Low Rate of Somatic Mutations in a Long-Lived Oak Tree

Namrata Sarkar[1,2,3,9], Emanuel Schmid-Siegert[4,9], Christian Iseli[4,9], Sandra Calderon[4], Caroline Gouhier-Darimont[5], Jacqueline Chrast[1], Pietro Cattaneo[5], Frédéric Schütz[1], Laurent Farinelli[6], Marco Pagni[4], Michel Schneider[7], Jérémie Voumard[8], Michel Jaboyedoff[8], Christian Fankhauser[1*], Christian S. Hardtke[5*], Laurent Keller[2*], John R. Pannell[2*], Alexandre Reymond[1*], Marc Robinson-Rechavi[2,3*], Ioannis Xenarios[1,4,7*], Philippe Reymond[5,10*]

[1]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland

[2]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

[3]Evolutionary Bioinformatics Group, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

[4]Vital-IT Competence Center, 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

[9]These authors contributed equally

[10]Lead contact

[*]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@unil.ch, philippe.reymond@unil.ch

SUMMARY

Because plants do not possess a proper germline, deleterious mutations that occur in the soma can be passed to gametes. It has generally been assumed that the large number of somatic cell divisions separating zygote from gamete formation in long-lived plants should lead to many mutations. However, a recent study showed that surprisingly few cell divisions separate apical stem cells from axillary stem cells in annual plants, challenging this view. To test this prediction, we generated and analysed the full genome sequence of two terminal branches of a 234-year-old oak tree and found very few fixed somatic single-nucleotide variants (SNVs), whose sequential appearance in the tree could reliably be traced back along nested sectors of younger branches. Our data indicate that the stem cells of shoot meristems in trees are robustly protected from accumulation of mutations, analogous to the germline in animals.

INTRODUCTION

Accumulation of deleterious mutations is a fundamental parameter in plant ageing and evolution [1, 2]. Because the pedigree of cell division that generates somatic tissue 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 in theory lead to a large number of DNA replication errors [3-5]. Tree architecture is determined by the modular growth of apical meristems, which contain stem cells. These cells divide and produce progenitor cells that undergo division, elongation and differentiation to form a vegetative shoot, the branch. Axillary meristems are formed at the base of leaf axils and are

50 responsible for the emergence of side branches. They are separated from apical meristems by
51 elongating internodes. In oak, early and repeated growth cessation of terminal apical
52 meristems leads to a branching pattern originating from such axillary meristems. In turn,
53 axillary meristems grow out and produce secondary axillary meristems. This process is
54 reiterated indeterminately to produce highly ramified trees of large stature, resulting in
55 thousands of terminal ramets [6, 7].

56 Classical studies of shoot apical meristem organization have reported that the most
57 distal zone has a significantly lower rate of cell division than more basal regions of the apex,
58 and might therefore be relatively protected from replication errors [8, 9]. In a recent study
59 that followed the fate of dividing cells in the apical meristems of *Arabidopsis thaliana* and
60 tomato, Burian et al. [10] showed that an unexpectedly low number of divisions separate
61 apical from axillary meristems. In these herbaceous plants, axillary meristems are separated
62 from apical meristem stem cells by seven to nine cell divisions, with internode growth
63 occurring through the division of cells behind the meristem. The number of cell divisions
64 between early embryonic stem cells and terminal meristems thus depends more on the
65 number of branching events than on absolute plant size. Moreover, only three or four semi-
66 permanent apical stem cells give rise to independent sectors of the growing shoot, so that a
67 mutation in one apical stem cell may propagate and be fixed in a domain of the apical shoot
68 that contains an axillary meristem [10]. Alternatively, mutations that arise in sub-apical stem
69 cells may be partly fixed in axillary meristems and are either lost or fully fixed in secondary-
70 order axillary meristems.

71 In trees, because axillary meristems replace apical meristems iteratively, the
72 cumulative number of cell divisions separating meristems determines the rate of genetic
73 aging and the potential accumulation of somatic mutations. If the same growth pattern
74 described above for *A. thaliana* and tomato applies to trees, their somatic mutation rate might

be much lower than is commonly thought, and the majority of fixed mutations should be found in relatively small sectors as nested sets of mutations. To test these predictions, we sequenced the full genome of two terminal branches of an iconic old oak tree (*Quercus robur*).

RESULTS AND DISCUSSION

Napoleon Oak Genome Sequencing

We conducted our study on an oak tree known as the 'Napoleon Oak' by the academic community of the University of Lausanne. 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. It was standing next to the road, or may have been transplanted in honour of the future emperor's passage, whence its campus nickname. At the time of sample harvest for our study, the dividing apical meristems of this magnificent tree (Figure 1, Figure S1) had been exposed for 234 years to potential environmental mutagens, such as UV 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 [11]. The oak genome encodes 49,444 predicted protein-coding loci (Table 1).

Identification of Somatic Mutations Between Genomes from Two Branches

We used two approaches to identify SNVs (single-nucleotide variants) between the sequenced genomes of the two terminal branches of our initial sample, for regions of both high and low coverage. In total, reads from ca. 650 Mb of non-Ns containing sequence could be assessed. First, we aligned Illumina paired-reads on the repeat-masked genome in combination with the GATK [12] variant caller. This allowed us to establish a list of 314,865 potential SNV candidates. On the basis of a confidence score ≥300 on the heterozygous sites and ≥200 on homozygous sites, we further selected 1,536 putative SNVs for experimental validation by both PCR-seq [13] and Sanger sequencing. Of the 1,536 candidates, only seven could be confirmed (see Experimental Procedures).

Second, we used fetchGWI [14] to map read pairs to the entire non-masked genome. We were able to call 5,330 potential SNVs from the mapped reads using a simple read pileup process followed by detection of positions where the pileup differed from the reference genome. We systematically browsed candidate positions 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 and potentially poorly assembled repeated regions. This approach allowed us to select 82 putatively variable positions, including the seven already identified using the repeat-masked genome analysis described above. Ten of the remaining 75 candidates identified using the second approach were confirmed by PCR-seq and Sanger sequencing, increasing the total number of confirmed SNVs separating the two genomes to 17 (Figure 1, Table S1). Based on a conservative estimate, we are likely to have missed no more than 17 further such sites (see Experimental Procedures). All confirmed SNVs were heterozygous, as expected for novel somatic mutations. Intriguingly, two SNVs were found on the same contig, separated by only 12 bp (Figure 1, Table S1).

125 **Nested Distribution of SNVs Throughout the Napoleon Oak**

126 Having confidently established 17 SNVs, we then assessed their occurrence throughout the
127 tree. We used Sanger sequencing to genotype the remaining 24 terminal branches sampled
128 from other parts of the tree and checked for the presence of each SNV. As might be expected,
129 the SNVs were found in different sectors of the tree in a nested hierarchy that clearly
130 indicates the accumulation of mutations along branches during development (Figure 1, Figure
131 S2). Specifically, SNV1 and SNV2 were located in a large sector of connected branches that
132 are distributed in the upper half of the tree (Figure 1). SNV3 was restricted to a portion of the
133 same sector, whereas SNV4 to SNV10 were found at the top of one of the two branches for
134 which the full genome had been sequenced. Similarly, SNV11 to SNV14 were restricted to
135 the terminal fork of a principle lower branch from which the other genome had been fully
136 sequenced, whereas SNV15 to SNV17 were present in only one sector arising from this fork
137 (Figure 1, Figure S2). These results both provide independent confirmation of the originally
138 identified SNVs, and demonstrate their gradual, nested appearance and fixation in
139 developmentally connected branches during growth of the oak tree. Thus, while the exact
140 ontogeny of the Napoleon Oak may be difficult to reconstruct, our SNV analysis generated a
141 nested set of lineages supported by derived mutations, analogous to a phylogenetic tree.

142

143 **Somatic Mutation Rate is Low in the Napoleon Oak**

144 The spontaneous mutation rate in plants has been estimated to range from 5×10^{-9} to 30×10^{-9}
145 substitutions/site/generation, based on mutations accumulated during divergence between
146 monocots and dicots [15] or divergence between independent lines of *A. thaliana* maintained
147 in the laboratory [16, 17]. *A. thaliana* is an annual plant that reaches approximately 30 cm in
148 height before producing seeds. In contrast, the physical distance traced along branches
149 between the terminal branches we sequenced for the Napoleon Oak is about 40 m (Figure 1).

Assuming similar cell sizes between oak and *A. thaliana*, there should have been about 133 (4,000 cm/ 30 cm) times more mitotic divisions separating the extremes of somatic lineages in oaks than in *A. thaliana*. Under the assumption that the per-generation mutation rate is correlated with the number of mitotic divisions from zygote to gametes of the next generation [3, 8], the mitotic mutation rate for the Napoleon Oak should be between 6.6×10^{-7} and 4.0×10^{-6} substitution/site/generation. With such a somatic mutation rate, there should be between 433 and 2,600 SNVs across the 650 Mb sequenced between branch ends of the tree, or conservatively between 305 and 1,832 SNVs across the 458,143,725 nucleotides with a read coverage ≥8 in both samples. These values are much higher that the 17 SNVs we actually found. Even under the conservative assumption that we missed 17 other SNVs from the list of potential variants and between 12.9 and 79 false negatives (see Experimental Procedures), a higher-end estimate of the total number of oak SNVs ranges from 46.9 to 113. Both values are still substantially (and significantly) lower than the expected values (305 to 1,832) (Fisher's exact test, both $P < 2.2E\text{-}16$). Our finding seems to imply either much lower generational or mitotic mutation rates than generally assumed, or a different model for the accumulation of somatic mutations in trees, such as that proposed by [10]. Although the oak lineages sampled have not been separated by any meiosis events, which in yeast was found to elevate the generational mutation rate [19], they have been exposed to the natural environment, which in *A. thaliana* is known to significantly enhance mutation rate when compared to a controlled lab environment [20].

Most of SNVs identified in the Napoleon Oak were found only in single branches. Early sequestration of axillary meristems has been found in *A. thaliana* and tomato, two herbaceous plants that differ markedly in the developmental fate of their apical meristems [10]. Whereas *A. thaliana* forms a rosette of leaves and extends an inflorescence shoot at the time of flowering, the tomato apical meristems are separated by elongated internodes. Oak

trees have apical meristems of similar diameters to those of tomato [10, 21, and Figure 2] and show similar ontogeny. It thus seems reasonable to suppose that the growth pattern described in *A. thaliana* and tomato is quite general in flowering plants and might also apply to long-lived trees. If so, it would be consistent with the low number of SNVs we have identified in the oak.

We found that G:C→A:T transitions were the most frequent class of SNVs observed in the oak (Figure 3). Ultraviolet (UV) light causes G:C→A:T transitions at dipyrimidine sites in plants [22]. Among the 11 G:C→A:T transitions that we observed, seven were in a dipyrimidine context (Table S1). 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 (Table S1). 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.

Taken together, our results suggest that mutations due to replication errors in trees may be less important than environmentally induced mutations, as hypothesized by Burian et al. [10]. Mutations accumulate with age, irrespective of plant stature, and long-term exposure to UV radiation contributes to such changes. However, oaks protect their meristems in buds under multi-layered leaf-like structures (Figure 2), potentially reducing the incidence of UV mutagenesis. If so, the surprisingly low somatic mutation rate implied by our data would be consistent both with the pattern of cell division hypothesised by Burian et al. [10], as well as with the protective nature of oak bud morphology. In this context, it is noteworthy that there was no evidence for an expansion of DNA-repair genes in the oak genome compared to *A. thaliana* (Table S2).

Sixteen of the 17 SNVs identified in the Napoleon Oak occurred in introns or non-coding sequences that are probably neutral. The remaining one (SNV1), which occurred in a

large sector of the tree, generates an arginine-to-glycine conversion in a putative E3-ubiquitin ligase (Table S1). The functional impact of exchanging a positively charged arginine with a non-charged and smaller glycine residue is unknown and deserves further analysis.

To our knowledge, only two examples of functional mosaicism have been reported in trees, a low incidence that might be attributable to the low somatic mutation rate implied by our analysis. 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 [23, 24]. 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.

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 [25], often involving a loss of self-incompatibility systems [26], long-lived woody species are more likely to be fully outcrossing [27], including oaks [28]. 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 [3]. Inbreeding depression is indeed higher in long-lived woody species than annuals [29], 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 [5]. 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. [10], 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.

Our data give an unprecedented view on the limited role played by somatic mutations in long-lived plants, and support the view that stem cells in trees, although vulnerable to environment-induced mutations, are probably quite well protected from them. Consistent with this finding, a recent study in *A. thaliana* has shown that the number of cell divisions from germination to gametogenesis is independent of life span and vegetative growth [30]. Our study 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 somatic mutations in both plant and animal tissues [31-34].

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and seven tables.

AUTHOR CONTRUBUTIONS

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.

ACKNOWLEDGMENTS

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, Woodtli+Leuba SA for sample collection, and Jean-Jacques Strahm and Marco Bonetti for providing oak images. All Illumina reads and SMRT sequences have been deposited in GenBank under accession BioProject PRJNA327502.

REFERENCES

1. Klekowski, E.J., and Godfrey, P.J. (1989). Ageing and mutation in plants. Nature *340*, 389-391.
2. Schultz, S.T., Lynch, M., and Willis JH. (1999). Spontaneous deleterious mutation in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA *96*, 11393–11398.
3. Scofield, D.G., and Schultz, S.T. (2006). Mitosis, stature and evolution of plant mating systems: low-Φ and high-Φ plants. Proc. Royal Soc. London B: Biol. Sci. 273:275–282.
4. Ally., D., Ritland, K., and Otto, S.P. (2010). Aging in a long-lived clonal tree. PLoS Biol. *8*, e1000454.
5. Bobiwash, K., Schultz, S.T., and Schoen, D.J. (2013). Somatic deleterious mutation rate in a woody plant: estimation from phenotypic data. Heredity *111*, 338–344.
6. Halle, F. (1986). Modular growth in seed plants. Phil. Trans. Royal Soc. London B: Biol. Sci. *313*, 77–87.
7. Millet, J. (2012). L'architecture des arbres des régions tempérées: son histoire, ses concepts, ses usages. (MultiMondes).
8. Romberger, J.A., Hejnowicz, Z., and Hill, J.F. (1993). Plant structure: function and development. (Springer-Verlag).

9. Kwiatkowska, D. (2008). Flowering and apical meristem growth dynamics. J. Exp. Bot. *59*, 187-201.
10. Burian, A., Barbier de Reuille, P., and Kuhlemeier, C. (2016). Patterns of stem cell divisions contribute to plant longevity. Curr. Biol. *26*, 1385-1394.
11. Plomion, C., Aury, J.M., Amselem, J., Alaeitabar, T., Barbe, V., Belser, C., Bergès, H., Bodénès, C., Boudet, N., Boury, C., et al. (2016). Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. Mol. Ecol. Resour. *16*, 254-265.
12. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297-303.
13. Howald, C., Tanzer, A., Chrast, J., Kokocinski, F., Derrien, T., Walters, N., Gonzalez, J.M., Frankish, A., Aken, B.L., Hourlier, T., et al. (2012). Combining RT-PCR-seq to catalog all genic elements encoded in the human genome. Genome Res. *22*, 1698-1710.
14. Iseli, C., Ambrosini, G., Bucher, P., and Jongeneel, C.V. (2007). Indexing strategies for rapid searches of short words in genome sequences. PLoS One *2*, e579.
15. Wolfe, K.H., Li, W.H., and Sharp, P.M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA *84*, 9054–9058.
16. Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science *327*, 92–94.
17. Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.Q., Hurst, L.D., and Tian, D. (2015). Parent–progeny sequencing indicates higher mutation rates in heterozygotes. Nature *523*, 463-467.
18. Scofield, D.G. (2006). Medial pith cells per meter in twigs as a proxy for mitotic growth rate (Φ/m) in the apical meristem. Am. J. Bot. *93*, 1740-1747.
19. Rattray, A., Santoyo, G., Shafer, B., and Strathern, J,N. (2015). Elevated mutation rate during meiosis in *Saccharomyces cerevisiae*. PLoS Genet. *11*, e1004910.
20. Rutter, M.T., Shaw, F.H., and Fenster, C.B. (2010). Spontaneous mutation parameters for *Arabidopsis thaliana* measured in the wild. Evolution *64*, 1825-1835.
21. Barton, M.K. (2010). Twenty years on: The inner workings of the shoot apical meristem, a developmental dynamo. Dev. Biol. *341*, 95-113.
22. Britt, A.B. (1996). DNA damage and repair in plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. *47*, 75–100.
23. Edwards, P.B., Wanjura, W.J., Brown, W.V., and Dearn, J.M. (1990). Mosaic resistance in plants. Nature *347*, 434.
24. Padovan, A., Lanfear, R., Keszei, A., Foley, W.J., and Külheim, C. (2013). Differences in gene expression within a striking phenotypic mosaic *Eucalyptus* tree that varies in susceptibility to herbivory. BMC Plant Biol. *13*, 29.
25. Stebbins, G.L. (1950). Variation and evolution in plants. (Columbia University Press).
26. Goldberg, E.E., Kohn, J.R., Lande, R., Robertson, K.A., Smith, S.A., and Igić, B. (2010). Species selection maintains self-incompatibility. Science *330*, 493-495.
27. Barrett, S.C.H., Harder, L.D., and Worley, A.C. (1996). The comparative biology of pollination and mating in flowering plants. Phil. Tran. Royal Soc. London B: Biol. Sci. *351*, 1271-1280.
28. Streiff, R., Ducousso, A., Lexer, C., Steinkellner, H., Gloessl, J. and Kremer, A. (2009). Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. Mol. Ecol. *8*,831-841.

29. Goodwillie, C., Kalisz, S., and Eckert, C.E. (2005). The evolutionary enigma of mixed mating systems in plants: occurrence, theoretical explanations, and empirical evidence. Annu. Rev. Ecol. Evo. Syst. *36*, 47-79.
30. Watson, J.M., Platzer, A., Kazda, A., Akimcheva, S., Valuchova, S., Nizhynska, V., Nordborg, M., and Riha, K. (2016). Germline replications and somatic mutation accumulation are independent of vegetative life span in *Arabidopsis*. Proc. Natl. Acad. Sci. USA *113*, 12226–12231.
31. James, R.L. (2013). Genome mosaicism - One human, multiple genomes. Science *341*, 358–359.
32. Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D.C., Tamuri, A.U., Martincorena, I., Petljak, M., Alexandrov, L.B., Gundem, G., et al. (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature *513*, 422–425.
33. Lodato, M.A., Woodworth, M.B., Lee, S., Evrony, G.D., Mehta, B.K., Karger, A., Lee, S., Chittenden, T.W., D'Gama, A.M., Cai, X., et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. Science *350*, 94–98.
34. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S,, Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., et al. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science *348*, 880–886.
35. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics *30*, 2114-2120.
36. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience *1*, 18.
37. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.
38. Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. Genome Biol. *13*, R56.
39. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods *9*, 811-814.
40. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403-410.
41. English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., et al. (2012). Mind the gap: Upgrading genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS One *7*, 11.
42. Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucl. Acids Res. *32*, W309-312.
43. Tran, V.D., De Coi, N., Feuermann, M., Schmid-Siegert, E., Băguţ, E.T., Mignon, B., Waridel, P., Peter, C., Pradervand, S., Pagni, M., et al. (2016). RNA sequencing-based genome reannotation of the dermatophyte *Arthroderma benhamiae* and characterization of its secretome and whole gene expression profile during infection. mSystems *1*:, e00036-16.
44. Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.
45. Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones S., Eddy, S.R., et al. (2009). Rfam: updates to the RNA families database. Nucl. Acids Res. *3*7, D136-140.

46. Sonnhammer, E.L.L., Eddy, S.R., and Durbin, R. (1997). Pfam: a comprehensive database of protein families based on seed alignments. Proteins *28*, 405-420.
47. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie2. Nature Methods *9*, 357-359.
48. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics *43*, 491-498.
49. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protocols Bioinfo. *43*, 11.10.1-11.10.33.
50. Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. *132*, 365–386.
51. Myers, E.W., and Miller, W. (1988). Optimal alignments in linear space. Comput. Appl. Biosci. *4*, 11-17.
52. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics *25*, 2078-9.
53. Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicated genes. Science *290*, 1151–1155.
54. Vanneste, K., Van de Peer, Y., and Maere, S. (2013). Inference of genome duplications from age distributions revisited. Mol. Biol. Evol. *30*, 177-190.
55. Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. Plant Physiol. *171*, 2294-2316.
56. Altenhoff, A.M., Škunca, N., Glover ,N., Train, C.M., Sueki, A., Piližota, I., Gori, K., Tomiczek, B., Müller, S., Redestig, H., et al. (2015). The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucl. Acids Res. *43*, D240-D249.

FIGURE AND TABLE 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 Fig. S1. (**B**) Location of all identified SNVs. Sectors of the tree containing each group of SNVs are represented by different colours.

Figure 2. Napoleon Oak Apical Meristem

(**A**) Cross-section of an apical meristem. Stem 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.

Figure 3. Spectrum of Somatic Mutations Between Two Napoleon Oak Genomes

The type of substitution for 17 confirmed oak SNVs is shown.

Table 1. ***Quercus robur*** **Genome Statistics**

EXPERIMENTAL PROCEDURES

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 were constructed and sequenced. Single-molecule real-time (SMRT) sequencing (Pacific Biosciences) was performed on 25 SMRT cells according to the manufacturer's instructions (University of Lausanne Genomics Technologies Platform).

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) [35] and assembled using SOAPdenovo2 (v2.04.240, kmer 49) [36]. 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 (25 smartcells, XL-C2 and P4-C2 chemistry) 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 [37] (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) [38].

We screened the paired-end libraries for potential non-oak sequences using metaphlan (v1.7.7) [39]. 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) [40]. The genome was next scaffolded again using the PacBio reads and PBJelly (v14.1.14) [41]. 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) [42], as described in Tran et al. [43]. 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) [44]. 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) [45] 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 *A. thaliana*. PFAM [46] was used additionally to obtain functional domain annotations for the proteome and the concatenated proteome annotation was transferred onto the oak genome.

SNV Identification

First, short Illumina paired-end sequencing reads (298,735,463 and 316,299,457 for sample 0 and 66, respectively) were aligned to the masked *de novo* assembly (RepeatMasker, v4.05) with Bowtie2 (v2.2.2, https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.2.2) [47] using default parameters. GATK [12] was used for local realignment and variant calling using standard hard filtering parameters according to GATK Best Practices recommendations [48, 49]. Prior to variant calling, each sample was screened for duplicates using PICARD tools (http://broadinstitute.github.io/picard/). 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, and an exponential distribution, possibly representing true positives, and the decreasing number of false positives with regard to increasing confidence score, respectively. 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. We selected 1,536 putative SNVs out of 314,865 for validation, on the basis of a confidence score ≥300 on the heterozygous sites and ≥200 on homozygous sites. To validate putative differences between both samples, we used PCR-seq, a modification of the published RT-PCR-seq method [13]. Briefly, pairs of primers for 50-150 bp amplicons containing the targeted sequence were designed using Primer3 [50]. 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

519 manufacturer's instructions. The KAPA LTP Library Preparation Kit (Kapa Biosystems) was
520 used, starting with 500 ng of purified PCR products, to create a library compatible with an
521 Illumina sequencing platform. Clean-ups between enzymatic steps were performed with
522 Nucleospin PCR Clean-up columns (Macherey-Nagel). After ligation of pentabase adapters,
523 libraries were run on a 2 % agarose gel and extracted using the MinElute Gel Extraction kit
524 (Qiagen). Libraries were sequenced on HiSeq 2000 after six cycles of amplification
525 (Lausanne Genomic Technologies Facility). Amplicon reads were aligned, with no
526 mismatches allowed, to a compendium of the expected amplimers that bore the reference
527 allele, the alternate allele identified in the heterozygote sample, as well as the remaining two
528 nucleotides at the variable position; this allowed an unbiased estimation of the error rate
529 generated by the sequencing itself. As this method might have missed *bona fide* changes
530 between the two sampled branches that present other heterozygous sites close by, we also
531 aligned amplicon sequencing reads directly to the reference genome, with mismatches
532 allowed. Only seven of the 1,536 candidates assessed were validated by PCR-seq. Sanger
533 sequencing further confirmed them.

534 Second, to identify potential differences between both samples including masked
535 sequences, Illumina reads of samples 0 and 66 were mapped against the non-masked oak
536 genome assembly. The genome was 719,779,348 bp long, but 69,130,634 (9.52%) of those
537 nucleotides were gaps and were discarded, leaving an actual search space of 650,648,714 bp.
538 Of the latter, 458,143,725 nucleotides with a read coverage ≥8 in both samples were analysed
539 further. The mapping process was performed at the read pair level by the genome mapping
540 tool, fetchGWI [14], followed by a detailed sequence alignment tool, align0 [51]. Potential
541 SNVs were called from the mapped reads by a simple read pileup process followed by
542 detection of positions where the pileup shows variations with respect to the reference
543 genome; this produced a list of 5,330 positions. Those positions were browsed through a

local adaptation of the samtools pileup browser [52] 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.

Estimation of the Possible Missed SNVs.

About half of the putative variable sites with confidence scores ≥200 were assessed experimentally (1,536 out of 3,488 sites). Given the confidence scores of the tested sites, we estimated that we missed fewer than six such sites not evaluated by PCR-seq. To evaluate the number of true positives missed within candidates with confidence scores <200, we fitted a mixture of two distributions modelled on the 1,832,554 sites that were predicted to be heterozygous on both sequenced branches (Figure S3). This suggested that we had missed fewer than eleven true SNVs. We thus estimate a total of 17 missed SNVs. 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 Possible False Negatives.

We postulated that the false negative rate corresponds to the number of true SNVs that could have been missed in the sites that were called homozygous in both samples due to random sampling of both alleles. Assuming similar cell sizes between oak and *Arabidopsis*, there should have been ca. 133 (4,000 cm/30 cm) more mitotic divisions separating the extremes of somatic lineages in the analyzed oak, which were distant by 40 m, than in *Arabidopsis*, which reaches approximately 30 cm in height. Under the assumption that the per-generation mutation rate is correlated with the number of mitotic divisions from zygote to gametes of the next generation, the mitotic mutation rate for the oak should be between 6.6×10^{-7} and 4.0×10^{-6} substitution/site/generation based on the estimated mutation rates in plants (5×10^{-9} to 30×10^{-9} substitutions/site/generation [15-17]). Since the fraction of the genome that we analyzed consisted of ca. 458 Mb with ≥8 reads in each sample, the assumption is that there should be between 305 and 1,832 SNVs. We thus calculated the probability of having missed between 305 and 1,832 true SNVs in the ca. 458 Mb (read coverage ≥8) of sites considered homozygous in both samples. Our selection criterion for calling a heterozygous site was that it should contain ≥30% of the minor allele. Given the distribution of both alleles in an heterozygous DNA sample, the probability of obtaining less than 30% of a given allele after sequencing follows a binomial distribution, decreasing as the percentage of the minor allele decreases. Thus, the added probability of obtaining less than 30% of the minor allele decreases with the number of reads. We therefore analyzed the read coverage distribution for both oak samples and sorted the data into bins. Sample 0 shows an even distribution of read coverage between 8 and 70 whereas the coverage of sample 66 ranges from 8 to 50 (Figure S4). For each coverage bin we computed the mean added probability of obtaining less than 30% of the minor allele, and we multiplied this value by the fraction of the genome

represented in this particular bin. We next used the higher and lower values of expected SNVs (305 or 1,832) to calculate the number of SNVs possibly missed in each bin in each sample for low and high mutation rates because of random sampling of both alleles (i.e., for each bin we multiplied the added probability by either 305 or 1,832). The higher-end estimate of false-negatives for both samples was 12.9 (4.0 + 8.9) for a low mutation rate, and 79.0 (24.1 + 54.9) for a high mutation rate (Table S3).

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 (Table S4). Amplicons were then subjected to Sanger sequencing.

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) [53, 54]. 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, Figure S5). 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. [55],

(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. [55]). 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. [55], 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 (Figure S5). 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 *A. thaliana*, *Prunus persica* (peach) and *Q. robur* were called using the OMA database [56]. One-to-many orthologs, e.g., between *A. thaliana* and *Q. robur*, represent duplication in the oak lineage since the divergence from *A. thaliana*; 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 *A. thaliana* relative to several copies in both the peach and oak genomes), or whether it was peach- or oak-specific (i.e., one copy in *A. thaliana* 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

644 number of orthologs between each tree genome and *A. thaliana*, Table S5). We then
645 manually compiled a list of *A. thaliana* genes involved in DNA repair from
646 SwissProt/UniProtKB annotations (Table S6). We then counted specifically the number of
647 duplicates for genes involved in DNA repair and reported this as the number of orthologs
648 associated with this function (Table S2 and S7).
649

650

651 **Table 1. *Quercus robur* Genome Statistics**

652

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

653

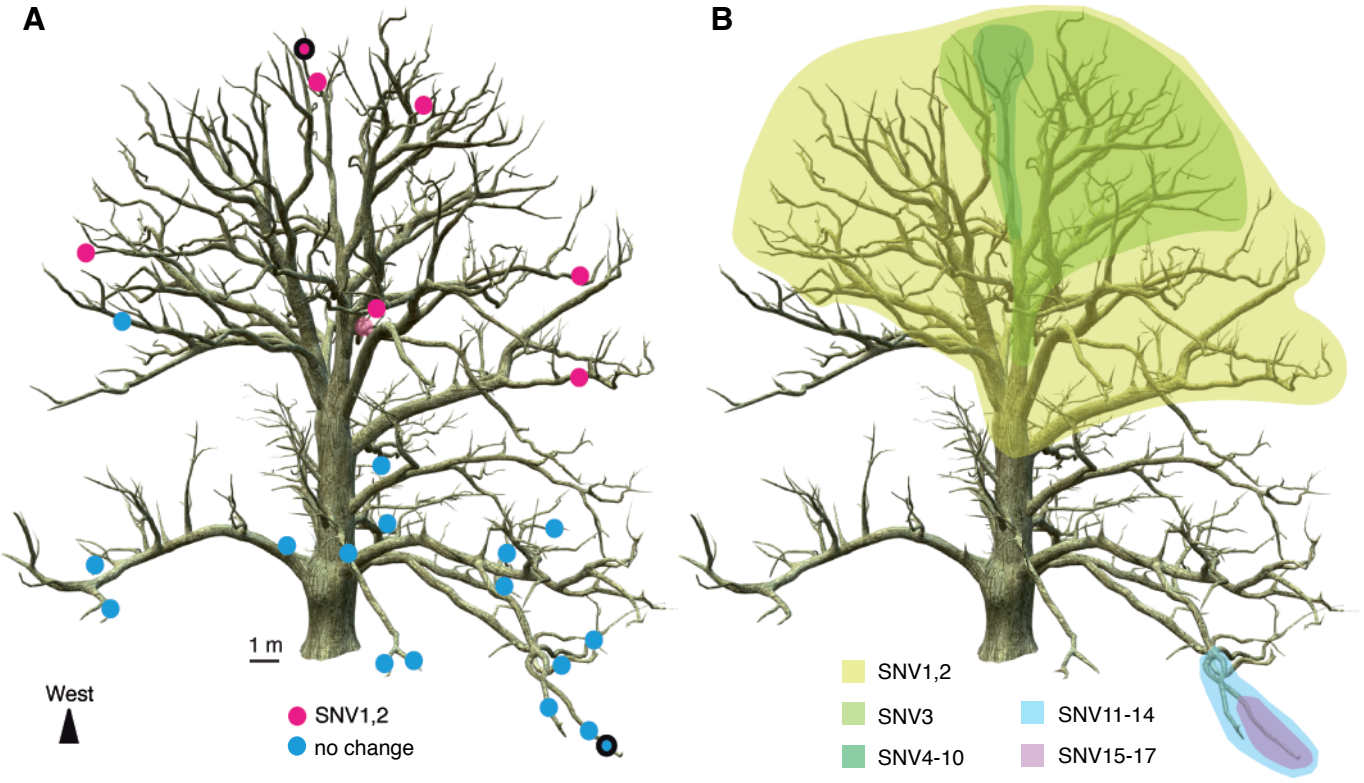

Figure 1. Distribution of Somatic Mutations in the Napoleon Oak

(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 Fig. S1. (**B**) Location of all identified SNVs. Sectors of the tree containing each group of SNVs are represented by different colours.

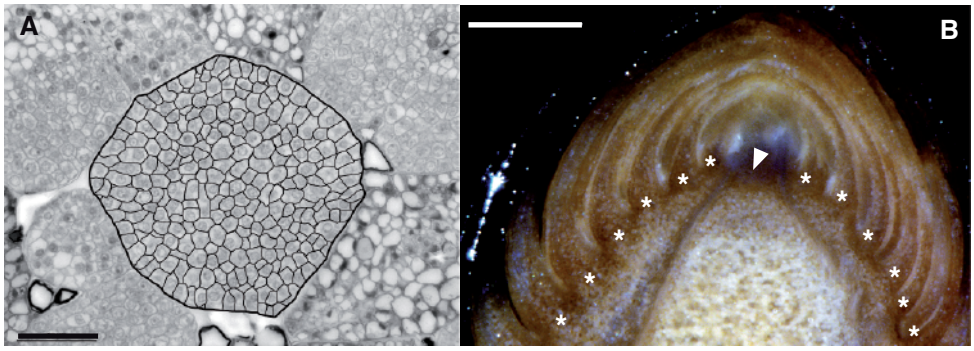

Figure 2. Napoleon Oak Apical Meristem

(**A**) Cross-section of an apical meristem. Stem 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.

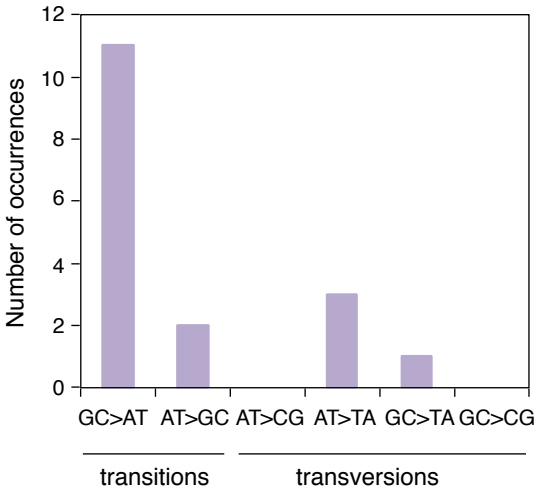

Figure 3. Spectrum of Somatic Mutations Between Two Napoleon Oak Genomes

The type of substitution for 17 confirmed oak SNVs is shown.