Supplementary Information for

Genomic Insights into Adaptation to Karst Limestone and Incipient Speciation in East Asian *Platycarya* spp. (Juglandaceae)

Yu Cao[1,2,3], Fabricio Almeida-Silva[2,3], Wei-Ping Zhang[1], Ya-Mei Ding[1], Dan Bai[1], Wei-Ning Bai[1*], Bo-Wen Zhang[1*], Yves Van de Peer[2,3,4,5*] and Da-Yong Zhang[1*]

[1]State Key Laboratory of Earth Surface Process and Resource Ecology and Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing, China.
[2]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium.
[3]Center for Plant Systems Biology, VIB, Ghent, Belgium.
[4]Center for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa.
[5]College of Horticulture, Nanjing Agricultural University, Nanjing, China.

***Corresponding authors**: baiwn@bnu.edu.cn, zhangbw@mail.bnu.edu.cn, yves.vandepeer@psb.ugent.be, zhangdy@bnu.edu.cn

This PDF file includes:

supplementary notes 1-11
supplementary figures 1 to 8

Additional supplementary information for this manuscript include:

supplementary table (an Excel file, with 18 single sheets in it).

Supplementary notes

supplementary note1: Genome assembly and annotation

A total of 167 Gb of subreads (240 × depth) from the PacBio Sequel platform and 70.83 Gb of short reads (102 × depth) from the Illumina NovaSeq 6000 were generated for *P. longipes*, and 129.55 Gb of subreads (184 × depth) from the PacBio Sequel platform and 90.04 Gb of short reads (128 × depth) from the Illumina NovaSeq 6000 were generated for *P. strobilacea* (supplementary table S1). The genome was initially de novo assembled and then polished by four rounds of Illumina short reads. The improved contigs were further assembled into scaffolds with a scaffold N50 of approximately 45 Mb for both species. Moreover, a total of 75.8 Gb and 62.6 Gb of Hi-C data were generated using the Illumina platform for *P. longipes* and *P. strobilacea,* respectively.

A total of 29,525 protein-coding genes were predicted in *P. longipes,* with an average CDS length, exon length and exon number of 1005.79, 232.77, and 4.32, respectively. For *P. strobilacea*, 29,330 protein-coding genes were predicted with an average CDS length, exon length and exon number of 948.63, 235.4, and 4.03, respectively, similar to reported parameters for other Fagales species (supplementary table S2).

For repeat annotation, approximately 47.99% of the *P. longipes* genome assembly and 44.54% of the *P. strobilacea* genome assembly were identified as repetitive elements based on de novo and homology-based methods (Fig. 1b and supplementary table S3). As in most plant genomes, long terminal repeat retrotransposons (LTR-RTs), accounting for 34.73% and 35.94% of the genome, are the most abundant elements. Among the LTR-RTs, Copia and Gypsy were the most common superfamilies, representing 19.44% and 12.66% of the *P. longipes* assembly and 23.44% and 10.99% of the *P. strobilacea* assembly, respectively (supplementary table S3). We further detected 511 and 509 microRNA (miRNA), 575 and 556 transfer RNA (tRNA), 279 and 2019 ribosomal RNA (rRNA), and 692 and 624 small nuclear RNA (snRNA) genes in the *P. longipes* and *P. strobilacea* genome sequences, respectively (supplementary table S4).

For function annotation, we annotated the protein-coding genes against the SwissProt (http://www.uniprot.org/), Nr (http://www.ncbi.nlm.nih.gov/protein), KEGG (http://www.genome.jp/kegg/), Pfam (http://pfam.xfam.org/) and InterPro (https://www.ebi.ac.uk/interpro/) databases (supplementary table S5).

supplementary note2: Sampling collection for whole genome resequencing

Fresh leaves of *P. strobilacea* (ZhangPu County, FuJian Province, China, 24°17′55.9″N, 117°56′45.78″E) and *P. longipes* (WangMo County, GuiZhou Province, China, 25°11′48.77″N, 106°8′31.3″E) were collected for extracting and sequencing genomic DNA. A permanent voucher for each species has been deposited in the BNU herbarium (*Cao* BNU0056917 and *Zhang* BNU0056918).

The DNA was extracted from green leaves using a routing protocol. The whole genome resequencing using paired-end libraries with an insert size of 350 bp was performed on Illumina HiSeq X-ten instruments by NovoGene (Beijing, China), with an average depth of approximately 30× for each sample.

supplementary note3: Genome assembly

The karyotype analysis revealed that both *P. strobilacea* and *P. longipes* have a chromosome base of 15 (2N= 30). Hi-C library was prepared and anchored to 15 chromosomes following the standard procedure (Lieberman-Aiden et al. 2009). The genome size was estimated through the analysis of 150 bp paired-end reads, computation of 17 bp K-mer frequencies using Jellyfish v2.3 (Marcais and Kingsford 2011), then the resulting histogram was exported into findGSE (Sun et al. 2018). The PacBio single-molecule long reads were then assembled, and BUSCO (Simao et al. 2015) (http://busco.ezlab.org/) with a plant database of 1,440 conserved plant genes was used to estimate the completeness of the assembly.

supplementary note4: Quality control for genome-wide SNP

To control the quality of genome-wide SNPs, sites with a mapping depth of less than a third or more than double an average depth of an individual, non-biallelic sites, and sites with missing data were removed. For heterozygous genotype calling, if the proportion of an alternative allele was between 20% and 80% when the depth of an allele was >20×, or if the proportion of an alternative allele was between 10% and 90% when the depth of an allele was >10×, the heterozygous genotype calling remained; otherwise, a homozygous genotype was called (Nielsen et al. 2011). Individuals with close relationships (more closely related than 3rd-degree relationships) were excluded

according to the kinship estimation of the KING program (Manichaikul et al. 2010) (supplementary fig. S8), resulting in a final sample size of 130 unrelated individuals.

supplementary note5: Population structure analysis

In the population genetic structure analysis, the genetic Clusters *K* were predefined from 1 to 6, and each assignment was performed 20 times to ensure a stable result. The Markov chain Monte Carlo analyses were run for 50,000 iterations after a burn-in period of 20,000 iterations.

supplementary note6: Data preparation and model estimation with IMa3

To prepare the noncoding sequence for IMa3 analysis, we mapped the reads of each individual to *P. strobilacea* reference genomes using the BWA-MEM algorithm from BWA v. 0.7.12 (Li and Durbin 2009). We then performed variant calling using SAMTOOLS v.1.19 (Li 2011), and SNPs were filtered with the quality adjuster -C setting at 50. Both the minimal base quality -Q and mapping quality of alignments -q were set to 20. Indels or any SNPs within 3 bp around a gap were removed. The filtration criteria of mapping depth and calibration of heterozygous sites were the same as the above-mentioned. The consensus genome of each individual was built based on the SNPs. After masking sites located in or near approximately 25 kb of the coding sequences, we extracted 300-1000 bp noncoding loci, which was at least 25 kb apart from each.

To perform the "Isolation with Migration" Bayesian framework of IMa3, the HKY mutation model (Hasegawa et al. 1985) was employed, and the migration and divergence time parameters were set as -q6, -m1 and -t4, respectively. Isolation with migration analysis was performed using Markov chain Monte Carlo sampling with 400 chains distributed across 80 processors and a geometric chain heating scheme with first and second heating parameters of 0.995 and 0.4, respectively. The program was run for 48 h following a 24-h burn-in to ensure effective sample sizes exceeding 200.

supplementary note7: Identification of Genomic Islands of Divergence

To test the extent to which historical demographic events can explain the observed patterns of genetic divergence between the two species, ms (Hudson 2002) was used to compare the observed patterns of differentiation to those expected under the demographic model inferred by *IMa3* (fig.

2e, supplementary table S7). $\rho = 4N_e c$ was assumed under exponential distribution with the mean $\rho = 60$ (sd= 62) per 25 kb by LDhat v2.2 (McVean et al. 2004; Auton and McVean 2007). A total of 500,000 replications of genotypes corresponding to a 25 kb region were simulated, and the output of ms was transformed into a VCF file using the script (https://github.com/Flavia95/Thesis (last visit at Feb. 9, 2022).

supplementary note8: Population genomic statistics in stepping windows

Population genomic statistics were calculated in non-overlapping 25 Kbp windows. π and Tajima's *D* were calculated using VCFtools v0.1.17 (Danecek et al. 2011). The *CLR* statistic was calculated using SWEEPFINDER2 (DeGiorgio et al. 2016) with a precomputed empirical spectrum and recombination map for *P. strobilacea* and *P. longipes*, respectively. *XP-EHH* was calculated using R package REHH v.3.2.2 (Gautier et al. 2017) to find a selective sweep region in *P. longipes*, assuming *P. strobilacea* as a neutral population. π and *CLR* statistics were computed based on the information given by *P. strobilacea* (Innan and Kim 2008).

supplementary note9: Decorrelated composite of multiple signals method

DCMS is a composite method for the detection of selection that combines molecular signals of different tests and considers potential correlations among the different tests. For each statistic, we tested whether its distribution fit the normal distribution using the R package Cmplot v4.0.0 (Yin 2022). If not, we performed a two-step normalization approach (Templeton 2011). In the first step, the variable is transformed into a percentile rank, which results in uniformly distributed probabilities. In the second step, the inverse-normal transformation is applied to the percentile ranks to form a variable consisting of normally distributed *Z score*s. Normalized scores for each statistic were *Z*- transformed, and a *p* was derived from this transformation.

supplementary note10: Sampling collection for transcriptome sequencing

For transcriptome sequencing, we collected roots, stems, and leaves of 36 samples of each species, and performed RNA-seq for these samples with high calcium-magnesium treatment for 0h (baseline without treatment), 6h, 1d, and 7d (fig. 6a), and also additional field samples (~5 years old; 18 leaves samples, 9 *P. longipes* vs. 9 *P. strobilacea*, 16 roots samples, 8 *P. longipes* vs. 8 *P.*

strobilacea; see supplementary table S17 for detail). For field samples, fresh tissues (roots and leaves) of *P. strobilacea* and *P. longipes* were procured from CeHeng County and WangMo County in Guizhou Province, TianLin County in Guangxi Zhuang Autonomous Region. Seeds for laboratory treatments were collected from TianLin County, 24°31′16.38″N, 106°23′30.55″E.

Following a period of chilling at 4 °C for 3 months, the seeds were put in a porous culture panel with a Hogland nutrient solution to germinate at 25 °C until the plant reached a height of approximately 10 cm (1-month-old seedlings). These 1-month-old seedlings were treated with 30 mM calcium-10 mM magnesium resolution. Leaves, roots, and stems with high calcium-magnesium treatment for 0 h, 6 h, 1 d and 7 d were collected and immediately frozen in liquid nitrogen and stored at -80 °C. Three biological replicates were obtained for each sample. The RNA was extracted and sequenced by the Illumina NovaSeq platform.

supplementary note11: Transcriptome analysis

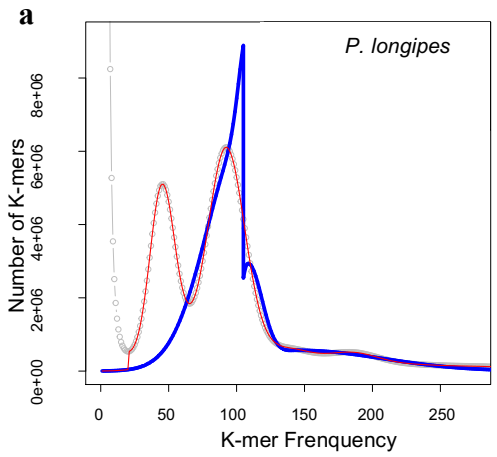
We used field samples to estimate the gene expression variation across species (*P. longipes* vs *P. strobilacea*). Differentially expressed genes (DEGs) of *P. longipes* were defined as those having at least a two-fold change in expression compared with the samples of *P. strobilacea* in the same field population (FDR < 0.05), both for root and leaf tissues. The same protocol was used to test the gene expression variation over time (6h, 1d, and 7d vs 0h) within each species.

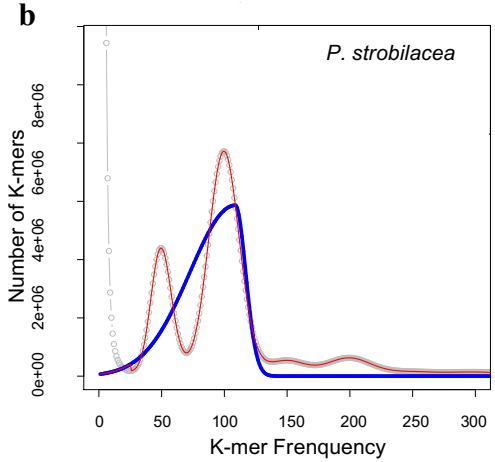
Following the methodology outlined by He et al. (2021), we specifically evaluated the gene expression variation as a result of time nested within species, which was further computed as the log2ratio of gene expression plasticity measured in each species, that is, the ratio of allele count after t h over allele count at t0 = 0 h in the first species divided by the ratio of allele count after t h over allele count at t0 = 0 h in the second species (e.g., log2 [(*P. longipes*=6h/ *P. longipes* =0h)/ (*P. strobilacea*=6h/ *P. strobilacea* =0h)]). A positive value of this ratio indicates a stronger response to stress (FDR < 0.1). A negative value of this ratio indicates either a weaker response to the stress or a reversal in the direction of the plastic change, where for example, a gene would be up-regulated in one species and down-regulated in the other species. The significance of gene expression variation was tested with a generalized linear model implemented in DESeq2 with read

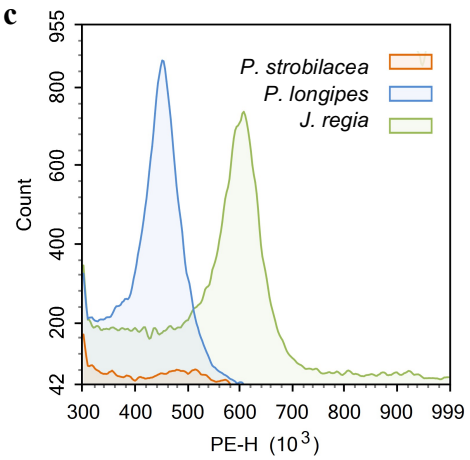
count as the dependent variable and time nested within species as the independent variable (note that DESeq2 normalizes read counts across samples).

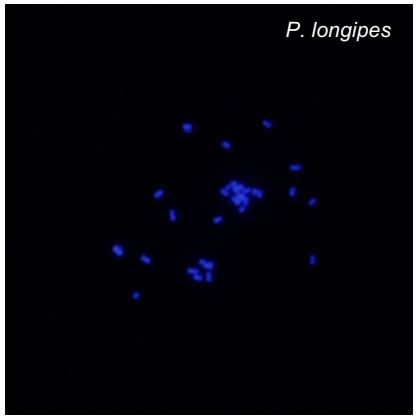
The gene co-expression network analysis was applied across the time points on the two species. We inferred a signed gene co-expression network with the R package BioNERO (Almeida-Silva and Venancio 2022) and compared the differences between the two species. For those modules that were unique in P. longipes, module enrichment analyses were performed using BioNERO. The expression levels of genes under positive selection were explored with the transcriptome data both in the field and laboratory, and differential co-expression analysis was performed with the R package diffcoexp (Wei et al. 2022) to check if the positively selected genes are differentially co-expressed between the two species. We also explored the expression profiles of nine categories of the key genes, which are reported to be related to calcium concentration regulation, at different time points under high calcium stress. It is worth mentioning that here we only selected the samples of stem tissue with the lowest heterogeneity in the hierarchical clustering results and genes that were differentially expressed at least in one period compared to a control group to display, considering the heterogeneity between biological replicates may interfere with the overall change trend.

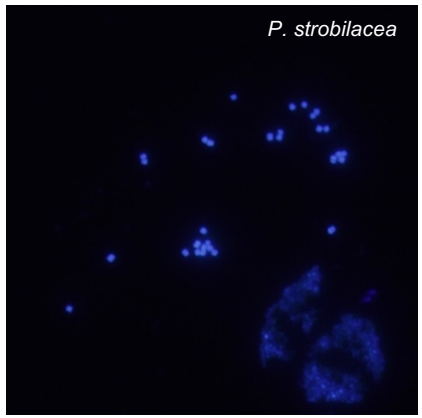
supplementary figures

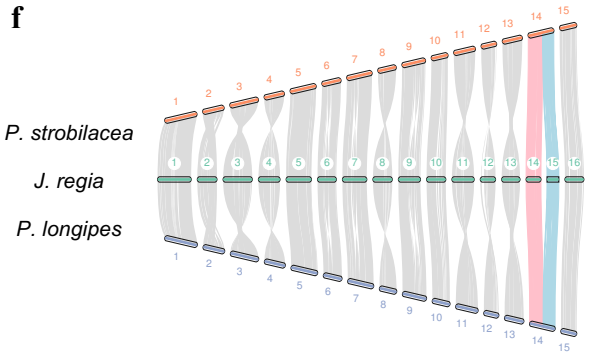




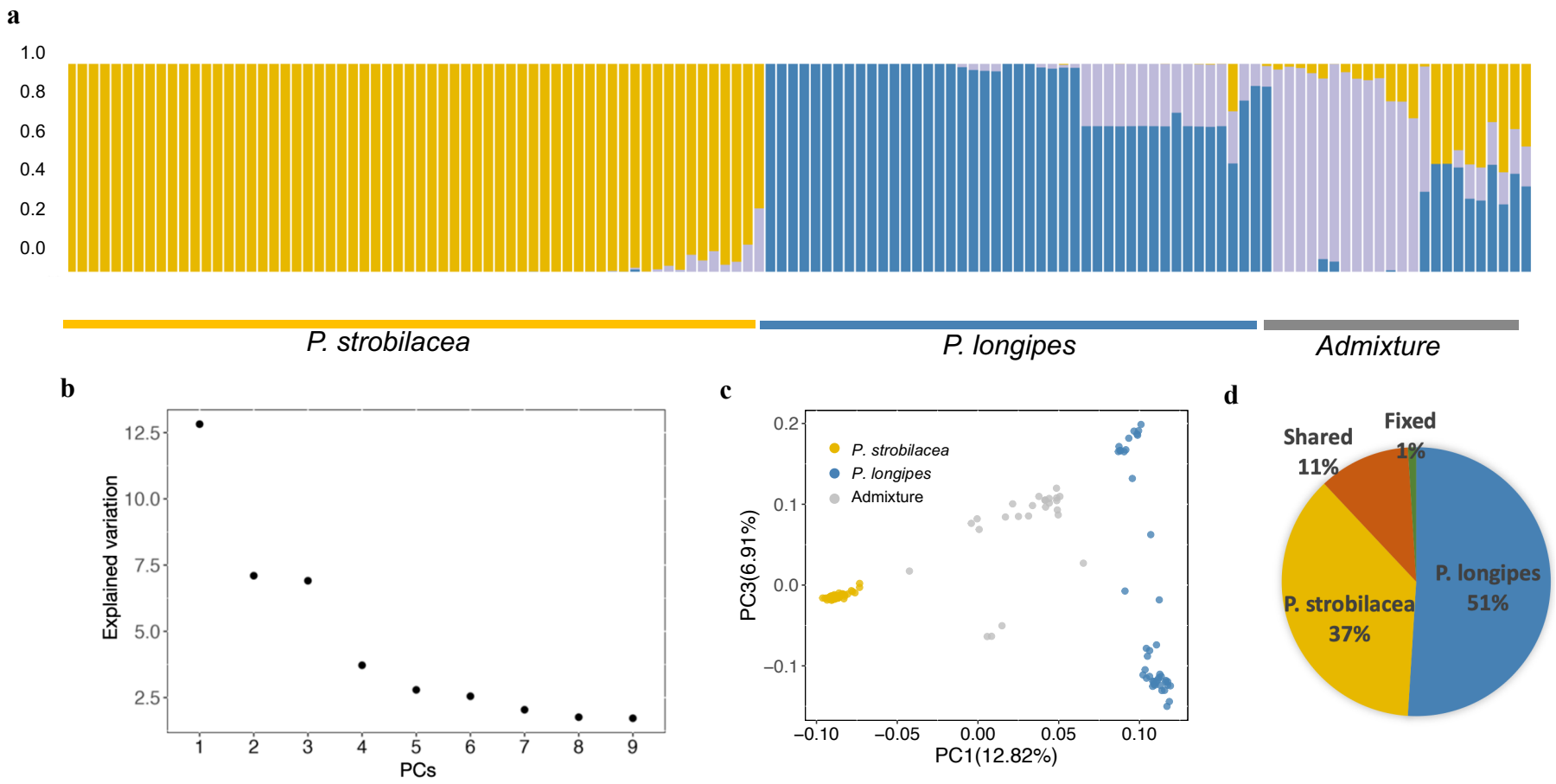




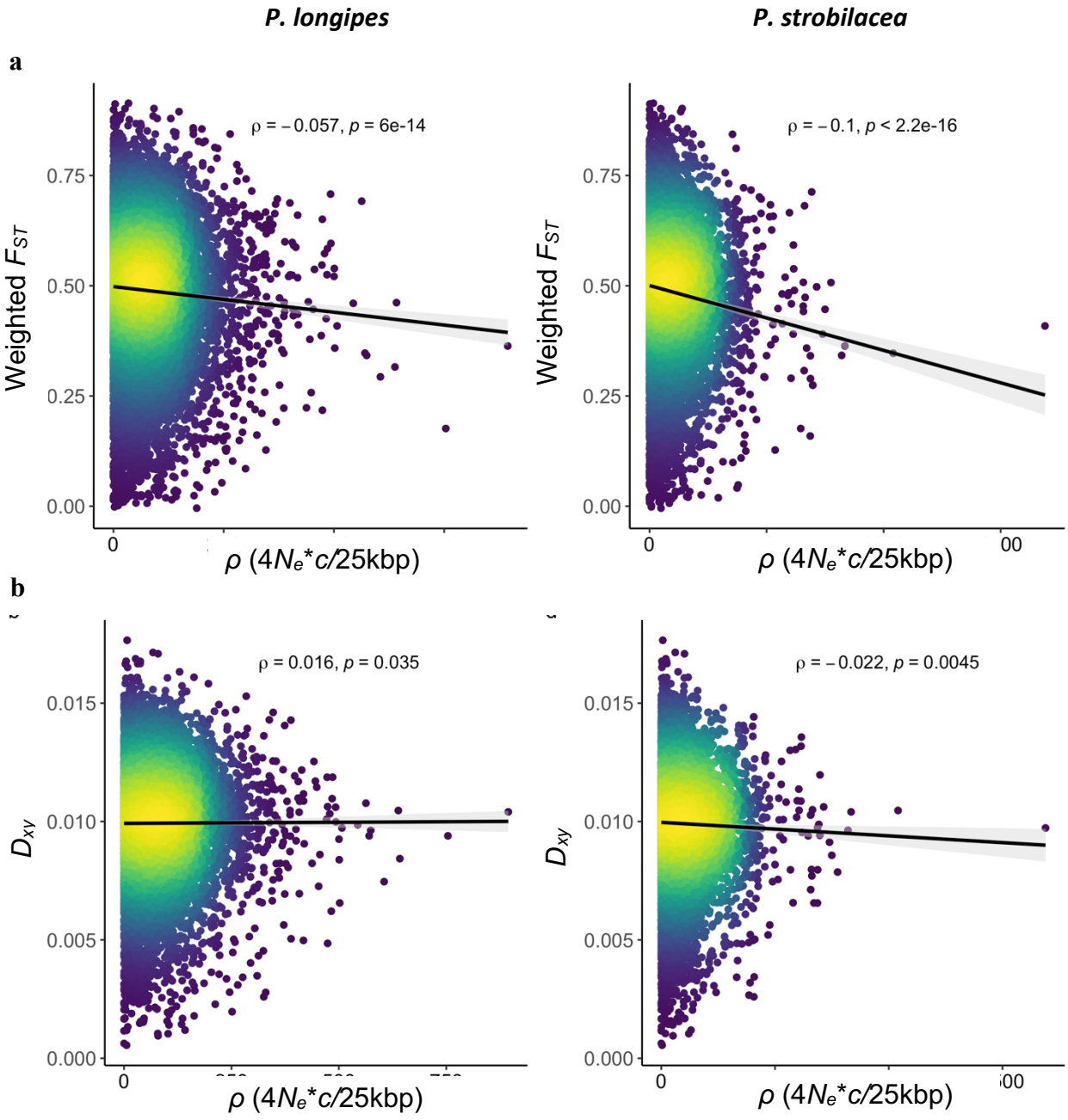




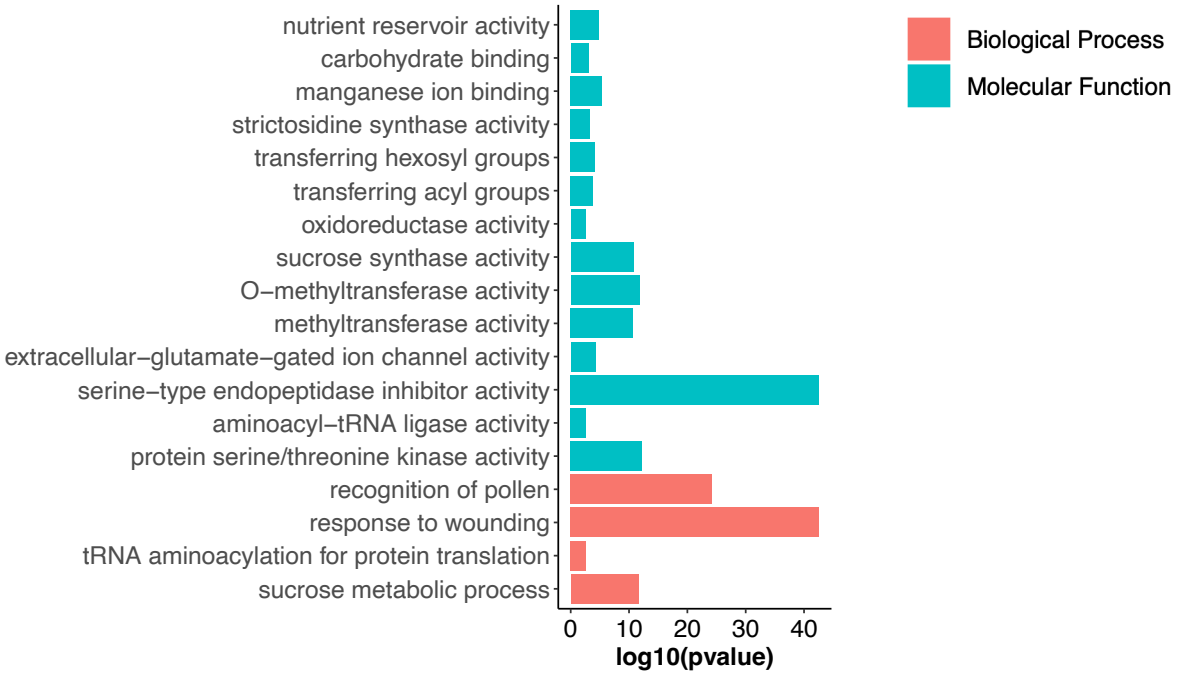
supplementary fig. S1 | Genome size and chromosome number of *Platycarya longipes* and *P. strobilacea*. Genome size estimation of **a**) *P. longipes* and **b)** *P. strobilacea.* by findGSE. K-mer size was set as 17 and the default parameters were used. Gray line is the observed K-mer coverage, blue line is the fitted count with fitted K-mer coverage, red line is the final corrected K-mer coverage. The genome size of *P. longipes* and *P. strobilacea* are 695 Mb and 703 Mb respectively. **c)** Genome size estimation using flow cytometry. The relative genome size of the genus *Platycarya* was smaller than *J. regia*, which imply that the chromosome base variation maybe occurred in *Platcarya*. FISH karyotype analysis of **d**) *P. longipes*. and **e**) *P. strobilaceae*. Both species have chromosome base N= 15 (2N=30). **f)** Inter-species macro- synteny of *P. longipes*, *P. strobilacea* and *J. regia*.



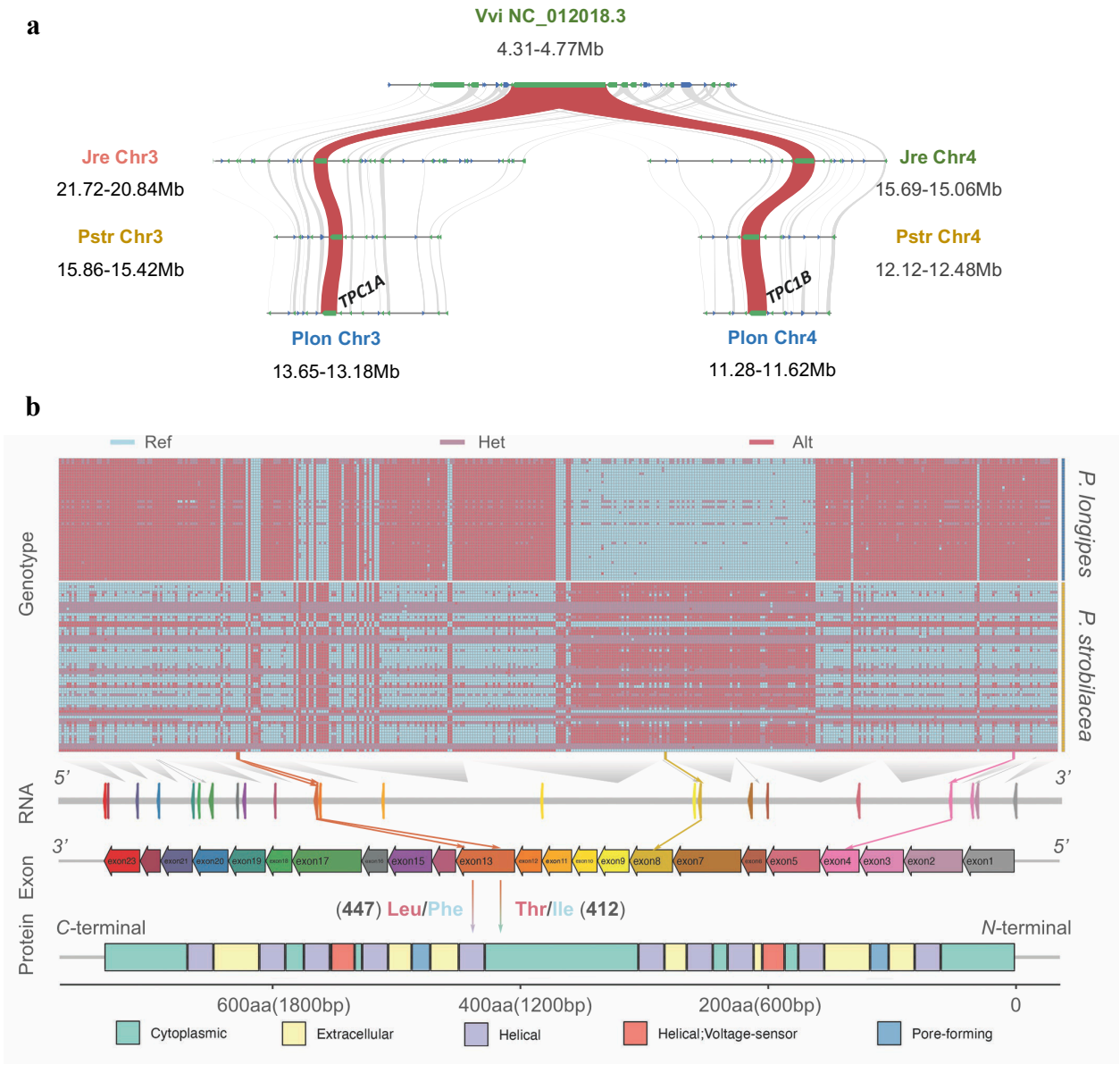
supplementary fig. S2 | Population structure. a) The population cluster in K= 3 of Structure analysis. **b)** Elbow plot of the explained variation in PCA analyses, here we show the first nine dimensions. **c)** PCA plot showing the first (PC1) and third (PC3) principal components for 130 individuals of *P. longipes* (blue dots), *P. strobilacea* (yellow dots) and the admixture (grey dots). **d)** Pie chart summarizing the proportion of fixed, shared, and exclusive polymorphisms of the two species.



supplementary fig. S3 | Relationships between population-scaled recombination rates (ρ= 4N_e*c/25kbp) and F_{ST} (a), D_{xy} (b) in both *P. longipes* (left panel) and *P. strobilacea* (right panel). Scatter plots display genome-wide values of two variables in dots over 25 Kbp non-overlapping windows. The yellow to green to blue gradient indicates decreased density of observed events at a given location in the graph.

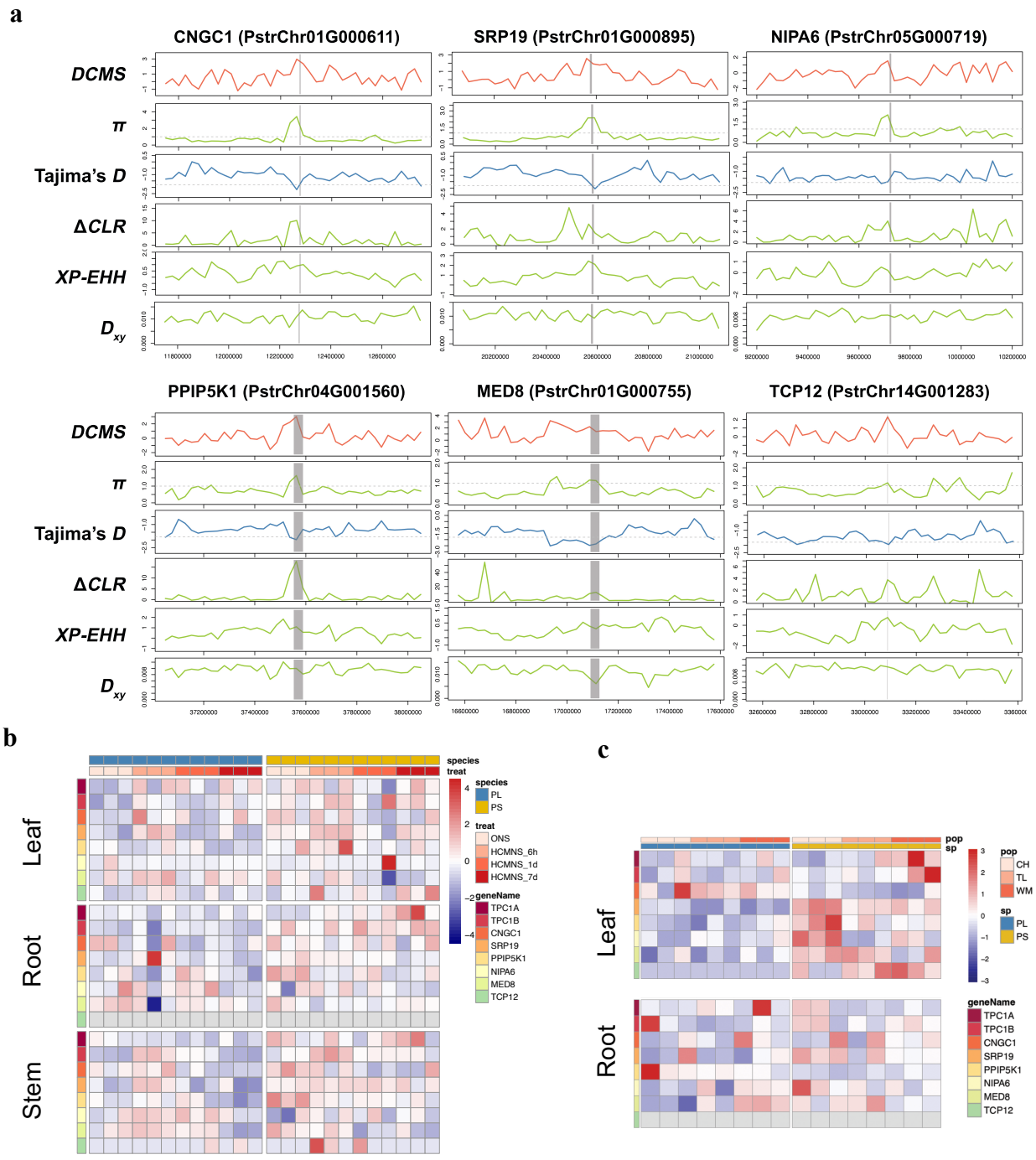


supplementary fig. S4 | Comparative genomics of the genus *Platycarya*. The GO enrichment of 305 genes in 46 expanded gene families in *P. longipes*.

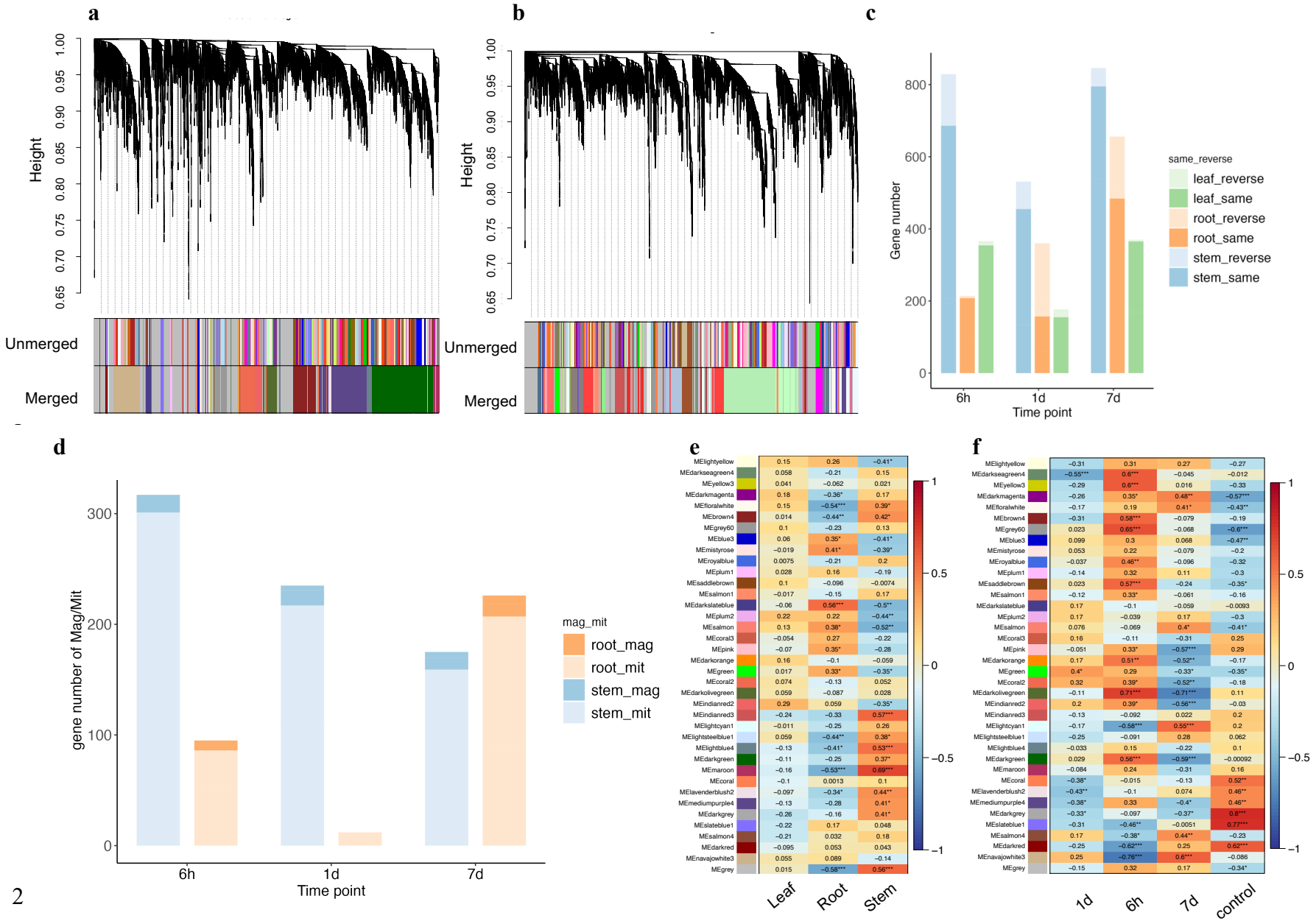


supplementary fig. S5 | The convergent adaptive signals of *TPC1*. a) Co-localization of the *TPC1* genes across the Juglandaceae and *Vitis vinifera*. *TPC1A* and *TPC1B* genes are co-localized in the genomes of *P. longipes*, *P. strobilacea* and *J. regia*: Vvi, *Vitis vinifera*; Jre, *J. regia*; Plon, *P. longipes*; Pstr, *P. strobilacea*. Blue and green rectangles indicate predicted gene models, with colour showing the gene orientation (blue, − strand; green, + strand). Orthologous gene pairs are linked by grey lines, with red and blue lines linking orthologous *TPC1*. **b**) Genotypic diversity, annotated gene structure, mRNA structure, and protein domain of *TPC1A*. Top line: Haplotype diversity of 396 highly diverged SNPs (interspecific $F_{ST} > 0.5$) within the *TPC1A* gene in *P. longipes* and *P. strobilacea*. Ref, reference allele (*P. strobilacea*); Alt, alternative allele; Het, heterozygous allele. Second line: structure

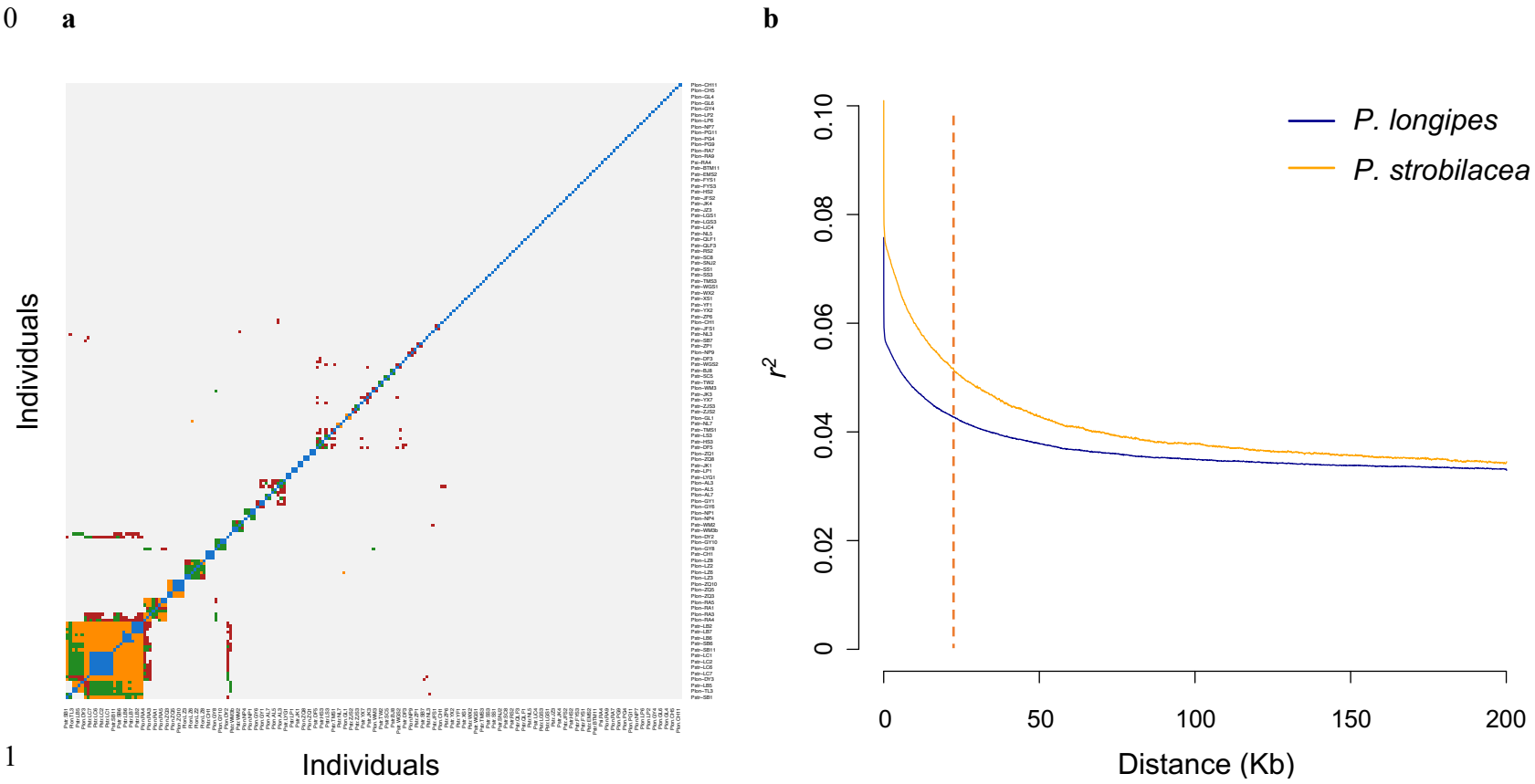
of the *TPC1A* gene in *P. longipes*. The boxes in different color represent exons, and lines between boxes represent introns. Highly diverged SNPs are marked by colorful lines (which belong to exons and may be missense or synonymous) and grey blocks (which belong to introns). Third line: mRNA structure, the transcript is composed of 23 exons. Two missense variations were located in exon 13, and two synonymous variations were located in exon 4 and exon 8. Bottom line: protein domains of *TPC1A* (refer to *Nicotiana* tabacum though). Light green, topological domain that might function in cytoplasm; light yellow, topological domain that is extracellular; purple, helical transmembrane domain; coral, helical transmembrane region acts as voltage-sensor; blue, intramembrane domain performs as the pore-forming. The two missense variations were localized in the cytoplasmic topological and helical transmembrane domain, respectively.



supplementary fig. S6 | Positive selection signal of six other genes except *TPC1A*. **a**) A zoom in of population genetics statistics and *DCMS* scores in each gene region and its upstream and downstream 500kb extension. Each data point is based on a sliding window analysis using non-overlapping 25-kb windows. The expression level of seven candidate genes under positive selection in *P. longipes* in **b**) laboratory-collected and **c**) field-collected transcriptome samples.


a
b
c
Height
Unmerged
Merged
Gene number
Time point
6h
1d
7d
same_reverse
leaf_reverse
leaf_same
root_reverse
root_same
stem_reverse
stem_same
d
e
f
gene number of Mag/Mit
mag_mit
root_mag
root_mit
stem_mag
stem_mit
Leaf
Root
Stem
control

supplementary fig. S7 | Transcriptomic analysis. Dendrograms of module assignments of gene co-expression networks (GCNs) of **a)** *P. longipes* and **b**) *P. strobilacea***. c)** Number of genes with a same direction (dark color) and reverse direction (light color) response in *P. longipes*, compared to *P. strobilacea*. **d)** Number of genes with a magnified (dark color) and mitigated (light color) response in *P. longipes*, compared to *P. strobilacea*. Different colour represents different tissue: leaf (green), root (orange) and stem (blue). Module-trait associations for preserved modules in the *P. longipes*'s GCNs **e**) across different tissues and **f**) across different time points.



supplementary fig. S8 | The data filtering details. **a)** The pairwise kinship of all 207 re-sequenced samples by KING. The color of each tile means the kinship between the individual pairs. Blue, self or mono-zygote twins; orange, parent-ofspring or full siblings; green, half siblings; red, grandchild-grandparent of first cousins; gray, no relationship. **b)** The LD decay of two species. The dashed broken lines marked distance equal 25Kbp.

References

Almeida-Silva F, Venancio TM. 2022. BioNERO: an all-in-one R/Bioconductor package for comprehensive and easy biological network reconstruction. Funct Integr Genomic 22:131-136.

Auton A, McVean G. 2007. Recombination rate estimation in the presence of hotspots. Genome research 17:1219-1227.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. 2016. SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics 32:1895-1897.

Gautier M, Klassmann A, Vitalis R. 2017. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol Ecol Resour 17:78-90.

Hasegawa M, Kishino H, Yano T-a. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174.

He F, Steige KA, Kovacova V, Gobel U, Bouzid M, Keightley PD, Beyer A, de Meaux J. 2021. Cis-regulatory evolution spotlights species differences in the adaptive potential of gene expression plasticity. Nat Commun 12:3376.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Innan H, Kim Y. 2008. Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. Genetics 179:1713-1720.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987-2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Lieberman-Aiden E, Berkum NLv, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326:289-294.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867-2873.

Marcais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764-770.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581-584.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443-451.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212.
Sun H, Ding J, Piednoel M, Schneeberger K. 2018. findGSE: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies. Bioinformatics 34:550-557.
Templeton GF. 2011. A two-step approach for transforming continuous variables to normal: implications and recommendations for IS research. Commun Assoc Inf Sys 28:4.
Wei W, Amberkar S, Hide W. 2022. Diffcoexp: differential co-expression analysis. R package version 1.14.0.
Yin L. 2022. CMplot: Circle Manhattan Plot. R package version 4.0.0.