Title: The molecular clock in long-lived tropical trees is independent of growth rate

Authors: Akiko Satake[1†*], Ryosuke Imai[1†], Takeshi Fujino[2], Sou Tomimoto[1], Kayoko Ohta[1], Mohammad Na'iem[3], Sapto Indrioko[3], Widiyatno[3], Susilo Purnomo[4], Almudena Mollá–Morales[5], Viktoria Nizhynska[5], Naoki Tani[6,7], Yoshihisa Suyama[8], Eriko Sasaki[1], Masahiro Kasahara[2],

Affiliations:

[1] Department of Biology, Faculty of Science, Kyushu University, Fukuoka 819-0395, Japan.

[2] Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 277-8561, Chiba, Japan.

[3] Faculty of Forestry, Universitas Gadjah Mada, Jl. Agro No. 1 Bulaksumur Yogyakarta 55281, Indonesia.

[4] PT. Sari Bumi Kusuma, Sungai Raya, Pontianak Kota, Pontianak, West Kalimantan 78391, Indonesia.

[5] Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences, Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, 1030, Vienna, Austria.

[6] Forestry Division, Japan International Research Center for Agricultural Sciences, Tsukuba, Ibaraki, 305-8686, Japan.

[7] Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki, 305-8686, Japan.

[8] Field Science Center, Graduate School of Agricultural Science, Tohoku University, 232-3 Yomogida, Naruko-onsen, Osaki, Miyagi, 989-6711, Japan.

[†] These two authors contributed equally to this work

*Correspondence to:

Akiko Satake (akiko.satake@kyudai.jp)

Abstract

The rates of appearance of new mutations play a central role in evolution. However, mutational processes in natural environments and their relationship with growth rates are largely unknown, particular in tropical ecosystems with high biodiversity. Here, we examined the somatic mutation landscapes of two tropical trees, *Shorea laevis* (slow-growing) and *S. leprosula* (fast-growing), in central Borneo, Indonesia. Using newly-constructed genomes, we identified a greater number of somatic mutations in tropical trees than in temperate trees. In both species, we observed a linear increase in the number of somatic mutations with physical distance between branches. However, we found that the

rate of somatic mutation accumulation per meter of growth was 3.7-fold higher in *S. laevis* than in *S. leprosula*. This difference in the somatic mutation rate was scaled with the slower growth rate of *S. laevis* compared to *S. leprosula,* resulting in a constant somatic mutation rate per year between the two species. We also found that somatic mutations are neutral within an individual, but those mutations transmitted to the next generation are subject to purifying selection. These findings suggest that somatic mutations accumulate with absolute time and older trees have a greater contribution towards generating genetic variation.

Significance Statement

The significance of our study lies in the discovery of an absolute time-dependent accumulation of somatic mutations in long-lived tropical trees, independent of growth rate. Through a comparative analysis of somatic mutation landscapes in slow- and fast-growing species, we observed a clock-like accumulation of somatic mutations in both species, regardless of their growth rates. Although the majority of somatic mutations were restricted to a single branch, we also identified mutations present in multiple branches, likely transmitted during growth. Our findings suggest that older trees make a greater contribution towards generating genetic variation.

Keywords

Somatic mutations; mutational spectrum; genetic diversity; tropical ecosystems; Shorea

Main text

Biodiversity ultimately results from mutations that provide genetic variation for organisms to adapt to their environments. However, how and when mutations occur in natural environments is poorly understood[1–3]. Recent genomic data from long-lived multicellular species have begun to uncover the somatic genetic variation and the rate of naturally occurring mutations[4,5]. The rate of somatic mutations per year in a 234-year-old oak tree has been found to be surprisingly low[6] compared to the rate in an annual herb[7]. Similar analyses in other long-lived trees have also shown low mutation rates in both broadleaf trees[8–12] and conifers[13]. Despite the growing body of knowledge of somatic mutation landscapes in temperate regions, there is currently no knowledge on the somatic mutation landscapes in organisms living in tropical ecosystems, which are among the most diverse biomes on Earth.

Mutations can arise from errors during replication[14], or from DNA damage caused by exogenous mutagens or endogenous reactions at any time during cell growth[15]. While DNA replication errors have long been assumed to be major sources of mutations[16,17], a modeling study that relates the mutation rate to rates of DNA damage, repair and cell division[15] and experimental studies in yeast[18], human[19], and other animals[20] have shown the importance of

mutagenic processes that do not depend on cell division. Consequently, it remains largely unknown which source of mutations, whether replicative or non-replicative, predominates in naturally growing organisms.

To investigate the rates and patterns of somatic mutation and their relation to growth rates in tropical organisms, we studied the somatic mutation landscapes of slow- and fast-growing tropical trees in a humid tropical rain forest of Southeast Asia. By comparing the somatic mutation landscape between slow- and fast-growing species in a tropical ecosystem, we can gain insights into the mutagenesis that occurs in a natural setting. This comparison provides a unique opportunity to understand the impact of growth rate on somatic mutations and its potential role in driving evolutionary processes.

Detecting somatic mutations in slow- and fast-growing tropical trees

The humid tropical rainforests of Southeast Asia are characterized by a preponderance of trees of the Dipterocarpaceae family[21]. Dipterocarp trees are highly valued for both their contribution to forest diversity and their use in timber production. For the purposes of this study, we selected *Shorea laevis* and *S. leprosula*, both native hardwood species of the Dipterocarpaceae family (Supplementary Fig. 1a). *S. laevis* is a slow-growing species[22], with a mean annual increment (MAI) of diameter at breast height (DBH) of 0.38 cm/year (as measured over a 20 year period in $n = 2$ individuals; Supplementary Data 1). In contrast, *S. leprosula* exhibits a faster growth rate[22], with an MAI of 1.21 cm/year ($n = 18$; Supplementary Data 1), which is 3.2 times greater than that of *S. laevis*. We selected the two largest individuals of each species (S1 and S2 for *S.laevis* and F1 and F2 for *S. leprosula*; Fig. 1a) at the study site, located just below the equator in central Borneo, Indonesia (Supplementary Fig. 1b). We collected leaves from the apices of seven branches and a cambium from the base of the stem from each tree (Fig. 1a; Supplementary Fig. 2), resulting in a total of 32 samples. To determine the physical distance between the sampling positions, we measured the length of each branch (Supplementary Data 2) and DBH (Supplementary Table 1). The average heights of the slow- and fast-growing species were 44.1 m and 43.9 m, respectively (Fig. 1a; Supplementary Data 1). While it is challenging to accurately estimate the age of tropical trees due to the absence of annual rings, we used the DBH/MAI to approximate the average age of the slow-growing species to be 256 years and the fast-growing species to be 66 years (Supplementary Table 1).

To identify somatic mutations, we constructed new reference genomes of the slow- and fast-growing species. We generated sequence data using long-read PacBio RS II and short-read Illumina sequencing and assembled the genome using DNA extracted from the apical leaf at branch 1-1 of the tallest individual of each species (S1 and F1). The genomes were estimated to contain 52,935 and 40,665 protein-coding genes, covering 97.9% and 97.8% of complete BUSCO genes (eudicots_odb10) for the slow- and fast-growing species (Supplementary Table 2). Genome sizes estimated using k-mer distribution were 347 and 376 Mb for the slow- and fast-growing species, respectively.

To accurately identify somatic mutations, we extracted DNA from each sample twice to generate two biological replicates (Supplementary Fig. 2). A total of 64 DNA samples were

sequenced, yielding an average coverage of 69.3 and 56.5× per sample for the slow- and fast-growing species, respectively (Supplementary Data 5). We identified Single Nucleotide Variants (SNVs) within the same individual by identifying those that were identical within two biological replicates of each sample (Supplementary Fig. 2). We identified 728 and 234 SNVs in S1 and S2, and 106 and 68 SNVs in F1 and F2, respectively (Supplementary Fig. 2; Supplementary Data 4). All somatic mutations were unique and did not overlap between individuals. Experimental validation of a subset of the inferred SNVs revealed that 93.9% were accurately annotated (Supplementary Data 5).

Somatic mutation rates per year is independent of growth rate

Phylogenetic trees constructed using somatic mutations were almost perfectly congruent with the physical tree structures (Fig. 1a). The majority of somatic mutations were present at a single branch, but we also identified somatic mutations present in multiple branches (Fig. 1b) which are likely transmitted to new branches during growth. We also observed somatic mutations that did not conform to the branching topology (Fig. 1b), as theoretically predicted due to the stochastic loss of somatic mutations during branching[23]. Unlike previous empirical studies[9,10], we did not incorporate knowledge of the branching topology of the tree in the SNV discovery process. The strong concordance between the phylogenetic trees and physical tree structure in the absence of prior knowledge of tree branching topology suggests the reliability of the SNV detection pipelines developed in this study.

Our analysis revealed that the number of SNVs increases linearly as the physical distance between branch tips increases (Fig. 2a). The somatic mutation rate per site per meter was determined by dividing the slope of the linear regression of the number of SNVs against the physical distance between branch tips by the number of callable sites from the diploid genome of each tree (Fig. 2b; Supplementary Table 3). The somatic mutation rate per nucleotide per meter was 7.08×10^{-9} (95% CI: $6.41–7.74 \times 10^{-9}$) and 4.27×10^{-9} (95% CI: $3.99–4.55 \times 10^{-9}$) for S1 and S2, and 1.77×10^{-9} (95% CI: $1.64–1.91 \times 10^{-9}$) and 1.29×10^{-9} (95% CI: $1.05–1.53 \times 10^{-9}$) for F1 and F2, respectively. The average rate of somatic mutation for the slow-growing species was 5.67×10^{-9} nucleotide^{-1} m^{-1}, which is 3.7-fold higher than the average rate of 1.53×10^{-9} nucleotide^{-1} m^{-1} observed in the fast-growing species (Fig. 2b; Supplementary Table 3). This result indicates that the slow-growing tree accumulates more somatic mutations compared to the fast-growing tree to grow the unit length. This cannot be explained by differences in the number of cell divisions, as the length and diameter of fiber cells in both species are not significantly different (1.29 mm and 19.0 μm for the slow-growing species[24] and 0.91mm and 22.7 μm for the fast-growing species[25]).

Based on the estimated age of each tree, somatic mutation rate per nucleotide per year was calculated for each tree. On average, resultant values were largely similar between the two species, with 7.71×10^{-10} and 8.05×10^{-10} nucleotide^{-1} year^{-1} for the slow- and fast-growing species, respectively (Fig. 2b; Supplementary Table 3). This result suggests that somatic mutation accumulates in a clock-like manner as they age regardless of tree growth. The result suggests that somatic mutation accumulates in a clock-like manner as they age regardless of tree

growth. Our estimates of somatic mutation rates per nucleotide per year in *Shorea* are higher than those previously reported in other long-lived trees such as *Quercus robur*[6], *Populus trichocarpa*[11], *Eucalyptus melliodora*[10] and *Picea sitchensis*[13]. This suggests that long-lived trees in the tropics do not necessarily suppress somatic mutation rates to the same extent as their temperate counterparts.

Mutational spectra are similar between slow- and fast-growing trees

Somatic mutations may be caused by exogenous factors such as ultraviolet and ionizing radiation, or endogenous factors such as oxidative respiration and errors in DNA replication. To identify characteristic mutational signatures caused by different mutagenic factors, we characterized mutational spectra by calculating the relative frequency of mutations at the 96 triplets defined by the mutated base and its flanking 5' and 3' bases (Fig. 3; Supplementary Fig. 3). Across species, the mutational spectra showed a dominance of cytosine-to-thymine (C>T and G>A on the other strand, noted as C:G>T:A) substitutions at CpG sites with CG (Fig. 3a, b). This is believed to result from the spontaneous deamination of 5-methylcytosine[26,27]. Methylated CpG sites spontaneously deaminate, leading to TpG sites and increasing the number of C>T substitutions[28].

We compared the mutational spectra of our tropical trees to single-base substitution (SBS) signatures in human cancers using the Catalogue Of Somatic Mutations In Cancer (COSMIC) compendium of mutation signatures (COSMICv.2[29–31]). The mutational spectra were largely similar to the dominant mutation signature in humans known as SBS1 (cosine similarity = 0.789 and 0.597 for the slow- and fast-growing species; Supplementary Data 6). SBS1 is believed to result from the spontaneous deamination of 5-methylcytosine. The mutational spectra were also comparable to another dominant signature in humans, SBS5 (cosine similarity = 0.577 and 0.558 for the slow- and fast-growing species; Supplementary Data 6), the origin of which remains unknown. Our finding that somatic mutations in tropical trees accumulate in a clock-like manner (Fig. 2a) is consistent with the clock-like mutational process observed in SBS1 and SBS5 in human somatic cells[32,33]. This suggests that the mutational processes in plants and animals are conserved, despite the variation in their life forms and environmental conditions.

Somatic mutations are neutral but inter-individual SNVs are subject to selection

We tested whether the somatic mutations and inter-individual SNVs are subject to selection (Fig. 4a). The observed rate of non-synonymous somatic mutations did not deviate significantly from the expected rate under the null hypothesis of neutral selection in both the slow- (binomial test: $P = 0.71$) and fast-growing (binomial test: $P = 1.0$) species (Fig. 4b; Supplementary Table 4). In contrast, the number of inter-individual SNVs were significantly smaller than expected ($P < 10^{-15}$ for both species: Fig. 4c). These results indicate that somatic mutations are largely neutral within an individual, but mutations passed to next generation are subject to strong purifying selection during the process of embryogenesis, seed germination and growth.

Overall, the mutational spectra were similar between somatic and inter-individual SNVs (Supplementary Fig. 3). However, the fraction of C>T substitutions, in particular at CpG sites,

was lower in inter-individual SNVs compared to somatic SNVs (Fig. 4d). This observation may be indicative of the potential influence of GC-biased gene conversion during meiosis[34] or biased purifying selection for C>T inter-individual nucleotide substitutions.

Discussion

Our study demonstrates that while the somatic mutation rate per meter is higher in the slow- than in fast-growing species, the somatic mutation rate per year is independent of growth rate. To gain deeper understanding of these findings, we developed a simple model that decomposes the mutation rate per site per cell division (μ) into the two components: DNA replication dependent (α) and replication independent (β) mutagenesis. This can be represented as $\mu = \alpha + \beta\tau$, where τ is the duration of cell cycle measured in years. The replication dependent mutation emanates from errors that occur during DNA replication, such as the misincorporation of a nucleotide during DNA synthesis. The replication independent mutation arises from DNA damage caused by endogenous reactions or exogenous mutagens at any time of cell cycle. Since the number of cell division per year is given as $r = 1/\tau$, the mutation rate per year becomes $r\mu = \alpha/\tau + \beta$. From the relationship, the number of nucleotide substitution per site accumulated over t years, denoted as $m(t)$, is given by $m(t) = (\alpha/\tau + \beta)t$. The formula indicates that when β is significantly greater than α, somatic mutations accumulate with tree age rather than with tree growth.

We estimated the relative magnitudes of α and β by using the results obtained from our study. Given that the cell cycle duration is likely inversely proportional to MAI, we have τ_S/τ_F = 3.2 (Supplementary Data 1), where τ_S and τ_F denote the cell cycle duration for the slow- and fast-growing species, respectively. It is also reasonable to assume that the same number of cell divisions are required to achieve 1 m of growth in both species as the cell size is similar between the two species. Based on our estimates of the somatic mutation rate per site per meter for the slow- (μ_S) and fast-growing species (μ_F), we have $\mu_S/\mu_F = (\alpha + \beta\tau_S)/(\alpha + \beta\tau_F)$ = 3.7, which is close to the ratio of cell cycle duration τ_S/τ_F. This consistency can be explained by the substantial contribution of the replication independent mutagenesis to the somatic mutation rate (i.e. $\beta \gg \alpha$), as long as the magnitudes for α and β are similar between the two species. The time required for a unit length to grow can vary even within the same species, depending on microenvironmental conditions such as the availability of light and nutrients. These variations could explain the differences in somatic mutation rates per unit growth between two individuals within the same species (Fig. 2).

This argument concords with previous studies in human and other animals, which showed the presence of mutations that do not track cell division[19,20]. This study contributes to understanding the importance of non-replicative mutagenesis in naturally grown trees by decoupling the impacts of growth and time on the rate of somatic mutation. The preponderance of non-replicative mutational process can be attributed to its distinct molecular origin, the accumulation of spontaneous CpG mutations with absolute time. The neutral nature of newly arising somatic mutations within the tree results in a molecular clock, a constant rate of molecular evolution[35–37]. For our argument, we made an intuitive assumption that the number of

stem cell divisions increases with distance regardless of species when cell size is similar. However, to further validate this assumption, we require mathematical models that consider the asymmetric division of stem cells[38,39] and complex stem cell population dynamics during elongation and branching in tree growth[23,40].

In trees, somatic mutations can be transmitted to seeds[8], resulting in a rich genetic variation that can facilitate the effectiveness of natural selection to subsequent generations. The humid tropical rainforests of Southeast Asia, which are renowned for their high species diversity, are currently facing threats from deforestation and climate change. Our study suggests that long-lived trees may contribute to the generation of genetic variation in tropical systems.

References

1. Whitham, T. G. & Slobodchikoff, C. N. Evolution by individuals, plant-herbivore interactions, and mosaics of genetic variability: The adaptive significance of somatic mutations in plants. *Oecologia* **49**, 287–292 (1981).
2. Gill, D. E., Chao, L., Perkins, S. L. & Wolf, J. B. Genetic mosaicism in plants and clonal animals. *Annu Rev Ecol Syst* **26**, 423–444 (1995).
3. Schoen, D. J. & Schultz, S. T. Somatic mutation and evolution in plants. *Annu. Rev. Ecol. Evol. Syst* **50**, 49–73 (2019).
4. Yu, L. *et al.* Somatic genetic drift and multilevel selection in a clonal seagrass. *Nature Ecology & Evolution 2020 4:7* **4**, 952–962 (2020).
5. Reusch, T. B. H., Baums, I. B. & Werner, B. Evolution via somatic genetic variation in modular species. *Trends Ecol Evol* **36**, 1083–1092 (2021).
6. Schmid-Siegert, E. *et al.* Low number of fixed somatic mutations in a long-lived oak tree. *Nat Plants* **3**, 926–929 (2017).
7. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
8. Plomion, C. *et al.* Oak genome reveals facets of long lifespan. *Nat Plants* **4**, 440–452 (2018).
9. Wang, L. *et al.* The architecture of intra-organism mutation rate variation in plants. *PLoS Biol* **17**, 1–29 (2019).
10. Orr, A. J. *et al.* A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. *Proceedings of the Royal Society B: Biological Sciences* **287**, (2020).
11. Hofmeister, B. T. *et al.* A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial *Populus trichocarpa*. *Genome Biol* **21**, 1–27 (2020).
12. Duan, Y. *et al.* Limited accumulation of high-frequency somatic mutations in a 1700-year-old Osmanthus fragrans tree. *Tree Physiol* **42**, 2040–2049 (2022).
13. Hanlon, V. C. T., Otto, S. P. & Aitken, S. N. Somatic mutations substantially increase the per-generation mutation rate in the conifer *Picea sitchensis*. *Evol Lett* **3**, 348–358 (2019).
14. Reijns, M. A. M. *et al.* Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518**, 502–506 (2015).
15. Gao, Z., Wyman, M. J., Sella, G. & Przeworski, M. Interpreting the dependence of mutation rates on age and time. *PLoS Biol* **14**, e1002355 (2016).
16. Makova, K. D. & Li, W.-H. *Strong male-driven evolution of DNA sequences in humans and apes. Nature* **416**, 624–626 (2002).
17. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).
18. Liu, H. & Zhang, J. Yeast Spontaneous mutation rate and spectrum vary with environment. *Current Biology* **29**, 1584-1591.e3 (2019).
19. Abascal, F. *et al.* Somatic mutation landscapes at single- molecule resolution.
20. De Manuel, M., Wu, F. L. & Przeworski, M. A paternal bias in germline mutation is widespread in amniotes and can arise independently of cell division numbers. *Elife* **11**, (2022).
21. J, G. *Dipterocarp Biology, Ecology, and Conservation.* (Oxford University Press, 2016).

22. Widiyatno, W., Soekotjo, S., Naiem, M., Purnomo, S. & Setiyanto, P. E. Early performance of 23 dipterocarp species planted in logged-over rainforest. *Journal of Tropical Forest Science* **26**, 259–266 (2014).
23. Tomimoto, S. & Satake, A. Modelling somatic mutation accumulation and expansion in a long-lived tree with hierarchical modular architecture. *bioRxiv* 2022.05.02.490262 (2023) doi:10.1101/2022.05.02.490262.
24. Usami, K. Tropical Woods as Pulp Stuffs. *Journal of Agricultural Research Quarterly* **12**, 109–114 (1978).
25. Praptoyo, H. & Mayaningsih, R. Anatomical features of wood from some fast growing red meranti. *Proceeding of The 4th International Symposium of IWoRs* **7**, 8 (2012).
26. Coulondre, C., Miller, J. H., Farabaugh, P. J., & Gilbert, W. Molecular basis of base substitution hotspots *Escherichia coli*. *Nature* **274**, 568–571 (1978).
27. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560–561 (1980).
28. Cooper, D. N. & Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum Genet* **83**, 181–188 (1989).
29. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246–259 (2013).
30. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
31. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
32. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. Europe PMC Funders Group. *Nat Genet* **47**, 1402–1407 (2015).
33. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
34. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**, 285–311 (2009).
35. Zuckerkandl, E. & Pauling, L. *Evolving Genes and Proteins*. (Academic Press, 1965).
36. Kimura, M. & Ohta, T. On the rate of molecular evolution. *J Mol Evol* **1**, 1–17 (1971).
37. Kimura, M. The Neutral theory of molecular evolution. (Cambridge University Press, 1983). doi:10.1017/CBO9780511623486.
38. Watson, J. M. *et al.* Germline replications and somatic mutation accumulation are independent of vegetative life span in *Arabidopsis*. *Proc Natl Acad Sci U S A* **113**, 12226–12231 (2016).
39. Lanfear, R. Do plants have a segregated germline? *PLoS Biol* **16**, e2005439 (2018).
40. Iwasa, Y., Tomimoto, S. & Satake, A. The genetic structure within a single tree is determined by the behavior of the stem cells in the meristem. *Genetics* (2023) doi:10.1093/GENETICS/IYAD020.

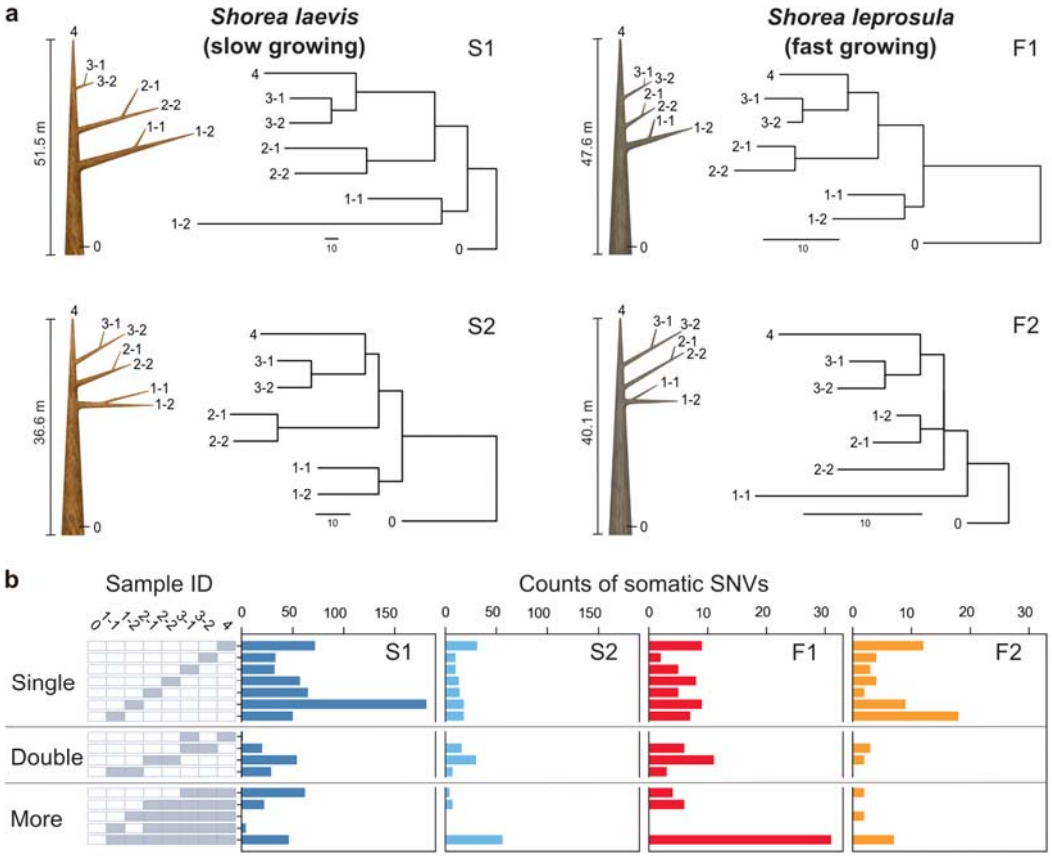


Fig. 1 | Physical tree structures and phylogenetic trees constructed from somatic mutations. **a,** Comparisons of physical tree structures (left, branch length in meters) and neighbor-joining (NJ) trees (right, branch length in the number of nucleotide substitutions) in two tropical tree species: *S. laevis*, a slow-growing species (S1 and S2), and *S. leprosula*, a fast-growing species (F1 and F2). IDs are assigned to each sample from which genome sequencing data were generated. Vertical lines represent tree heights. **b,** Distribution of somatic mutations within tree architecture. A white and gray panel indicates the presence (gray) and absence (white) of somatic mutation in each of eight samples compared to the genotype of sample 0. Sample IDs are the same between panels **a** and **b**. The distribution pattern of somatic mutations is categorized as Single, Double, and More depending on the number of samples possessing the focal somatic mutations. Among $2^7 - 1$ possible distribution patterns, the patterns observed in at least one of the four individuals are shown.

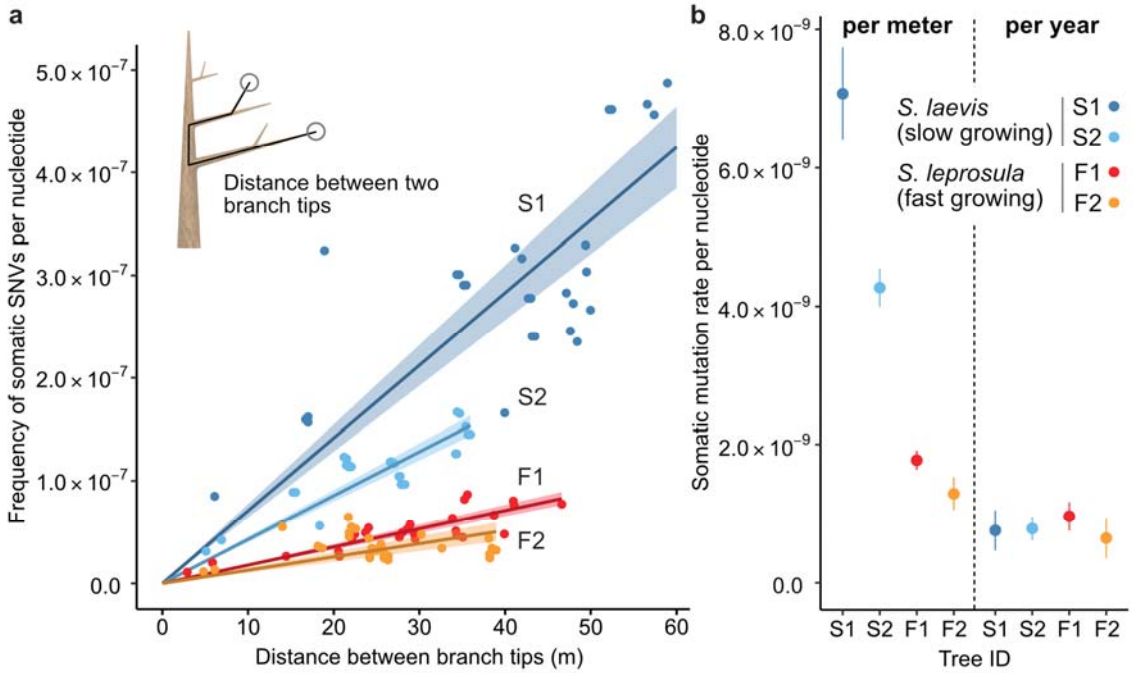


Fig. 2 | The relationship between the physical distance and the numbers of SNVs. a, Linear regression of the number of SNVs against the pair-wise distance between branch tipcs with an intercept of 0 for each tree (S1: blue, S2: right blue, F1: red, and F2: orange). Shaded areas represent 95% confidence intervals of regression lines. Regression coefficients are listed in Supplementary Table 3. **b**, Comparison of somatic mutation rates per nucleotide per growth and per year across four tropical trees. Bars indicate 95% confidence intervals.

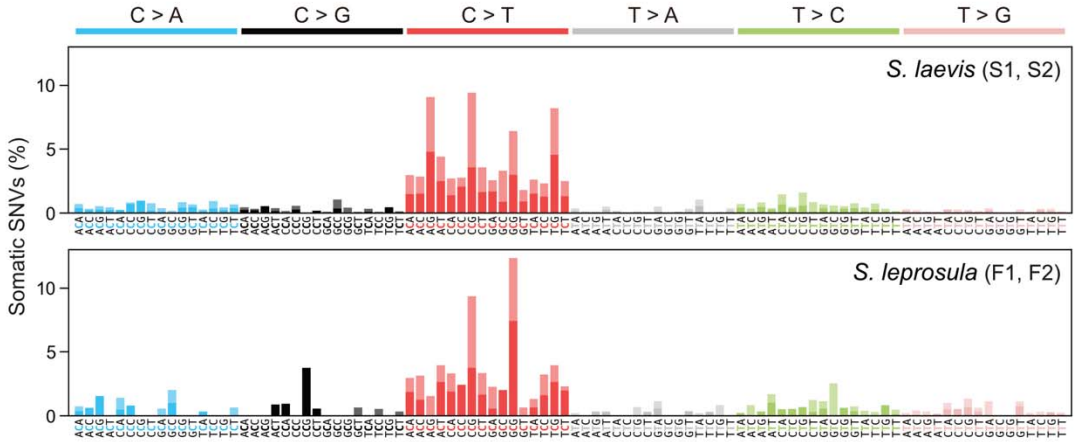


Fig. 3 | Mutational spectra of somatic SNVs. Somatic mutation spectra in *S. laevis* (upper panel) and *S. leprosula* (lower panel). The horizontal axis shows 96 mutation types on a trinucleotide context, coloured by base substitution type. Different colours within each bar indicate complementary bases. For each species, the data from two trees (S1 and S2 for *S. laevis* and F1 and F2 for *S. leprosula*) were pooled to calculate the fraction of each mutated triplet.

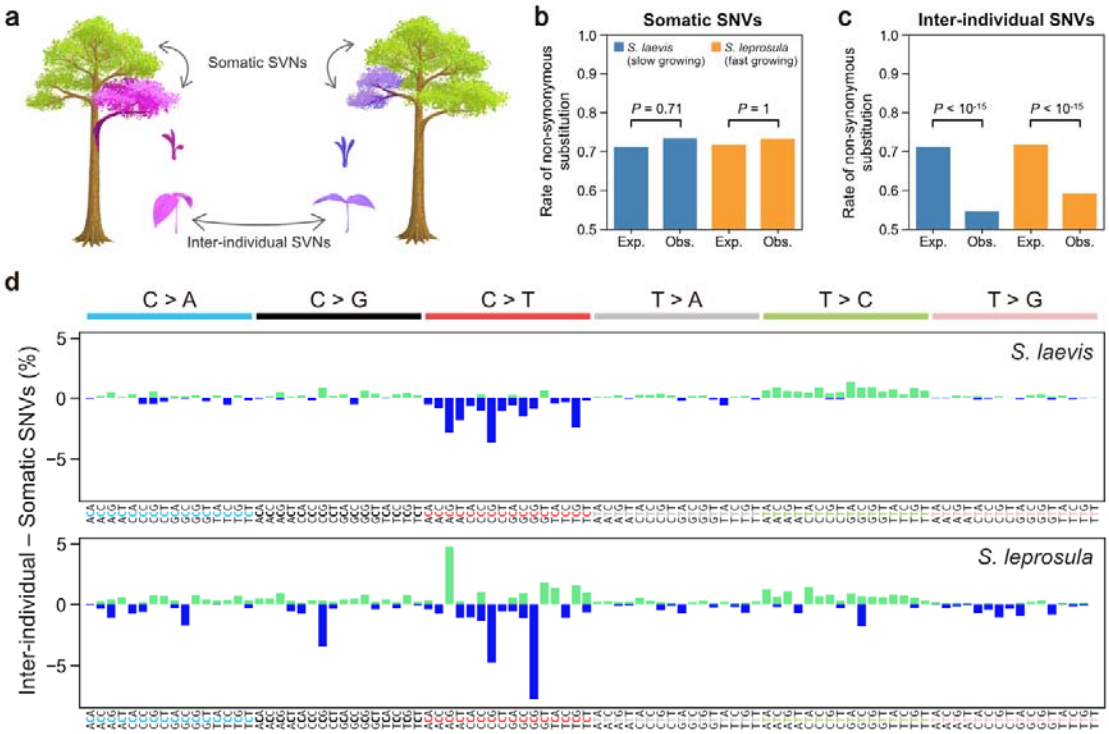


Fig. 4 | Detecting selection on somatic and inter-individual SNVs. a, An illustration of somatic and inter-individual SNVs. Different colours indicate different genotypes. **b**, Expected (Exp.) and observed (Obs.) rates of somatic non-synonymous substitutions. **c**, Expected (Exp.) and observed (Obs.) rates of inter-individual non-synonymous substitutions. **d**, The difference between the fractions of inter-individual and somatic substitutions spectra in *S. laevis* (upper panel) and *S. leprosula* (lower panel). The positive and negative values are plotted in different colours. The horizontal axis shows 96 mutation types on a trinucleotide context, coloured by base substitution type.

Materials and Methods

Study site and sampling methods

The study site is in a humid tropical rain forest in Central Borneo, Indonesia (00°49′ 45.7″ S, 112°00′ 09.5″ E; Supplementary Fig. 1b). The forest is characterized by a prevalence of trees of the Dipterocarpaceae family and is managed through a combination of selective logging and line planting (Tebang Pilih Tanam Jalur, TPTJ). The mean annual temperature range from 2001 to 2009 was between 22 to 28°C at night and 30 to 33°C during the day, with an average annual precipitation of 3376 mm[41].

The study focuses on two native Dipterocarpaceae species, *S. laevis* and *S. leprosula* (Supplementary Fig. 1a). We logged two individuals from each species (S1 and S2 for *S. laevis* and F1 and F2 for *S leprosula*; Supplementary Fig. 1a) on July 17–18, 2018 and collected samples prior to their transportation for timber production. Approximately 0.4–1.0 g of leaf tissue was collected from each of the apices of seven branches and approximately 5 g of cambium tissue was taken from the base of the stem per individual (Supplementary Fig. 2). To calculate the physical distance between sampling positions within the tree architecture, we measured the length of each branch (Supplementary Data 2). Samples were promptly preserved in a plastic bag with silica gel following harvest and transported to the laboratory within 4 days of sampling. During transportation, samples were kept in a cooler box with ice to maintain a low temperature. Once in the laboratory, samples were stored at −80°C until DNA and RNA extraction.

DBH have been recorded for the trees with DBH greater than 10 cm every two years since 1998 within three census plots of 1 hectare (100 × 100 m) in size located near the target trees. The mean growth was calculated by taking the average of MAI of DBH for 2 and 18 trees for the slow- and fast-growing species, respectively (Supplementary Data 1).

DNA extraction

For short-read sequencing, DNA extraction was performed using a modified version of the method described previously[42] as follows: Frozen leaves were ground in liquid nitrogen and washed up to five times with 1 mL buffer (including 100 mM HEPES pH 8.0, 1% PVP, 50 mM Ascorbic acid, 2% (v/v) β-mercaptoethanol)[43]. DNA was treated with Ribonuclease (Nippongene, Tokyo, Japan) according to the manufacture's instruction. DNA was extracted twice independently from each sample for two biological replicates. The DNA yield was measured on a NanoDrop ND-2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and Qubit4 Fluorometer (Thermo Fisher Scientific). For long-read sequencing, we extracted high molecular weight genomic DNA from branch 1-1 leaf materials of S1 and F1 individuals using a modified CTAB method[44].

RNA extraction and sequencing

For genome annotation, total RNA was extracted from the cambium sample of the S1 individual of *S. laevis* in accordance with the method described in a previous study[45]. RNA integrity was measured using the Agilent RNA 6000 Nano kit on a 2100 Bioanalyzer (Agilent Technologies,

Santa Clara, CA, USA), and the RNA yield was determined using a NanoDrop ND-2000 spectrophotometer (Thermo Fisher Scientific). The extracted RNA was sent to Pacific Alliance Lab (Singapore), where a cDNA library was prepared with a NEBNext® Ultra™ RNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, MA, USA) and 150 paired-end transcriptome sequencing was conducted using an Illumina NovaSeq6000 sequencer (Illumina, San Diego, CA, USA). For *S. leprosula*, we used published RNA-seq data[46].

Illumina short-read sequencing and library preparation

For Illumina short-read sequencing, the DNA sample from the first replicate of the S1 individual of *S. laevis* was sent to the Next Generation Sequencing Facility at Vienna BioCenter Core Facilities (VBCF), a member of the Vienna BioCenter (VBC) in Austria, for library preparation and sequencing on the Illumina HiSeq2500 platform (Illumina). The library was prepared using the on-bead tagmentation library prep method according to the manufacturer's protocol and was individually indexed with the Nextera index Kit (Illumina) by PCR. Insert size was adjusted to around 450 bp. The quantity and quality of each amplified library were analyzed using the Fragment Analyzer (Agilent Technologies) and the HS NGS Fragment Kit (Agilent Technologies).

The DNA sample from the second replicate of the S1 individual and two replicates from the S2, F1, and F2 individuals were sent to Macrogen Inc. (Republic of Korea) for sequencing on the Illumina HiseqX platform (Illumina). DNA was sheared to around 500 bp fragments in size using dsDNA fragmentase (New England BioLabs). Library preparation was performed using the NEBNext Ultra II DNA Library Prep Kit (New England BioLabs) according to the manufacturer's protocol, and the libraries were individually indexed with the NEBNext Multiplex Oligos for Illumina (New England BioLabs) by PCR. The quality and quantity of each amplified library were analyzed using the Bioanalyzer 2100 (Agilent Technologies), the High Sensitivity DNA kit (Agilent Technologies), and the NEBNext Library Quant Kit for Illumina (New England BioLabs). In total, 64 samples (16 samples per individual) were used for short-read sequencing.

PacBio long-read sequencing and library preparation

To construct the reference genome of *S. laevis* and *S. leprosula*, high molecular weight DNA samples were extracted from branch 1-1 leaf materials of S1 and F1 individuals of each species, and sequenced using PacBio platforms. For *S. laevis*, library preparation and sequencing were performed at VBCF. The library was prepared using the SMRTbell express Kit (PacBio, Menlo Park, CA, USA), and sequenced on the Sequel platform with six SMRT cells (PacBio). For *S. leprosula*, library preparation and sequencing were performed by Macrogen Inc. (Republic of Korea). The library for *S. leprosula* was prepared using the HiFi SMARTbell library preparation system (PacBio) according to the manufacturer's protocol, and was sequenced on the Sequel II platform (PacBio) with one SMRT cell.

Genome assembly

The PacBio continuous long reads of *S. laevis* were assembled using Flye 2.7-b1587[47] with 12 threads and with an estimated genome size of 350 Mbp. We subsequently used HyPo v1.0.3[48] for polishing the contigs. The Illumina read alignments provided to HyPo were created using Bowtie v2.3.4.3[49] with `--very-sensitive` option and using 32 threads. We used the Illumina reads from all branches of the individual S1 rather than utilizing exclusively those of branch 1-1, in order to capitalize on the increased aggregate sequencing depth.

The PacBio HiFi reads of *S. leprosula* with an average Quality Value (QV) 20 or higher were extracted, and subsequently assembled using Hifiasm 0.16.1-r375[50], with `-z10` option and using 40 threads. The primary assembly of *S. leprosula* was used for further analysis. The quality and completeness of the genome assembly were assessed by searching for a set of 2,326 core genes from eudicots_odb10 using BUSCO v5.3.0[51] for each species (Supplementary Table 2).

Genome annotation

We constructed repeat libraries of *S. laevis* and *S. leprosula* using EDTA v2.0.0[52]. Using the libraries, we ran RepeatMasker 4.1.2-p1[53] with `-s` option and with Cross_match as a search engine, to perform soft-masking of trepetitive sequences in the genomes. The estimated percentages of the repetitive sequences were 42.4% for *S. laevis* and 39.5% for *S. leprosula* (Supplementary Table 2).

We ran BRAKER 2.1.6[54] to perform gene prediction by first incorporating RNA-seq data and subsequently utilizing a protein database, resulting in the generation of two sets of gene predictions for each species. To perform RNA-seq-based prediction, we mapped the RNA-seq reads (see RNA extraction in Methods section) to the genomes using HISAT 2.2.1[55], with the alignments subsequently being employed as training data for BRAKER. For protein-based prediction, we used proteins from the Viridiplantae level of OrthoDB v10[56] as the training data.

The two sets of gene predictions were merged using TSEBRA (commit `0e6c9bf` in the GitHub repository)[57] to select reliable gene predictions for each species. Although in principle TSEBRA groups overlapping transcripts and considers them as alternative spliced isoforms of the same gene, we identified instances where one transcript in a gene overlapped with another transcript in a separate gene. In such cases, we manually clustered these transcripts into the same gene.

We used EnTAP 0.10.8[58] with default parameters for functional annotation. The databases employed were: UniProtKB release 2022_05[59], NCBI RefSeq plant proteins release 215[60], EnTAP Binary Database v0.10.8[58] and EggNOG 4.1[61]. We constructed the standard gene model by utilizing the gene predictions of each species, eliminating any gene structures that lacked a complete ORF. Transcripts containing Ns were also excluded. Following the filtering process, the splice variant displaying the longest coding sequence (CDS) was selected as the primary isoform for each gene. The set of primary isoforms was used as the standard gene model.

Genome size estimation

We estimated genome size of two species using GenomeScope[62]. We counted k-mer from forward sequence data of branch 1-1 from the S1 and F1 individuals using KMC 3[63] (k = 21). The genome size and heterozygous ratio were estimated by best model fitting. Estimated genome sizes were 347 Mb for the slow-growing species and 376 Mb for the fast-growing species. These estimates were 8% and 7% smaller than the estimates obtained through flow cytometry[64], respectively. The genome size of the fast-growing species was nearly identical to that previously reported for *S. leprosula* in peninsular Malaysia[46].

Somatic (intra-individual) SNV discovery

We filtered low quality reads out and trimmed adapters using fastp v22.0[65] with following options: -q 20 -n 10 -t 1 -T 1 -l 20 -w 16. The cleaned reads were mapped to the reference genome using bwa-mem2 22.1[66] with default parameters. We removed PCR duplicates using fixmate and markdup function of samtools 1.13[67]. The sequence reads were mapped to the reference genome, yielding average mapping rates of 91.61% and 89.5% for the slow- and fast-growing species, respectively. To identify reliable SNVs, we utilized two SNP callers (bcftools mpileup[67,68] and GATK (4.2.4.0) HaplotypeCaller[69]) and extracted SNVs detected by both (Supplementary Fig. 2).

We first called SNVs with BCFtools 1.13[70] mpileup at three different thresholds; threshold 1 (T40): mapping quality (MQ) = 40, base quality (BQ) = 40; threshold 2 (T30): MQ = 30, BQ = 30; threshold 3 (T20): MQ = 20, BQ = 20. SNVs detected under each threshold were pooled for further analyses, with duplicates removed. We normalized indels using bcftools norm for vcf files. We removed indels and missing data using vcftools 0.1.16[71].

Second, we called SNVs using GATK (4.2.4.0) HaplotypeCaller and merged the individual gvcfs into a vcf file containing only variant sites. We removed indels from the vcf using the GATK SelectVariants. We filtered out unreliable SNVs using GATK VariantFiltration with the following filters: QD (Qual By Depth) < 2.0, QUAL (Base Quality) < 30.0, SOR (Strand Odds Ratio) > 4.0, FS (Fisher Strand) > 60.0, MQ (RMS Mapping Quality) < 40.0, MQRankSum (Mapping Quality Rank Sum Test) < -12.5, ReadPosRankSum (Read Pos Rank Sum Test) < -8.0. After performing independent SNV calling for each biological replicate using each SNP caller, we extracted SNVs that were detected in both replicates for each SNP caller. We further extracted SNVs that were detected by both bcftools mpileup and GATK HaplotypeCaller (Supplementary Fig. 2) using Tassel5[72] and a custom python script, generating potential SNVs for each threshold. Finally, SNVs detected at any of the three thresholds were extracted to obtain candidate SNVs. The number of SNVs at each filtering step can be found in Supplementary Data 4.

The candidate SNV calls were manually confirmed by two independent researchers using the IGV browser[73]. We removed sites from the list of candidates if there were fewer than five high-quality reads (MQ > 20) in at least one branch sample among the 16 samples. After labeling branches carrying the called variant as somatic mutations, we compared the observed pattern with the genotyping call and extracted SNVs that were supported more than one read in both biological replicates (Supplementary Fig. 4a). We illustrated three types of false positive

SNVs that were removed from the list of candidates in Supplementary Fig. 4b–d. The final set of SNVs can be found in Supplementary Data 7. Proportion of false positive SNVs for each threshold are illustrated in Supplementary Fig. 5.

The NJ tree for each individual was generated using MEGA11[74] based on the matrix of the number of sites with somatic SNVs present between each pair of branches and edited using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Most of the somatic SNVs were heterozygous, whereas 4% of the total SNVs (46/1136) were homozygous (Supplementary Data 7). The homozygous sites were treated as a single mutation due to the likelihood of a genotyping error being higher than the probability of two mutations occurring at the same site.

Inter-individual SNV discovery

We also identified SNVs between pairs of individuals within each species as inter-individuals SNVs. The method for calling inter-individual SNVs was the same as for intra-individual SNVs, except that only threshold 2 (MQ = 30, BQ = 30) for BCFtools 1.13[70] was used. We extracted SNVs that are present in all branches within an individual using Tassel5[72]. To exclude ambiguous SNV calls, we removed SNVs within 151 bp of indels that were called with BCFtools 1.13[70] with the option of threshold 2. We eliminated SNVs within 151 bp of sites with a depth value of zero that occur in more than ten consecutive sites. We also removed SNVs that had a depth smaller than five or larger than $d + 3\sqrt{d}$, where d represents the mean depth of all sites[75]. Due to the large number of candidates for inter-individual SNVs, the manual checking process was skipped.

Somatic SNVs confirmation by amplicon sequencing

We verified the reliability of the final set of somatic SNVs by amplicon sequencing approximately 5% of the SNVs in *S. laevis* (31 and 10 SNVs for S1 and S2, respectively). We used multiplexed phylogenetic marker sequencing method (MPM-seq[76]) with modifications to the protocol as follows: to amplify 152–280 bp fragments, the first PCR primers comprising tail sequences for the second PCR primers were designed on the flanking regions of each SNV. The first PCR was conducted using the Fast PCR cycling kit (Qiagen, Düsseldorf, Germany) under the following conditions: an initial activation step at 95°C for 5 minutes, followed by 30 cycles of denaturation at 96°C for 5 seconds, annealing at 50/54/56°C for 5 seconds, and extension at 68°C for 10 seconds. This was followed by a final incubation at 72°C for 1 minute. Subsequent next-generation sequencing was performed on an Illumina MiSeq platform using the MiSeq Reagent Kit v2 (300 cycles: Illumina).

Amplicon sequencing reads were mapped to the reference genome using bwa-mem2 22.1[66] with default parameters. Using bcftools mpileup[70], we called the genotypes of all sites on target regions and eliminated candidate sequences with MQ and BQ less than 10. The final set of sites selected for confirmation consisted of 24 for the S1 individual and 9 for the S2 individual. We manually confirmed the polymorphic patterns at the target sites using the IGV browser[73]. If the alternative allele was present or absent in all eight branches in the amplicon sequence, the site was determined as fixed. The site was determined as mismatch if the difference of

polymorphic patterns between the somatic SNV calls and amplicon sequence was supported by more than four reads per branch. The sites that were neither fixed nor mismatched were determined as true. 93.9% (31/33) of SNVs at the final target sites were confirmed to exhibit a polymorphic pattern that exactly matched between the somatic SNV calls and amplicon sequence (Supplementary Data 5).

Somatic mutation rates per growth and per year

To estimate the somatic mutation rate per nucleotide per growth (μ_g), a linear regression analysis of the number of somatic SNVs against the physical distance between sampling positions within an individual was conducted using the lm package, with an intercept of zero, in R version 3.6.2. The somatic mutation rate per nucleotide per growth was estimated as:

$$\mu_g = \frac{b}{2 \times R},$$

where b indicates the slope of linear regression and R denotes the number of callable sites, respectively. Note that the denominator includes a factor of two due to diploidy. A site was considered callable when it passed the filters as the polymorphic sites, that is, a mapping quality of at least 40 using GATK, a mapping quality of at least 20 using BCFtools, and a depth greater than or equal to 5. This resulted in 388,801,756 and 320,739,335 base pairs for S1 and S2 and 327,435,618 and 263,488,812 base pairs for F1 and F2, respectively.

The somatic mutation rate per nucleotide per year (μ_y) was estimated as:

$$\mu_y = \frac{M}{2 \times R \times A}.$$

Here, M indicate the total number of SNVs accumulated from the base (ID 0 in Fig. 1a; Supplementary Data 2) to the branch tip and A represents tree age, respectively. R denotes the number of callable sites that was also used to estimate μ_g. Because there are seven branch tips for each tree (Fig. 1a), we estimated μ_y for each of branch tips and then calculated the mean and 95% confidence interval for each tree (Supplementary Table 3).

Mutational spectrum

Mutational spectra were derived directly from the reference genome and alternative alleles at each variant site. There are a total of six possible classes of base substitutions at each variant site: A:T>G:C (T>C), G:C>A:T (C>T), A:T>T:A (T>A), G:C>T:A (C>A), A:T>C:G (T>G), and G:C>C:G (C>G), By considering the bases immediately 5′ and 3′ to each mutated base, there are a total of 96 possible mutation classes, referred to as triplets, in this classification. We used seqkit[77] to extract the triplets for each variant site. To count the number of each triplet, we used the Wordcount tool in the EMBOSS web service (https://www.bioinformatics.nl/cgi-bin/emboss/wordcount). We calculated the fraction of each mutated triplet by dividing the number of mutated triplets by the total number of triplets in the reference genome.

We compared the mutational signatures of our tropical trees to those of single-base substitution (SBS) signatures in human cancers using Catalogue Of Somatic Mutations In Cancer (COSMIC) compendium of mutation signatures (COSMICv.2[78–80], available at https://cancer.sanger.ac.uk/cosmic/signatures_v2). Cosine similarity was calculated between each tropical tree species and each SBS signature in human cancers.

Testing selection of somatic and inter-individual SNVs

To test whether somatic and inter-individual SNVs are subject to selection, we calculated the expected rate of non-synonymous mutation. For the CDS of length L_{cds}, there are possible numbers of mutations of length of $3L_{\mathrm{cds}}$ (Supplementary Fig. 6). We classified all possible mutations into three types based on the codon table: synonymous, missense, and nonsense (Supplementary Fig. 6). Each type of mutation was counted for each of the six base substitution classes (Supplementary Fig. 6). We generated count tables based on two distinct categories of CDS: those that included all isoforms and those that only encompassed primary isoforms (Supplementary Data 8). As the two tables were largely congruent, we employed the version which included all isoforms of CDS.

Using the count table and background mutation rate for each category of substitution class, we calculated the expected number of synonymous (λ_S) and non-synonymous mutations (λ_N) (Supplementary Fig. 6). As a background mutation rate, we adopted the observed somatic mutation rates in the six substitution classes in the intergenic region (Supplementary Table 5), assuming that the intergenic region is nearly neutral to selection. Because the number of nonsense somatic mutation is small, we combined missense and nonsense mutations as non-synonymous. The intergenic regions were identified as the regions situated between 1 kbp upstream of the start codon and 500 bp downstream of the stop codon. Expected rate of synonymous mutation (p_N) is given as $\lambda_N/(\lambda_S + \lambda_N)$. Given the observed number of non-synonymous and synonymous mutations, we rejected the null hypothesis of neutral selection using a binomial test with the significance level of 5% (Supplementary Table 4). We used the package binom.test in R v3.6.2.

We also used the observed somatic mutation rate in the whole genome (Supplementary Table 5), including genic and intergenic regions, as the background mutation rate and confirmed the robustness of our conclusion (Supplementary Tables 4). The somatic mutation rates in the intergenic region and the whole genome were calculated for each species by pooling the data from two individuals (Supplementary Table 5). While cancer genomics studies have accounted for more detailed context-dependent mutations, such as the high rate of C>T at CpG dinucleotides[81] or comprehensive analysis of 96 possible substitution classes in triplet context[82], the number of SNVs in our tropical trees is too small to perform such a comprehensive analysis. Therefore, we used the relatively simple six base substitution classes. The genes with somatic SNVs can be found in Supplementary Data 7.

Data availability

The raw sequencing data have been deposited to DDBJ under accessons DRX404986-

DRX405036 for *S. laevis* and DRX412534-DRX412566 for *S. leprosula*. The genome assembly and the gene annotation are available under accessions BSQA01000001-BSQA01007745 for *S. laevis* and BSQB01000001-BSQB01000070 for *S. leprosula*.

Code availability

The codes for the bioinformatics pipeline to process whole genome sequencing data is available from https://github.com/ku-biomath/Shorea_mutation_detection.

Acknowledgements

The authors would like to thank to M. Seki for his assistance with statistical analysis, S.K. Hirota for his technical support in molecular experiments We also thank Y. Iwasa, H. Tachida, M.M. Manuel, N. Spisak, M. Przeworski and M, Nordborg for their very insightful comments on the initial draft of our manuscript.

Authors contributions

A.S. conceived and designed the analysis; M.N, S.I, W., S.P., N.T., Y.S. and A.S. collected samples; K.O., R.I., A.M.M., V.N., and Y.S. performed molecular experiments; R.I., E.S., S.T. and A.S. analyzed data; T.F. and M.K. performed reference genome construction. A.S. leaded writing the paper with input from all authors. This study was funded by JSPS KAKENHI (JP17H06478 to A.S. and JP22H04925 (PAGS) to M.K.).

Additional information

Additional supporting information will be found in the online version of this article.

Methods references

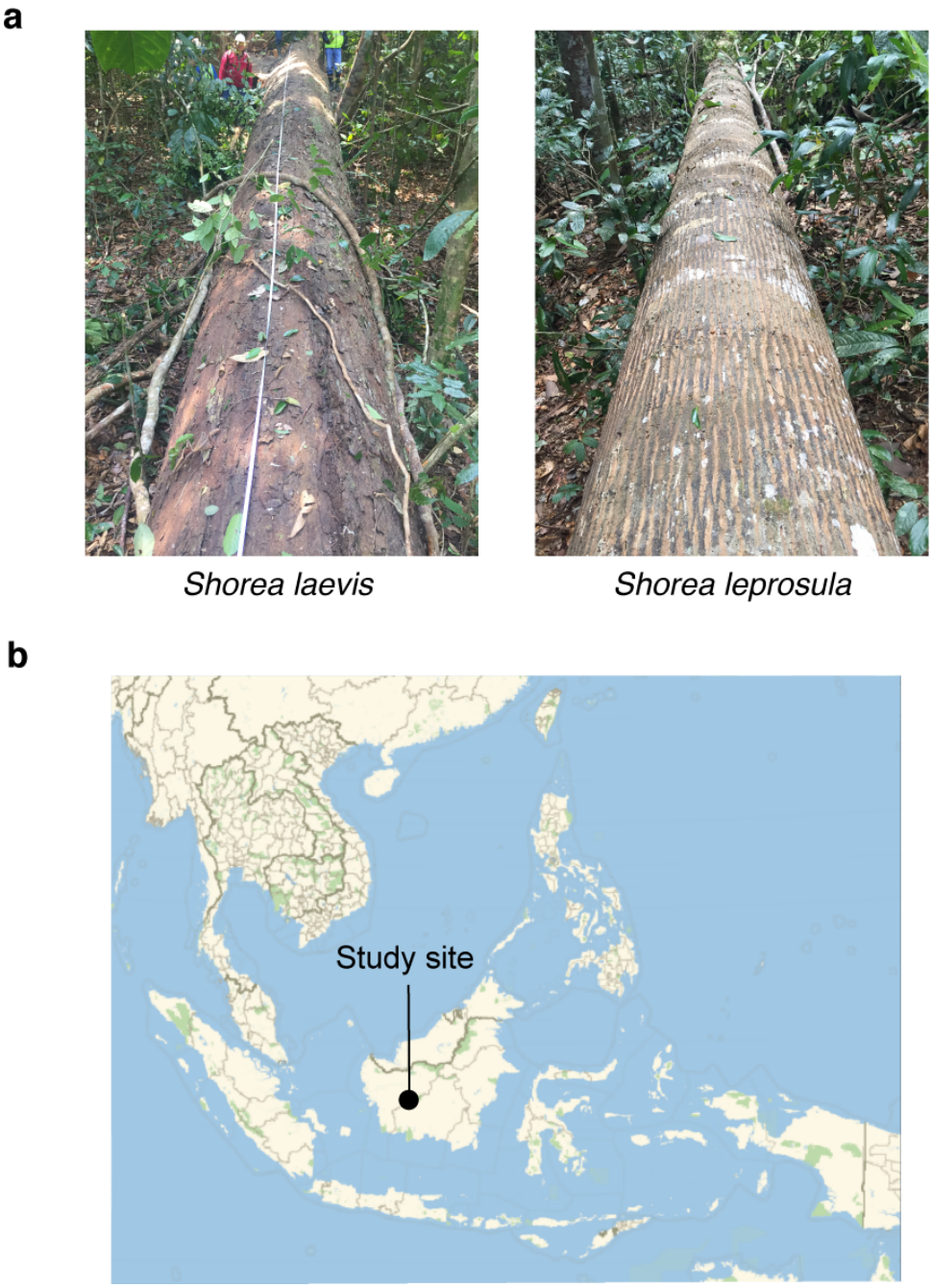
41. Widiyatno, W., Soekotjo, S., Naiem, M., Purnomo, S. & Setiyanto, P. E. Early performance of 23 dipterocarp species planted in logged-over rainforest. *Journal of Tropical Forest Science* **26**, 259–266 (2014).
42. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11–15 (1987).
43. Toyama, H. *et al.* Effects of logging and recruitment on community phylogenetic structure in 32 permanent forest plots of Kampong Thom, Cambodia. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 1–13 (2015).
44. Doyle, J. DNA Protocols for plants in molecular techniques in taxonomy. (Springer, 1991). doi:https://doi.org/10.1007/978-3-642-83962-7_18.
45. Yeoh, S. H. *et al.* Unravelling proximate cues of mass flowering in the tropical forests of South-East Asia from gene expression analyses. *Mol Ecol* **26**, 5074–5085 (2017).
46. Ng, K. K. S. *et al.* The genome of *Shorea leprosula* (Dipterocarpaceae) highlights the ecological relevance of drought in aseasonal tropical rainforests. *Commun Biol* **4**, 1–14 (2021).
47. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019).
48. Kundu, R., Casey, J. & Sung, W.-K. HyPo: Super Fast & Accurate polisher for long read genome assemblies. *bioRxiv* (2019).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, (2012).
50. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, (2021).

51. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* **38**, (2021).
52. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, (2019).
53. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. *http://www.repeatmasker.org*.
54. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3**, (2021).
55. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, (2019).
56. Zdobnov, E. M. *et al.* OrthoDB in 2020: Evolutionary and functional annotations of orthologs. *Nucleic Acids Res* **49**, (2021).
57. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, (2021).
58. Hart, A. J. *et al.* EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol Ecol Resour* **20**, (2020).
59. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, (2021).
60. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, (2016).
61. Powell, S. *et al.* EggNOG v4.0: Nested orthology inference across 3686 organisms. *Nucleic Acids Res* **42**, (2014).
62. Vurture, G. W. *et al.* GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
63. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
64. Ng, C. H. *et al.* Genome size variation and evolution in Dipterocarpaceae. *Plant Ecol Divers* **9**, 437–446 (2016).
65. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
66. Vasimuddin, M., Sanchit, M., Heng, L. & Srinivas, A. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. *Proceedings - 2019 IEEE 33rd International Parallel and Distributed Processing Symposium, IPDPS 2019* 314–324 (2019) doi:10.1109/IPDPS.2019.00041.
67. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
68. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
69. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, (2010).
70. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).
71. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
72. Bradbury, P. J. *et al.* TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, (2007).
73. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Cancer Research* vol. 77 Preprint at https://doi.org/10.1158/0008-5472.CAN-17-0337 (2017).
74. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol Biol Evol* **38**, (2021).
75. Li, H. & Wren, J. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
76. Suyama, Y. *et al.* Complementary combination of multiplex high-throughput DNA sequencing for molecular phylogeny. *Ecol Res* **37**, 171–181 (2022).
77. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, 1–10 (2016).
78. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246–259 (2013).
79. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).

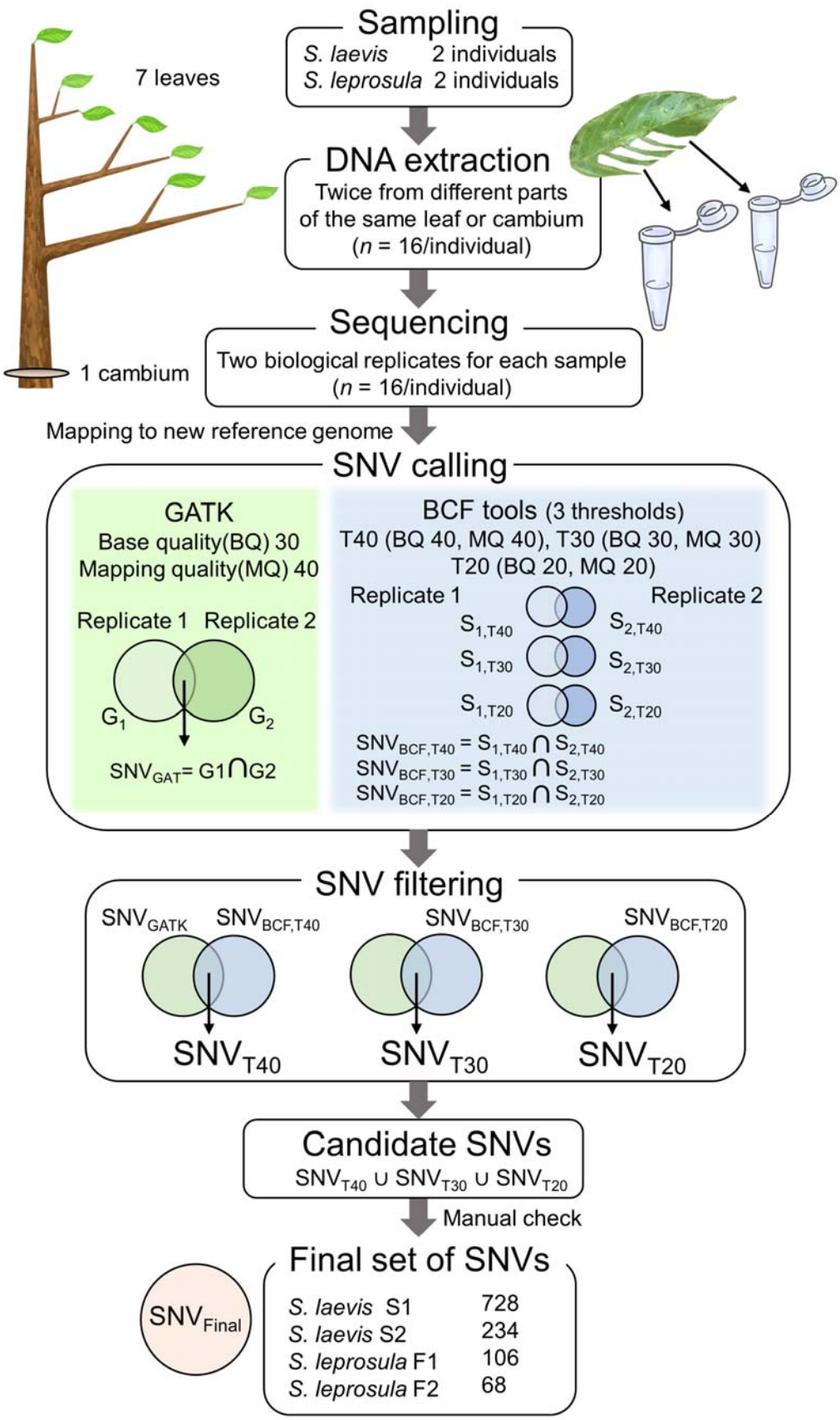
80. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
81. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173**, 2187–2198 (2006).
82. Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029-1041.e21 (2017).

Competing interest

The authors declare that they have no competing financial and non-financial interests.

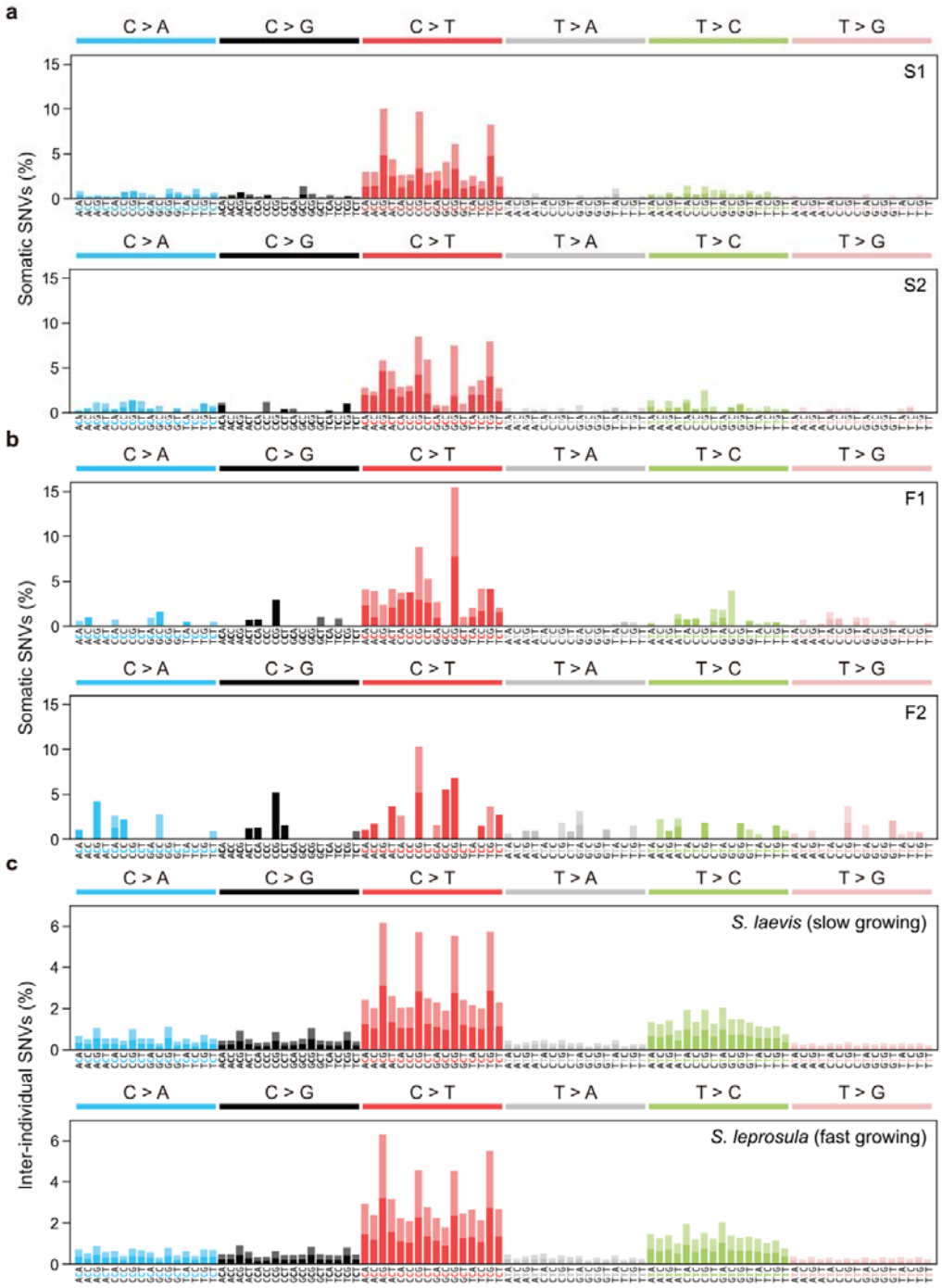


Supplementary Fig. 1 | Target tropical trees and location of study site. a, Images of *S. laevis* (S1), a slow-growing species, and *S. leprosula* (F1), a fast-growing species. **b**, Location of the study site in central Borneo, Indonesia.

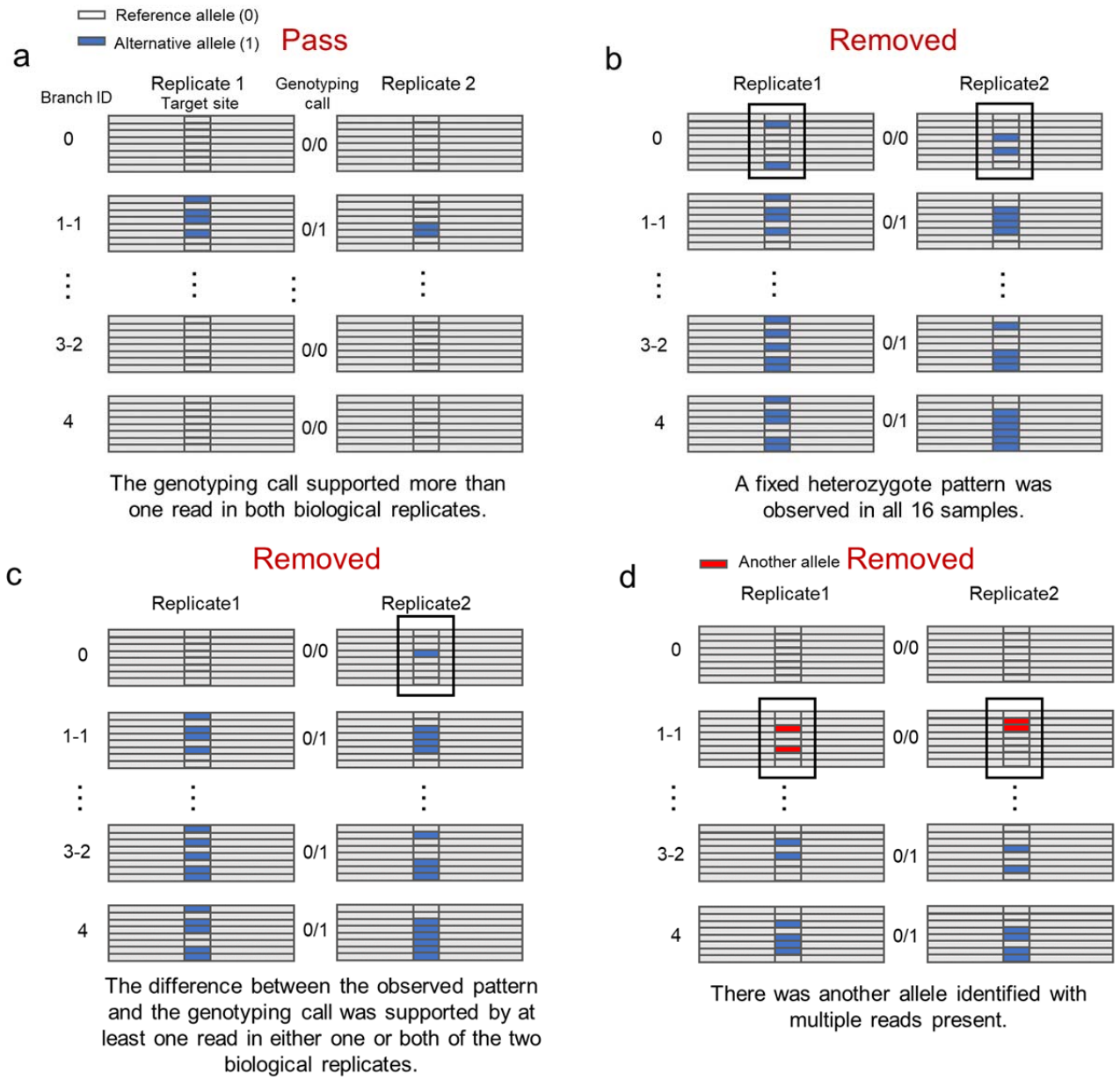

Sampling
S. laevis 2 individuals
S. leprosula 2 individuals
7 leaves
1 cambium
DNA extraction
Twice from different parts of the same leaf or cambium
(n = 16/individual)
Sequencing
Two biological replicates for each sample
(n = 16/individual)
Mapping to new reference genome
SNV calling
GATK
Base quality(BQ) 30
Mapping quality(MQ) 40
Replicate 1 Replicate 2
G1 G2
SNVGAT= G1∩G2
BCF tools (3 thresholds)
T40 (BQ 40, MQ 40), T30 (BQ 30, MQ 30)
T20 (BQ 20, MQ 20)
Replicate 1 Replicate 2
S1,T40 S2,T40
S1,T30 S2,T30
S1,T20 S2,T20
SNVBCF,T40 = S1,T40 ∩ S2,T40
SNVBCF,T30 = S1,T30 ∩ S2,T30
SNVBCF,T20 = S1,T20 ∩ S2,T20
SNV filtering
SNVGATK SNVBCF,T40
SNVBCF,T30
SNVBCF,T20
SNVT40 SNVT30 SNVT20
Candidate SNVs
SNVT40 ∪ SNVT30 ∪ SNVT20
Manual check
Final set of SNVs
SNVFinal
S. laevis S1 728
S. laevis S2 234
S. leprosula F1 106
S. leprosula F2 68

774

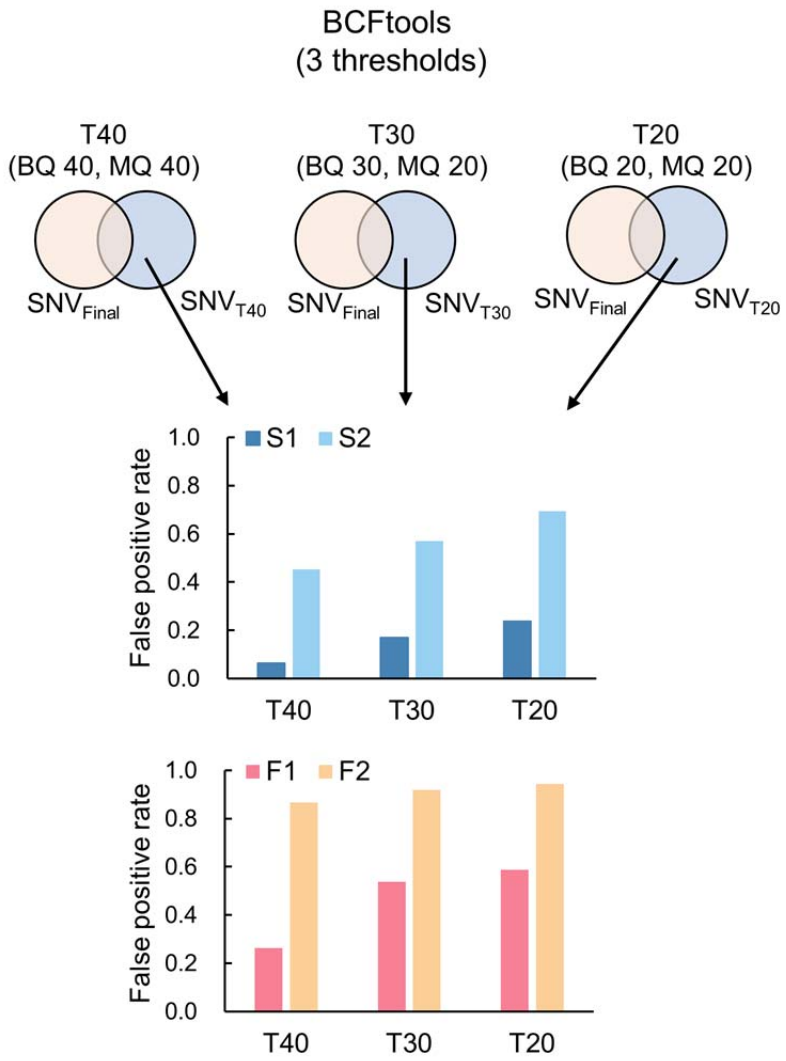
775 **Supplementary Fig. 2 | Workflow for identifying *de novo* somatic SNVs.** 8 samples (seven
776 leaves and one cambium) were collected from four trees (two trees from each species). DNA was
777 extracted twice independently from each sample and sequenced independently. Reads were
778 mapped to the reference genome and used for SNV calling and filtering. SNVs over 8 samples
779 were called using GATK HaplotypeCaller (GATK) and Bcftools mpileup (BCF tools) for each
780 set of biological replicates from 7 branches and 1 cambium independently, generating potential
781 SNVs for each set of replicates and for each SNP caller (G_1 and G_2 for GATK, B_1 and B_2 for
782 BCF tools). For BCF tools, we set three thresholds (T40, T30, and T20) with different base
783 quality (BQ) and mapping quality (MQ). SNVs detected in both replicates were extracted for
784 each SNP callers and generated potential SNVs for each SNP caller, SNV_{GATK} for GATK and
785 SNV_{BCF} for Bcftools with three thresholds. These SNVs were filtered by extracting SNVs
786 detected in both SNP callers, generating potential SNVs for each threshold: SNV_{T40}, SNV_{T30},
787 and SNV_{T20}. Finally, SNVs detected at any of the three thresholds were extracted to obtain
788 candidate SNVs. We checked the candidate SNVs manually and obtained a final set of SNVs,
789 SNV_{Final}.
790



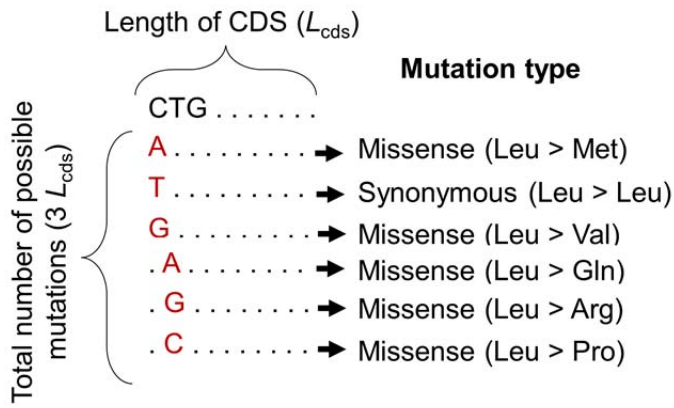
Supplementary Fig. 3 | Mutational spectra of somatic and inter-individual substitutions. a, Somatic mutation spectra for S1 and S2 individuals in *S. laevis*. **b**, Somatic mutation spectra for F1 and F2 individuals in *S. leprosula*. **c**, Inter-individual SNVs between S1 and S2 (upper panel) and between F1 and F2 (lower panel). The horizontal axis shows 96 mutation types on a trinucleotide context, coloured by base substitution type. Different colours in each bar indicate complementary bases.



Supplementary Fig. 4 | Manual confirmation of candidate SNVs. a, SNVs that passed manual confirmation. **b,** SNVs that were removed due to their fixed heterozygote pattern. **c**, SNVs that have been removed due to the difference between the observed pattern and the genotyping call. **d**, SNVs that were removed due to the presence of another allele with multiple reads.



Supplementary Fig. 5 | Proportion of false positive SNVs for *S. laevis* (S1, S2) and *S. leprosula* (F1, F2). Proportion of false positive SNVs was calculated by extracting the subset of potential SNVs that were not included in the final set for each threshold and dividing by the total number of potential SNVs for each threshold.



Base substitution class	Base substitution class ID (k)	Background mutation rate	Possible number of mutation type		
			synonymous	missence	nonsense
A:T > G:C	1	r_1	$N_{\mathrm{S}}^{(1)}$	$N_{\mathrm{M}}^{(1)}$	$N_{\mathrm{Non}}^{(1)}$
G:C > A:T	2	r_2	$N_{\mathrm{S}}^{(2)}$	$N_{\mathrm{M}}^{(2)}$	$N_{\mathrm{Non}}^{(2)}$
A:T > T:A	3	r_3	$N_{\mathrm{S}}^{(3)}$	$N_{\mathrm{M}}^{(3)}$	$N_{\mathrm{Non}}^{(3)}$
G:C > T:A	4	r_4	$N_{\mathrm{S}}^{(4)}$	$N_{\mathrm{M}}^{(4)}$	$N_{\mathrm{Non}}^{(4)}$
A:T > C:G	5	r_5	$N_{\mathrm{S}}^{(5)}$	$N_{\mathrm{M}}^{(5)}$	$N_{\mathrm{Non}}^{(5)}$
G:C > C:G	6	r_6	$N_{\mathrm{S}}^{(6)}$	$N_{\mathrm{M}}^{(6)}$	$N_{\mathrm{Non}}^{(6)}$

Possible number of mutation type k

$$N_{\mathrm{T}}^{(k)} = N_{\mathrm{S}}^{(k)} + N_{\mathrm{M}}^{(k)} + N_{\mathrm{Non}}^{(k)}$$

Expected number of synonymous mutation

$$\lambda_S = L_{\mathrm{cds}} \sum_{k=1}^{6} r_k \frac{N_{\mathrm{S}}^{(k)}}{N_T^{(k)}}$$

Expected number of non-synonymous mutation

$$\lambda_N = L_{\mathrm{cds}} \sum_{k=1}^{6} r_k \left(\frac{N_{\mathrm{M}}^{(k)} + N_{\mathrm{Non}}^{(k)}}{N_T^{(k)}} \right)$$

Expected rate of non-synonymous mutation

$$p_N = \lambda_N / (\lambda_S + \lambda_N)$$

Supplementary Fig. 6 | A calculation scheme for the expected rate of non-synonymous mutation. The possible numbers of synonymous (N_{S}), missense (N_{M}), and nonsense (N_{Non}) mutations were counted for each of six base substitution classes from all possible mutations in CDS of length L_{cds} and used for the calculation of expected rate of non-synonymous mutation. For non-synonymous mutation, we pooled the number for missense and nonsense mutations. The background mutation rate for each substitution class i (r_i) is calculated from the observed somatic substitutions in intergenic regions.

Supplementary Table 1.
Summary statistics of the studied trees.
Height and DBH were directly measured for two individuals of *S. laevis* and *S. leprosula*. Age was estimated as DBH divided by a mean annual increment (MAI).

Species	Individual ID	Height (m)	DBH (cm)	Estimated age (year)
S. laevis	S1	51.5	122.8	324.9
S. laevis	S2	36.6	70.7	187.0
S. leprosula	F1	47.6	95.5	78.8
S. leprosula	F2	40.1	65.6	54.1

Supplementary Table 2.
Summary statistics of genome assemblies for *S. laevis* and *S. leprosula*. We assembled the genome using DNA extracted from the apical leaf at branch 1-1 of the tallest individual of each species (S1 and F1). Summary statistics of genome assemblies are listed here.

Category	Features	*S. laevis*	*S. leprosula*
Assembly	Sequencing Technology	PacBio CLR + Illumina	PacBio HiFi
	Number of Contigs	7,695	70
	Total length of contigs (nt)	539,910,052	378,538,404
	Contig N50 (nt)	206,905	39,554,423
	Number of Scaffolds	7,745	70
	Total length of scaffolds (nt)	540,441,910	378,538,404
	Scaffold N50 excluding gaps (nt)	219,154	39,554,423
	Maximum scaffold length excluding gaps (nt)	3,442,711	58,083,003
	Number of scaffolds > 1kbp	7,427	70
	Number of N's per 100 kbp	1	0
	GC-content (%)	33.71	33.14
Annotation	Number of protein-coding genes	52,935	40,665
	% of repetitive sequences (identified by EDTA)	42.4	39.5
	% of Complete BUSCO genes	97.9	97.8
	% of Complete and single-copy BUSCO genes	65.8	76.7

Supplementary Table 3.

Somatic mutation rates. The somatic mutation rate per nucleotide per meter was estimated as $\mu_g = \frac{b}{2\times R}$, where b indicates the slope of linear regression. The somatic mutation per nucleotide per year (μ_y) was estimated as $\mu_y = \frac{M}{2\times R\times A}$, where M indicates the total number of SNVs accumulated from the base to the branch tip and A represents tree age, respectively. R denotes the number of callable sites.

Species	***S. laevis***		***S. leprosula***	
individual ID	S1	S2	F1	F2
Coefficient of linear regression (95% confidence interval)	5.503 (4.985–6.020)	2.740 (2.561–2.919)	1.149 (1.06–1.239)	0.679 (0.552–0.806)
μ_g (95% confidence interval)	7.08×10^{-9} ($6.41–7.74\times10^{-9}$)	4.27×10^{-9} ($3.99–4.55\times10^{-9}$)	1.77×10^{-9} ($1.64–1.91\times10^{-9}$)	1.29×10^{-9} ($1.05–1.53\times10^{-9}$)
μ_y (95% confidence interval)	7.58×10^{-10} ($4.67–10.1\times10^{-10}$)	7.85×10^{-10} ($6.12–9.57\times10^{-10}$)	9.63×10^{-10} ($7.56–11.7\times10^{-10}$)	6.46×10^{-10} ($3.55–9.38\times10^{-10}$)
Average of μ_g within species	5.67×10^{-9}		1.53×10^{-9}	
Average of μ_y within species	7.71×10^{-10}		8.05×10^{-10}	

842 **Supplementary Table 4.**

843 **Results of the binomial test for selection on somatic and inter-individual SNVs.** To test
844 whether somatic and inter-individual SNVs are subject to selection, we calculated the expected
845 rate of non-synonymous mutation. Given the observed number of non-synonymous and
846 synonymous mutations, we rejected the null hypothesis of neutral selection using a binomial test
847 with the significance level of 5%. p_N_expected and p_N_observed represent the expected and
848 observed rate of non-synonymous substitutions.

849

Type of SNVs	**Genomic region used for background mutation inference**	**Summary statistics**	***S. laevis***	***S. leprosula***
Somatic SNVs	Intergenic	p_N_expected	0.712	0.717
		p_N_observed	0.734	0.733
		P value	0.711	1
	Whole genome	p_N_expected	0.716	0.723
		p_N_observed	0.734	0.733
		P value	0.803	1
Inter-individual SNVs	Intergenic	p_N_expected	0.712	0.717
		p_N_observed	0.546	0.592
		P value	**$<2.20\times10^{-16}$**	**$<2.20\times10^{-16}$**
	Whole genome	p_N_expected	0.716	0.723
		p_N_observed	0.546	0.592
		P value	**$<2.20\times10^{-16}$**	**$<2.20\times10^{-16}$**

850

Supplementary Table 5.

Somatic mutation rates for six substitution classes. Somatic mutation rates for six substitution classes were calculated based on the observed number of SNVs both from the intergenic region and the whole genome. S1+S2 and F1+F2 represent the use of pooled data from two individuals for each species: *S. laevis* (S1, S2) and *S. leprosula* (F1, F2). The values based on the pooled data (indicated in bold type) were used to calculate the expected rate of non-synonymous mutation.

Genomic region used to determine background mutation rate	Tree ID	Transition			Transversion			Total number of SNVs
		A:T > G:C	G:C >A:T	A:T > T:A	G:C > T:A	A:T > C:G	G:C > C:G	
Intergenic	S1	0.141	0.585	0.105	0.066	0.062	0.041	532
	S2	0.193	0.540	0.107	0.067	0.073	0.020	150
	S1+S2	**0.152**	**0.575**	**0.106**	**0.066**	**0.065**	**0.037**	**682**
	F1	0.225	0.588	0.025	0.038	0.088	0.038	80
	F2	0.297	0.297	0.162	0.108	0.054	0.081	38
	F1+F2	**0.248**	**0.496**	**0.068**	**0.060**	**0.077**	**0.051**	**118**
Whole genome	S1	0.168	0.541	0.099	0.087	0.058	0.048	728
	S2	0.205	0.496	0.103	0.090	0.064	0.043	234
	S1+S2	**0.177**	**0.530**	**0.100**	**0.087**	**0.059**	**0.047**	**962**
	F1	0.208	0.547	0.028	0.066	0.104	0.047	106
	F2	0.250	0.265	0.162	0.103	0.147	0.074	68
	F1+F2	**0.224**	**0.437**	**0.080**	**0.080**	**0.121**	**0.057**	**174**