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

Authors: 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], Akiko Satake[1*]

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.

*Correspondence to:

Akiko Satake (akiko.satake@kyudai.jp)

The rates and patterns of somatic mutations in wild plants, as well as how they relate to longevity, are largely unknown[1–3]. Here, we examined the somatic mutation landscapes of slow- and fast-growing tropical species in central Borneo, Indonesia. Using newly-constructed genomes, we identified an average of 480 mutations in the slow-growing species (265-year-old, 44.1 m in height), which was five times greater than that observed in the fast-growing species (66-year-old, 43.9 m). The number of somatic mutations increased linearly with branch length. The somatic mutation rate per meter was higher in the slow-growing species, yet the rate per year remained constant across both species. The mutational spectra exhibited a dominance of spontaneous mutations, specifically cytosine-to-thymine substitutions at CpG sites. An analysis of nucleotide substitutions at both the intra- and inter-individual level revealed that somatic mutations are neutral within an individual, but those mutations transmitted to the next

generation are subject to purifying selection. We developed a model to evaluate the relative contribution of cell division on mutational processes, and postulate that cell-division independent mutagenesis predominates. These findings deepen our understanding of mutational processes underlying the generation of genetic diversity in a tropical ecosystem.

Biodiversity ultimately results from mutations that provide genetic variation for organisms to adapt to their environments. However, how and when mutations occur is poorly understood. Recent studies on somatic mutations that accumulate in long-lived trees have begun to uncover the rate of naturally occurring mutations and their relationship to longevity. The rate of somatic mutations per year in a 234-year-old oak tree has been found to be surprisingly low[4] compared to the rate in an annual herb[5]. Similar analyses in other long-lived trees have also shown low mutation rates in both broadleaf trees[6–10] and conifers[11], consistent with the recent finding in mammals that the rate of somatic mutations exhibits a strong inverse relationship with lifespan[12].

Low mutation rates are advantageous to cope with deleterious mutation effects, such as aging[13] and mutational meltdown[14,15]. In contrast, low mutation rates reduce genetic variation and the rate of evolution because, unlike animals, somatic mutations in plants can be transmitted to offspring[6]. In tropical ecosystems, which are among the most diverse biomes on Earth, the role of somatic mutations in generating genetic variations can be more pronounced than in organisms living in higher latitudes[16]. However, there is currently no knowledge on the relationship between somatic mutation rate and lifespan in tropical organisms.

To investigate the rates and patterns of somatic mutation and their relation to lifespans 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. The comparative analysis of slow- and fast-growing species enables us to investigate the relationship between lifespan and mutagenesis because slow-growing trees have longer lifespans compared to their fast-growing counterparts[17,18].

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[19]. 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 (Extended Data Fig. 1a). *S. laevis* is a slow-growing species[20], 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 Table 1). In contrast, *S. leprosula* exhibits a faster growth rate, with an MAI of 1.21 cm/year ($n = 18$; Supplementary Table 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 (Extended Data 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; Extended Data 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 Table 2) and DBH (Supplementary Table 3). The average heights of the slow- and fast-growing

species were 44.1 m and 43.9 m, respectively (Fig. 1a; Supplementary Table 3). 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 age of the trees to be, on average, 256 and 66-year-old for the slow- and fast-growing species, respectively (Supplementary Table 3).

To identify somatic mutations, we constructed new reference genomes of the slow- and fast-growing species. We generated sequence data using long-range 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, respectively (Supplementary Table 4). 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 (Extended Data 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 Table 5). We identified Single Nucleotide Variants (SNVs) within the same individual by identifying those that were identical within two biological replicates of each sample (Extended Data Fig. 2). We identified 728 and 234 SNVs in S1 and S2, and 106 and 68 SNVs in F1 and F2, respectively (Extended Data Fig. 2; Supplementary Table 6). 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 Table 7).

Somatic mutations 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. Unlike previous studies[7,8], 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 branches 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 branches by the size of the reference genome (Fig. 2b). The average rate of somatic mutation per meter was 7.63×10^{-9} site^{-1} m^{-1} for the slow-growing species, which is 3.16-fold higher than the average rate of 2.41×10^{-9} site^{-1} m^{-1} observed in the fast-growing species (Fig. 2b; Supplementary Table 8). This result indicates that the slow-growing tree accumulates more *de novo* 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[21] and 0.91mm and 22.7 μm for the fast-growing species[22]).

Based on the estimated age of each tree, somatic mutation rates per site per year were calculated for each tree. On average, resultant values were largely similar between the two species, with 1.24×10^{-9} and 1.40×10^{-9} site^{-1} year^{-1} for the slow- and fast-growing species, respectively (Fig. 2b; Supplementary Table 8). This 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 site per year in *Shorea* are significantly higher than those previously reported in other long-lived trees such as *Quercus robur*[4] ($1.79–2.22 \times 10^{-10}$), *Populus trichocarpa*[9] (1.33×10^{-10}), *Eucalyptus melliodora*[8] ($1.16–11.2 \times 10^{-10}$) and *Picea sitchensis*[11] (7.4×10^{-11}). The somatic mutation rates of tropical trees were on the same order of magnitude as those of the annual herb *Arabidopsis thaliana*[5] (7.1×10^{-9} per generation). 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; Extended Data 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[23,24]. Methylated CpG sites spontaneously deaminate, leading to TpG sites and increasing the number of C>T substitutions[25].

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[26–28]). The mutational spectra were largely similar to the dominant mutation signature in human cancers known as SBS1 (cosine similarity = 0.789 and 0.597 for the slow- and fast-growing species; Supplementary Table 9). SBS1 is believed to result from the spontaneous deamination of 5-methylcytosine. The mutational spectra were also comparable to another dominant signature in human cancers, SBS5 (cosine similarity = 0.577 and 0.558 for the slow- and fast-growing species; Supplementary Table 9), 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[29,30]. 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 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 10). 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 (Extended Data 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[31] 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: cell-division dependent (α) and cell-division independent (β) mutagenesis. This can be represented as $\mu = \alpha + \beta\tau$, where τ is the duration of cell cycle measured in years. The cell-division dependent mutation emanates from errors that occur during DNA replication, such as the misincorporation of a nucleotide during DNA synthesis. The cell-division 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 above formula encompasses the effects from two distinct sources of somatic mutations: cell-division dependent (α/τ) and cell-division independent (i.e., time-dependent) mutagenesis (β). When β is significantly greater than α, the rate of somatic mutation accumulation becomes constant, independent of growth rate.

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 $\tau_S/\tau_F = 3.2$ (Supplementary Table 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 $\mu_S/\mu_F = (\alpha + \beta\tau_S)/(\alpha + \beta\tau_F) = 3.16$, which is nearly equivalent to the ratio of cell cycle duration τ_S/τ_F. This consistency can be explained by the substantial contribution of the cell-division independent mutagenesis (i.e. $\beta \gg \alpha$) to the somatic mutation rate,

as long as the magnitudes for α and β are similar between the two species. The relative contribution of cell-division dependent to cell-division independent mutagenesis (α/β) can be estimated as 0.015, a significantly small value. This suggests that the cell-division independent mutagenesis has a greater impact on the overall somatic mutation rate than the cell-division dependent mutagenesis.

DNA replication and cell division have long been assumed to be major sources of somatic mutations. Nevertheless, our study demonstrates that cell-division independent mutagenesis predominates in tropical trees. This finding concords with a recent report on somatic mutations in human, which showed that dominant mutational processes can occur independently of cell division[32]. The preponderance of non-replicative mutational process can be attributed to its distinct molecular origin, specifically the accumulation of spontaneous transitions at CpG sites over time. The dominance of endogenous mutational process leads to a molecular clock, a constant rate of molecular evolution[33,34].

In trees, somatic mutations can be transmitted to numerous seeds, resulting in a rich genetic variation that can facilitate the effectiveness of natural selection to subsequent generations. The high rate of somatic mutations observed in long-lived tropical trees suggests that these mutations may play a more significant role in plant evolution than previously thought. 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 the conservation of long-lived trees is crucial for preserving their evolutionary potential and sustaining ecosystem function.

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. *et al*. Genetic mosaicism in plants and clonal animals. *Annu. Rev. Ecol. Evol. Syst.* **26**, 423–444 (2009).
3. Schoen, D. J. & Schultz, S. T. Somatic mutation and evolution in plants. *Annu. Rev. Ecol. Evol. Syst.* **50**, 49–73 (2019).
4. Schmid-Siegert, E. *et al.* Low number of fixed somatic mutations in a long-lived oak tree. *Nat. Plants* **3**, 926–929 (2017).
5. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
6. Plomion, C. *et al.* Oak genome reveals facets of long lifespan. *Nat. Plants* **4**, 440–452 (2018).
7. Wang, L. *et al.* The architecture of intra-organism mutation rate variation in plants. *PLoS Biol.* **17**, 1–29 (2019).
8. Orr, A. J. *et al.* A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. *Proc. R. Soc. B Biol. Sci.* **287**, (2020).
9. 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).

10. 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).
11. 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).
12. Cagan, A. *et al.* Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
13. Larsson, N. G. Somatic mitochondrial DNA mutations in mammalian aging. *Annu. Rev. Biochem.* **79**, 683–706 (2010).
14. Klekowski, E. J. Plant clonality, mutation, diplontic selection and mutational meltdown. *Biol. J. Linn. Soc.* **79**, 61–67 (2003).
15. Lynch, M., Butcher, D., Bürger, R. & Gabriel, W. The mutational meltdown in asexual populations. *J. Hered.* **84**, (1993).
16. Rohde, K. Latitudinal gradients in species diversity : The search for the primary cause. *Oikos* **65**, 514–527 (1992).
17. Johnson, S. E. & Abrams, M. D. Age class, longevity and growth rate relationships: Protracted growth increases in old trees in the eastern United States. *Tree Physiol.* **29**, 1317–1328 (2009).
18. Black, B. A., Colbert, J. J. & Pederson, N. Relationships between radial growth rates and lifespan within North American tree species. *Ecoscience* **15**, 349–357 (2008).
19. Ghazoul, J. *Dipterocarp Biology, Ecology, and Conservation*. (Oxford University Press, 2016).
20. Widiyatno *et al.* Early performance of 23 dipterocarp species planted in logged-over rainforest. *J. Trop. For. Sci.* **26**, 259–266 (2014).
21. Usami, K. Tropical woods as pulp stuffs. *J. Agric. Res. Q.* **12**, 109–114 (1978).
22. Praptoyo, H. & Mayaningsih, R. Anatomical features of wood from some fast growing red meranti. *Proceeding 4th Int. Symp. IWoRs* **7**, 8 (2012).
23. Coloundre, C. *et al.* Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**, 568–571 (1978).
24. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560–561 (1980).
25. Cooper, D. N. & Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum. Genet.* **83**, 181–188 (1989).
26. Alexandrov, L. B. *et al.* Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3**, 246–259 (2013).
27. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
28. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
29. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells Europe. *Nat Genet* **47**, 1402–1407 (2015).

30. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
31. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).
32. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* **593,** 405-410 (2021).
33. Zuckerkandl, E. & Pauling, L. *Evolving Genes and Proteins*. (Academic Press, 1965).
34. Kimura, M. & Ohta, T. On the rate of molecular evolution. *J. Mol. Evol.* **1**, 1–17 (1971).

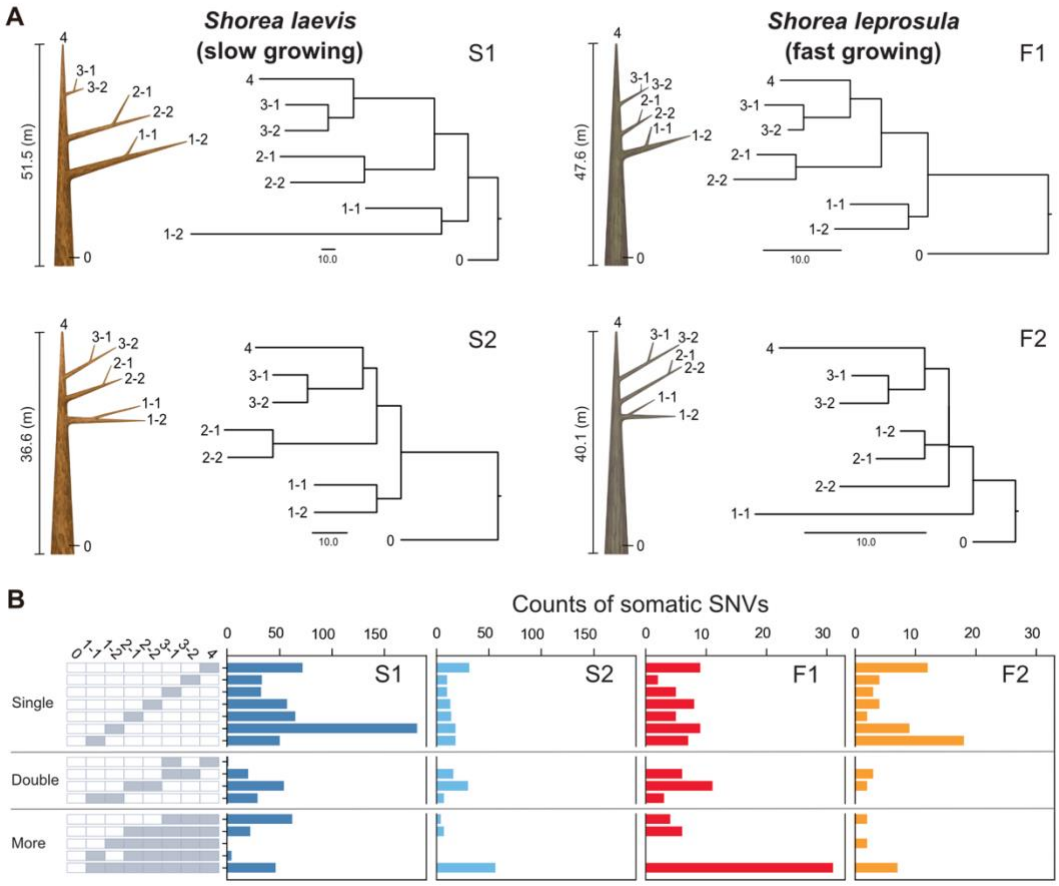


Fig. 1 | Physical tree structures and phylogenetic trees constructed from somatic mutations. a, Comparisons of physical tree structures (left, branch length in meters) and neighbor-joining (NJ) trees (right, branch length in the number of nucleotide substitutions) in two tropical tree species: *S. laevis*, 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.

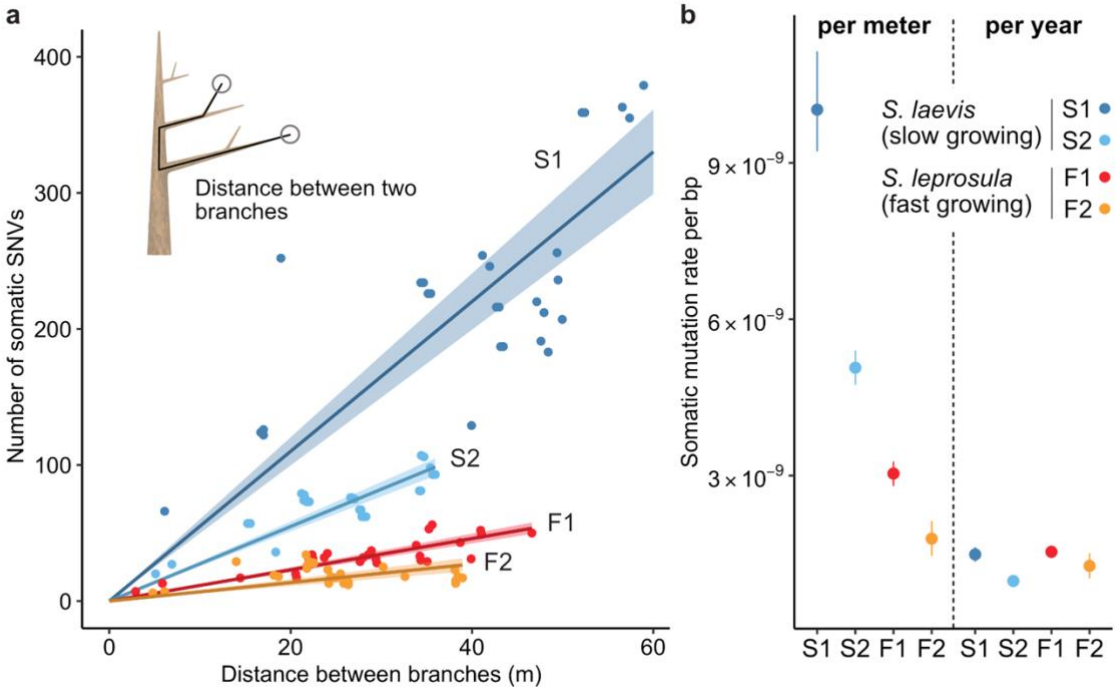


Fig. 2 | The relationship between the physical distance and the numbers of SNVs. a, Linear regression of the number of SNVs against the pair-wise distance between branches with an intercept of 0 for each tree (S1: blue, S2: right blue, F1: red, and F2: orange). Shaded areas represent 95% confidence intervals of regression lines. Regression coefficients are listed in Supplementary Table 8. **b**, Comparison of somatic mutation rates per site to grow 1 m and per year across four tropical trees. Bars indicate 95% confidence intervals.

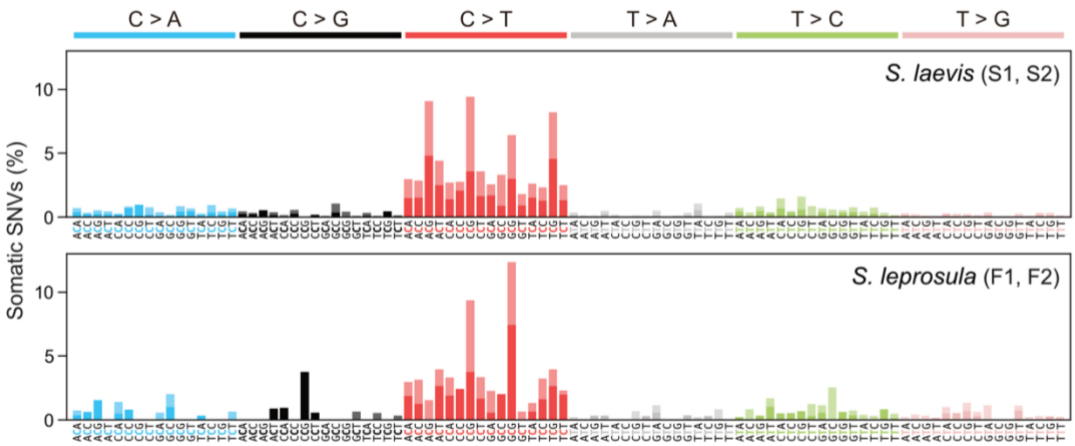


Fig. 3 | Mutational spectra of somatic SNVs. Somatic mutation spectra in *S. laevis* (upper panel) and *S. leprosula* (lower panel). The horizontal axis shows 96 mutation types on a trinucleotide context, coloured by base substitution type. Different colours within each bar indicate complementary bases. 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.

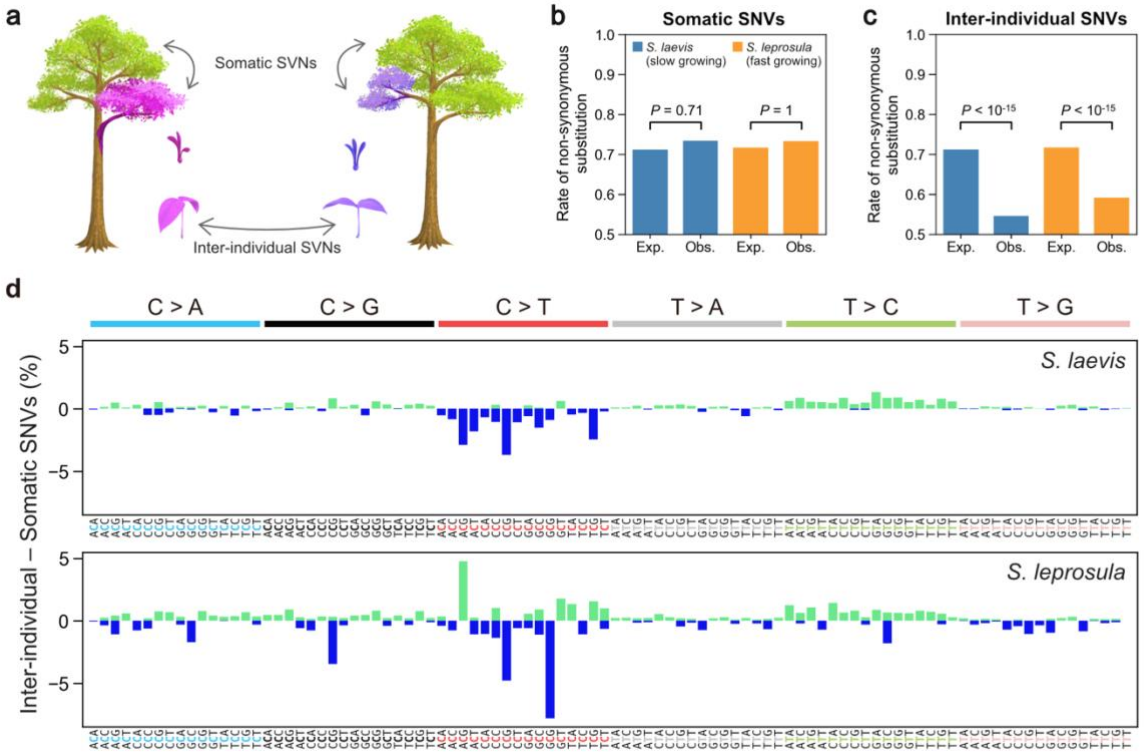


Fig. 4 | Detecting selection on somatic and inter-individual SNVs. a, An illustration of somatic and inter-individual SNVs. Different colours indicate different genotypes. **b**, Expected (Exp.) and observed (Obs.) rates of somatic non-synonymous substitutions. **c**, Expected (Exp.) and observed (Obs.) 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.

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; Extended Data 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[20].

The study focuses on two native Dipterocarpaceae species, *S. laevis* and *S. leprosula* (Extended Data Fig. 1a). We logged two individuals from each species (S1 and S2 for *S. laevis* and F1 and F2 for *S leprosula*; 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 (Extended Data Fig. 2). To calculate the physical distance between sampling positions within the tree architecture, we measured the length of each branch (Supplementary Table 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 Table 1).

DNA extraction

For short-read sequencing, DNA extraction was performed using a modified version of the method described previously[35] 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)[36]. 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[37].

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[38]. 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[39].

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 F2 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[40] with 12 threads and with an estimated genome size of 350 Mbp. We subsequently used HyPo v1.0.3[41] for polishing the contigs. The Illumina read alignments provided to HyPo were created using Bowtie v2.3.4.3[42]

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[43], 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[44] for each species (Supplementary Table 4).

Genome annotation

We constructed repeat libraries of *S. laevis* and *S. leprosula* using EDTA v2.0.0[45]. Using the libraries, we ran RepeatMasker 4.1.2-p1[46] 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 4).

We ran BRAKER 2.1.6[47] 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[48], 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[49] as the training data.

The two sets of gene predictions were merged using TSEBRA (commit `0e6c9bf` in the GitHub repository)[50] 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[51] with default parameters for functional annotation. The databases employed were: UniProtKB release 2022_05[52], NCBI RefSeq plant proteins release 215[53], EnTAP Binary Database v0.10.8[51] and EggNOG 4.1[54]. 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[55]. We counted k-mer from forward sequence data of branch 1-1 from the S1 and F1 individuals using jellyfish[56] (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[57], respectively. The genome size

of the fast-growing species was nearly identical to that previously reported for *S. leprosula* in peninsular Malaysia[39].

Somatic (intra-individual) SNV discovery

We filtered low quality reads out and trimmed adapters using fastp v22.0[58] with default settings. The cleaned reads were mapped to the reference genome using bwa-mem2 22.1[59] with default parameters. We removed PCR duplicates using fixmate and markdup function of samtools 1.13[60]. 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[60,61] and GATK HaplotypeCaller[62]) and extracted SNVs detected by both (Extended Data Fig. 2).

We first called SNVs with Bcftools 1.13[63] 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[63].

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 (Extended Data Fig. 2) using Tassel5[64] 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 Table 6.

The candidate SNV calls were manually confirmed by two independent researchers using the IGV browser[65]. 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, or if the percentage of high-quality reads was less than 50% of the total mapped reads in at least one branch sample. We compared the branch-wise pattern with the genotyping call after labeling branches carrying the called variant as somatic mutation. If the alternative allele was present in all 16 samples, we removed the site from the list as a fixed site rather than a variant. If the observed pattern did not match the genotyping call, we disregarded the site from the list of candidates as a variant call error if the following two conditions were satisfied: (1) the difference between the observed pattern and the genotyping call was supported by more than one read in either of the two biological replicates or by only one read in both biological replicates, (2) the alternative (or reference) allele was not present in one of the two biological replicates but the presence of the allele in another biological replicate was

supported by only one read. If IGV detected another allele and the sample carried more than one read of that allele, we deemed them genotyping errors and disregarded the site from the list of candidates. The final set of SNVs can be found in Supplementary Table 11. Proportion of false positive and undetected SNVs for each threshold are illustrated in Extended Data Fig. 4.

The NJ tree for each individual was generated using MEGA11[66] 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 Table 11). 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. 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.

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[63] was used. We extracted SNVs that are present in all branches within an individual using Tassel5[64]. To exclude ambiguous SNV calls, we removed SNVs within 151 bp of indels that were called with Bcftools 1.13[63] 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[67]. 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[68]) 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[59] with default parameters. Using Bcftools 1.13 mpileup[63], 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[65]. 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 Table 7).

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[69] 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[26–28], 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 $3L_{cds}$ (Extended Data Fig. 5). We classified all possible mutations into three types based on the codon table: synonymous, missense, and nonsense (Extended Data Fig. 5). Each type of mutation was counted for each of the six base substitution classes (Extended Data Fig. 5). We generated count tables based on two distinct categories of CDS: those that included all isoforms and those that only encompassed primary isoforms (Supplementary Table 12). 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) (Extended Data Fig. 5). As a background mutation rate, we adopted the observed somatic mutation rates in the six substitution classes in the intergenic region (Supplementary Table 13), 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 10). We used the package binom.test in R v3.6.2.

We also used the observed somatic mutation rate in the whole genome (Supplementary Tables 13), including genic and intergenic regions, as the background mutation rate and confirmed the robustness of our conclusion (Supplementary Tables 10). 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 Tables 13). While cancer genomics studies have accounted for more detailed context-dependent mutations, such as the high rate of C>T at CpG dinucleotides[70] or comprehensive analysis of 96 possible substitution classes in triplet context[71], 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 Table 14.

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, and VBC for providing the NGS facility. We also thank Y. Iwasa, H. Tachida, and M. Nordborg for their 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., I.R., A.M.M., V.M., and Y.S. performed molecular experiments; I.R., 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.).

Methods references

35. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11–15 (1987).
36. Toyama, H. *et al.* Effects of logging and recruitment on community phylogenetic structure in 32 permanent forest plots of Kampong Thom, Cambodia. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 1–13 (2015).
37. Doyle, J. *DNA Protocols for Plants in Molecular Techniques in Taxonomy*. (Springer, 1991).
38. 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).
39. 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).

40. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, (2019).
41. Kundu, R., Casey, J. & Sung, W.-K. HyPo: Super cast & accurate polisher for long read genome assemblies. *bioRxiv* (2019).
42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
43. 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**, 170–175 (2021).
44. 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**, 4647–4654 (2021).
45. Vurture, G. W. *et al.* GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
46. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
47. Ng, C. H. *et al.* Genome size variation and evolution in Dipterocarpaceae. *Plant Ecol. Divers.* **9**, 437–446 (2016).
48. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
49. Zdobnov, E. M. *et al.* OrthoDB in 2020: Evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **49**, D389–D393 (2021).
50. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. *http://www.repeatmasker.org*
51. 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 Genomics Bioinforma.* **3**, lqaa108 (2021).
52. 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**, 907–915 (2019).
53. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
54. Hart, A. J. *et al.* EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol. Ecol. Resour.* **20**, 591–604 (2020).
55. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
56. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–745 (2016).
57. Powell, S. *et al.* EggNOG v4.0: Nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42**, D231–D239 (2014).

644 58. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor.
645 *Bioinformatics* **34**, i884–i890 (2018).
646 59. Vasimuddin, M., Sanchit, M., Heng, L. & Srinivas, A. Efficient architecture-aware
647 acceleration of BWA-MEM for multicore systems. *Proc. 2019 IEEE 33rd Int. Parallel*
648 *Distrib. Process. Symp. IPDPS* 314–324 (2019).
649 60. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
650 2079 (2009).
651 61. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and
652 population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–
653 2993 (2011).
654 62. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing
655 next-generation DNA sequencing data. *Genome Res.* **20**, 1297–12303 (2010).
656 63. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).
657 64. Bradbury, P. J. *et al.* TASSEL: Software for association mapping of complex traits in diverse
658 samples. *Bioinformatics* **23**, 2633–2635 (2007).
659 65. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant
660 review with the integrative genomics viewer. *Cancer Research* **77**, e31–e34 (2017).
661 66. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular evolutionary genetics analysis
662 version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).
663 67. Li, H. & Wren, J. Toward better understanding of artifacts in variant calling from high-
664 coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
665 68. Suyama, Y. *et al.* Complementary combination of multiplex high-throughput DNA
666 sequencing for molecular phylogeny. *Ecol. Res.* **37**, 171–181 (2022).
667 69. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for
668 FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).
669 70. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis
670 of pathogenicity of somatic mutations in cancer. *Genetics* **173**, 2187–2198 (2006).
671 71. Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**,
672 1029-1041.e21 (2017).
673

674 **Competing interest**

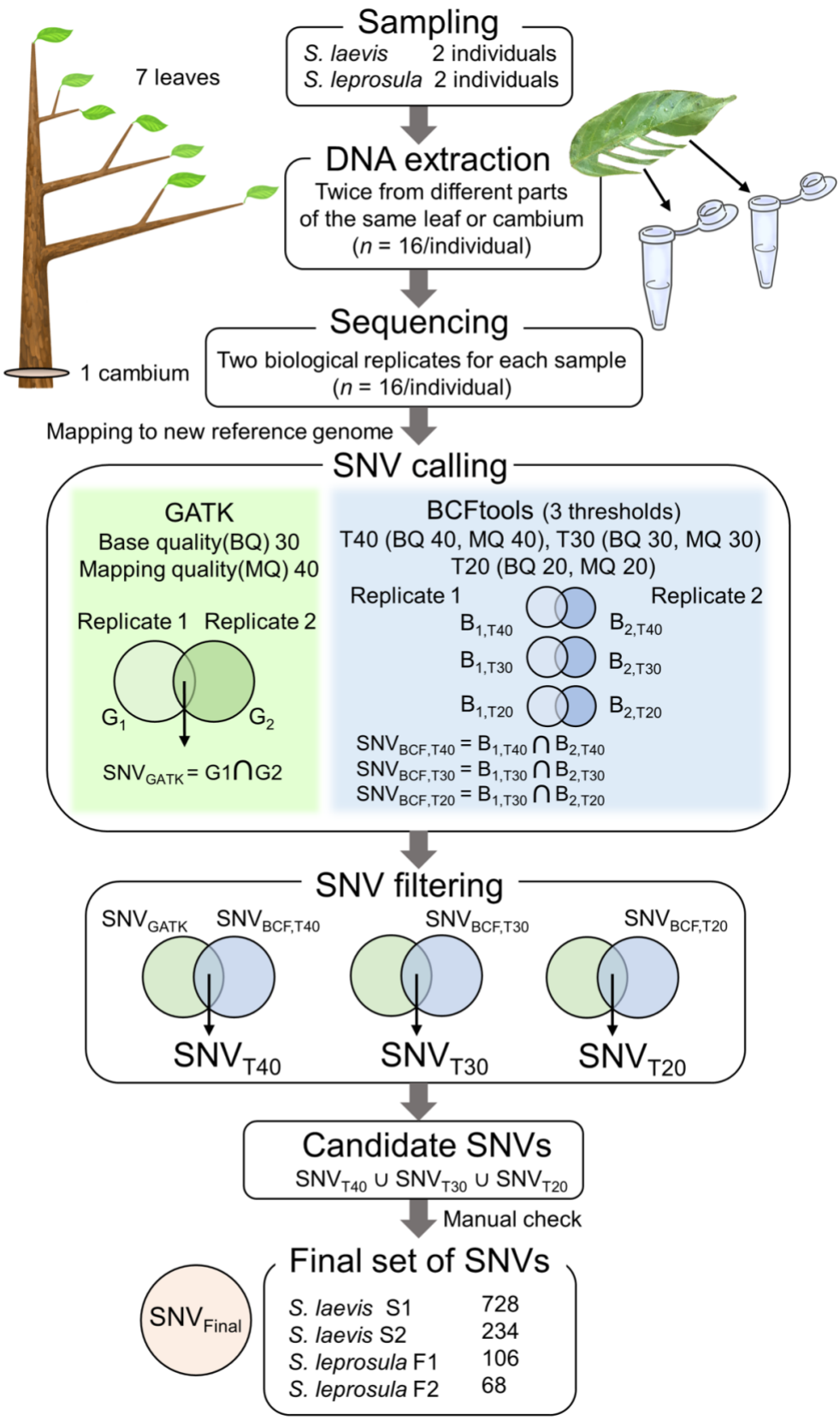
675 The authors declare that they have no competing financial and non-financial interests.
676

677

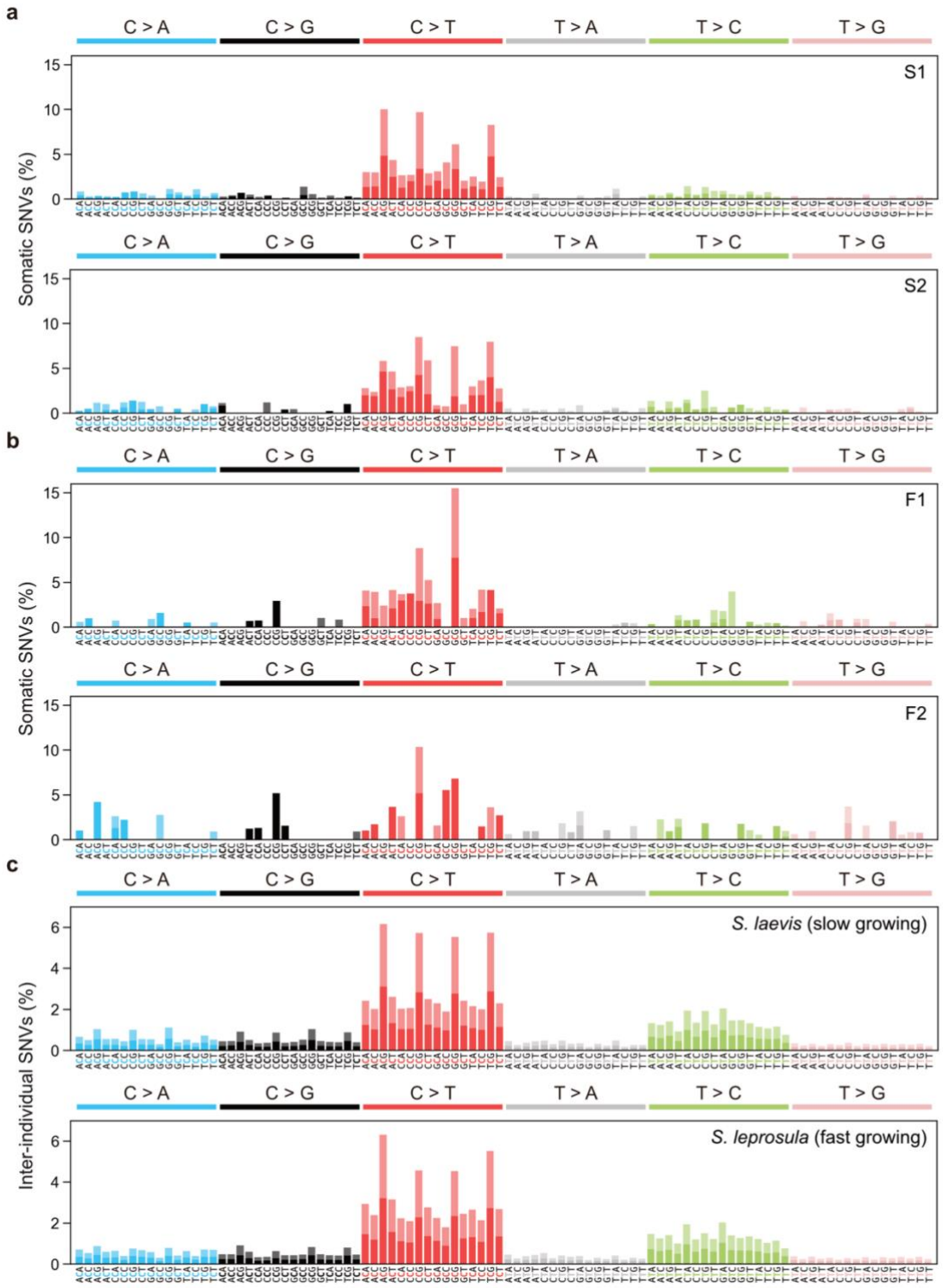
678
679

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

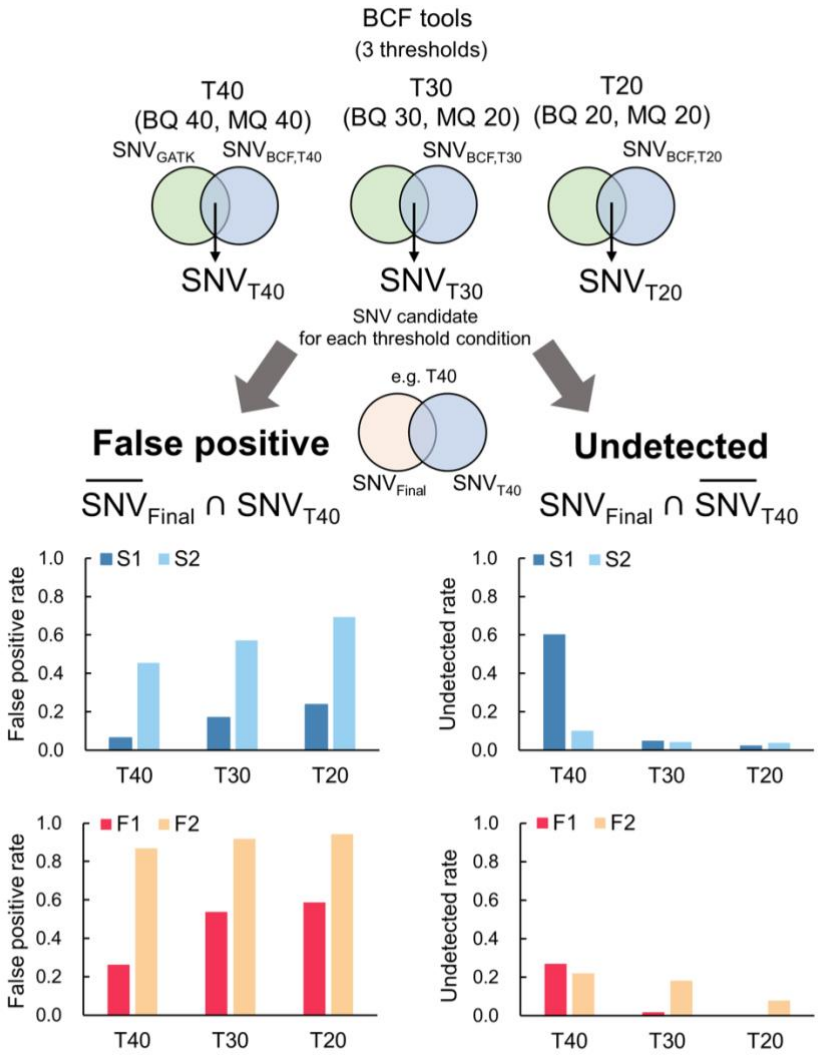
683



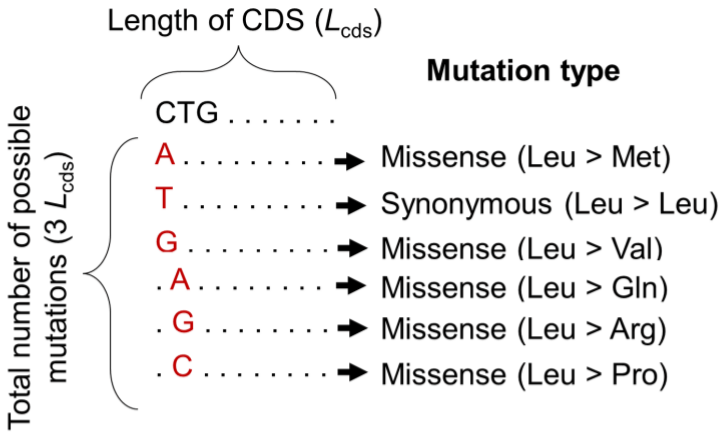
Extended Data Fig. 2 | Workflow for identifying *de novo* somatic SNVs. 8 samples (seven leaves and one cambium) were collected from four trees (two trees from each species). DNA was extracted twice independently from each sample and sequenced independently. Reads were mapped to the reference genome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



Extended Data Fig. 3 | Mutational spectra of somatic and inter-individual substitutions in each tree. 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.



Extended Data Fig. 4 | Proportion of false positive and undetected SNVs. Proportion of false positive SNVs were 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. Proportion of undetected SNVs were calculated by extracting the subset of potential SNVs that were included in the final set but not included in the potential SNVs for each threshold, and by dividing 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

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

Extended Data Fig. 5 | 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.