulation of different deleterious somatic mutations in different sectors of the plant[3]. However, our finding now challenges the notion that the breeding system of long-lived trees is constrained by a high rate of somatic mutations.

The results of our study, in conjunction with those of Burian et al.[14]*,* have important implications for how we should view one of the most fundamental ways in which plants differ from animals – their absence of a germline. In oak, iterative growth of axillary meristems produces terminal branches that carry stem cells. As in other plants, favourable conditions induce stem cells to produce floral buds and ultimately the gametes of the next generation. These stem cells are functionally analogous to germ cells in metazoans and result from a limited number of divisions that prevent an accumulation of replicative errors.

References

23. Bolger, A.M., Lohse, M. & Usadel, B. *Bioinformatics* **30,** 2114-2120 (2014).
24. Luo, R. *et al. GigaScience* **1,** 18 (2012).
25. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. S. *Bioinformatics* **27,** 578–579 (2011).
26. Boetzer, M. & Pirovano, W. *Genome Biol.* **13,** R56 (2012).
27. Segata, N. *et al. Nature Methods* **9,** 811-814 (2012).
28. English, A.C. *et al. PLoS One* **7,** 11 (2012).
29. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. *Nucl. Acids Res.* **32,** W309-312 (2004).
30. Tran, V.D. *et al. mSystems* **1,** e00036-16 (2016).
31. Slater, G.S.C. & Birney, E. *BMC Bioinformatics* **6,** 31 (2005).
32. Gardner, P.P. *et al. Nucl. Acids Res.* **37,** D136-140 (2009).
33. Sonnhammer, E.L.L., Eddy, S.R. & Durbin, R. *Proteins* **28,** 405-420 (1997).
34. Howald, C. *et al. Genome Res.* **22,** 1698-1710 (2012).
35. Keightley, P.D., Ness, R.B., Halligan, D.L. & Haddrill, P.R. *Genetics* **196,** 313-320 (2014).
36. Neph, S. *et al. Bioinformatics* **28,** 1919-1920 (2012).
37. https://github.com/alexpreynolds/sample
38. Delignette-Muller, M.L. & Dutang, C. *J. Stat. Softw.* **64,** 1-34 (2015).
39. Ewing, A.D. *et al. Nature Methods* **12,** 623–630 (2015).
40. Lynch, M. & Conery, J.S. *Science* **290,** 1151–1155 (2000).
41. Vanneste, K., Van de Peer, Y. & Maere, S. *Mol. Biol. Evol.* **30,** 177-190 (2013).
42. Panchy, N., Lehti-Shiu, M. & Shiu, S.-H. *Plant Physiol.* **171,** 2294-2316 (2016).
43. Altenhoff, A.M. *et al. Nucl. Acids Res.* **43,** D240-D249 (2015).
44. Britt, A.B. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47,** 75–100 (1996).
45. Rattray, A., Santoyo, G., Shafer, B. & Strathern, J. N. *PLoS Genet.* **11,** e1004910 (2015).
46. Rutter, M.T., Shaw, F.H. & Fenster, C.B. *Evolution* **64,** 1825-1835 (2010).
47. Stebbins, G.L. *Variation and evolution in plants* (Columbia Univ. Press, 1950).
48. Goldberg, E.E. *et al. Science* **330,** 493-495 (2010).
49. Barrett, S.C.H., Harder, L.D. & Worley, A.C. *Phil. Tran. Royal Soc. London B: Biol. Sci.* **351,** 1271-1280 (1996).
50. Streiff, R. *et al. Mol. Ecol.* **8,** 831-841 (2009).
51. Goodwillie, C., Kalisz, S. & Eckert, C.E. *Annu. Rev. Ecol. Evo. Syst.* **36,** 47-79 (2005).