New Phytologist Supporting Information

Article title: Genome-wide analysis of butterfly bush (*Buddleja alternifolia*) in three uplands provides insights into biogeography, demography and speciation

Authors: Yong-Peng Ma, Hafiz Muhammad Wariss, Rong-Li Liao, Ren-Gang Zhang, Quan-Zheng Yun, Richard G. Olmstead, John H. Chau, Richard I. Milne, Yves Van de Peer and Wei-Bang Sun

Article acceptance date: 19 July 2021

SUPPLEMENTARY FIGURES

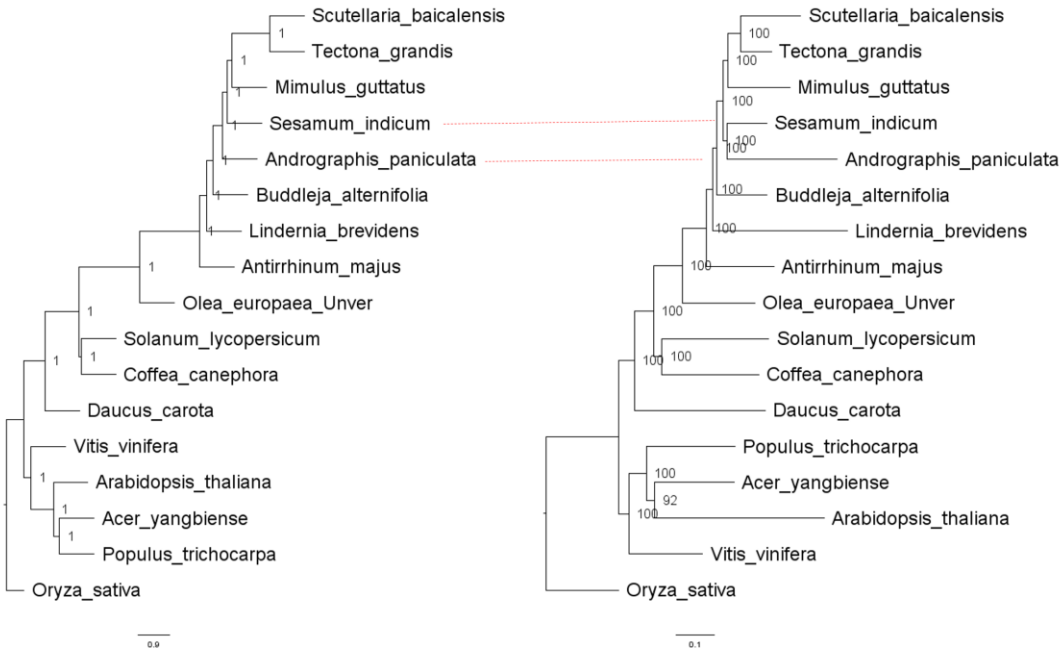

Fig. S1. Phylogenetic trees inferred by ASTRAL (left) and ML (right) based approaches, with the dot lines indicating the conflict position between *A. paniculata* and *S. indicum*, both of which clustered with a clade comprising *M. guttatus*, *T. grandis* and *S. baicalensis* in the ASTRAL tree, while formed a clade sister to the clade consisting of *M. guttatus*,*T. grandis* and *S. baicalensis* in the ML tree.

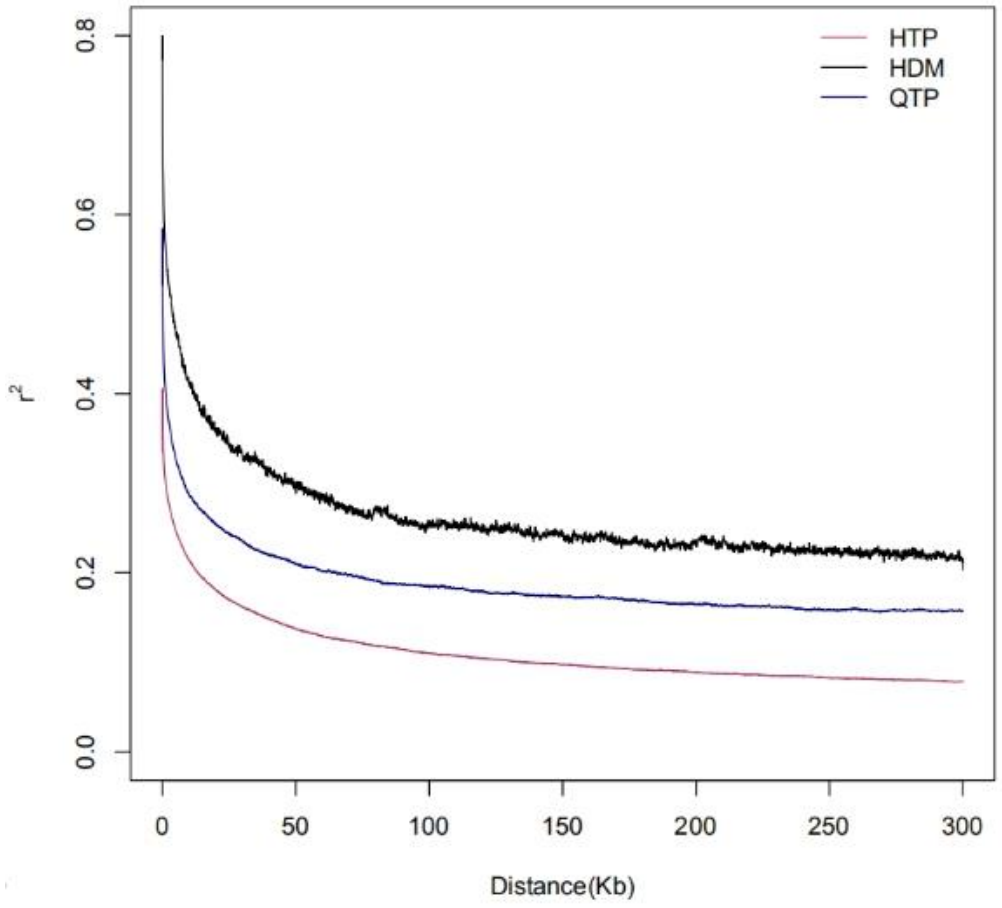

Fig. S2. Patterns of linkage disequilibrium (LD). LD decays as a function of genomic distance between polymorphisms in three distribution areas of *B. alternafolia*. LD was measured by r^2.

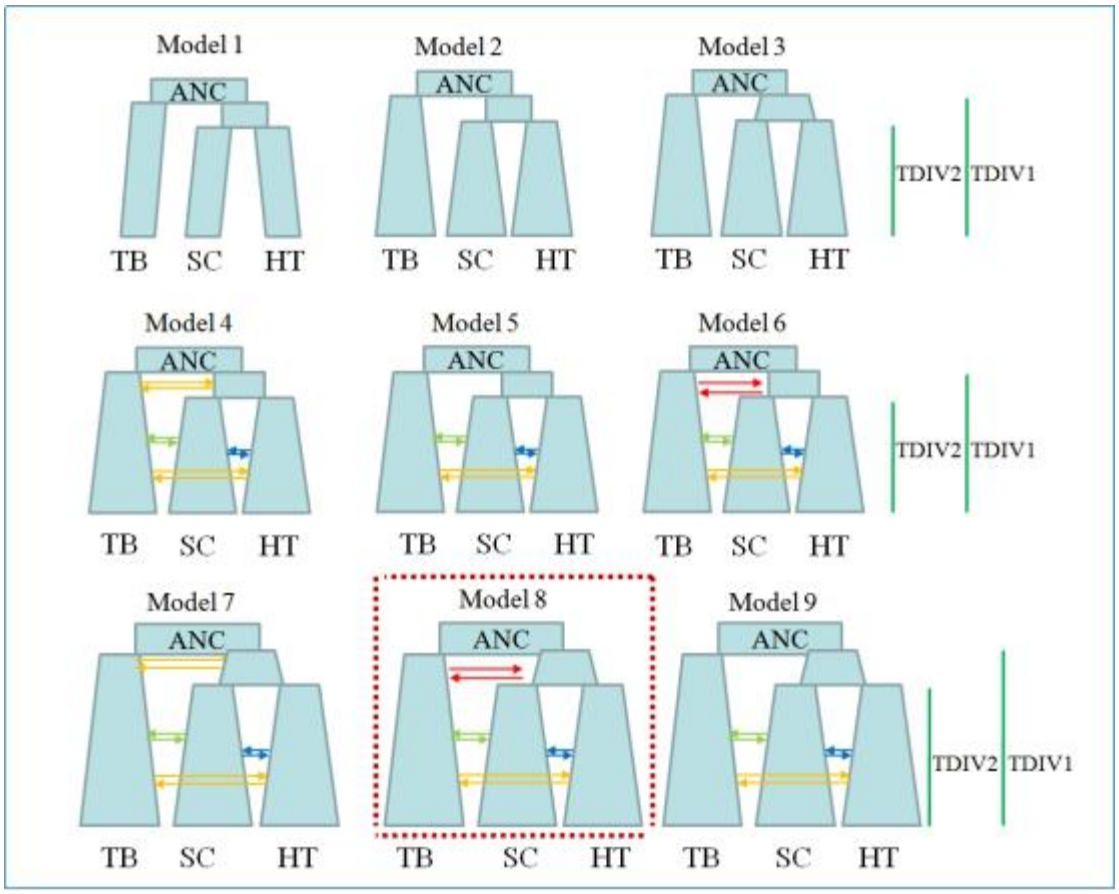

Fig. S3. Model 1-3, during the process of divergence among the three linkages, no gene flows with no changes in effective population size and (Model 1); with changes in effective population sizes starting from the divergence of TB (TDIV1), as well as SC and HT (TDIV2, Model 2); with changes in effective population sizes starting from TDIV1. Model 4-6, the same with model 2 plus three gene flow scenarios including gene flows all the time with the same (Model 4) and different patterns (Model 6) of gene flow in ancestral and recent times (i.e. TDIV1-TDIV2 v.s.TDIV2), and gene flows after divergence of SC and HT (i.e. TDIV2, Model 5). Model 7-9, the same with model 3 plus three gene flow scenarios including gene flows all the time with the same (Model 7) and different patterns (Model 8) of gene flow in ancestral and recent times (i.e. TDIV1-TDIV2 v.s.TDIV2), and gene flows after divergence of SC and HT (i.e. TDIV2, Model 9). Different coloured lines and arrows (yellow, blue, green, red) assume occurrence and time of gene flows. The red dotted square showed the best model.

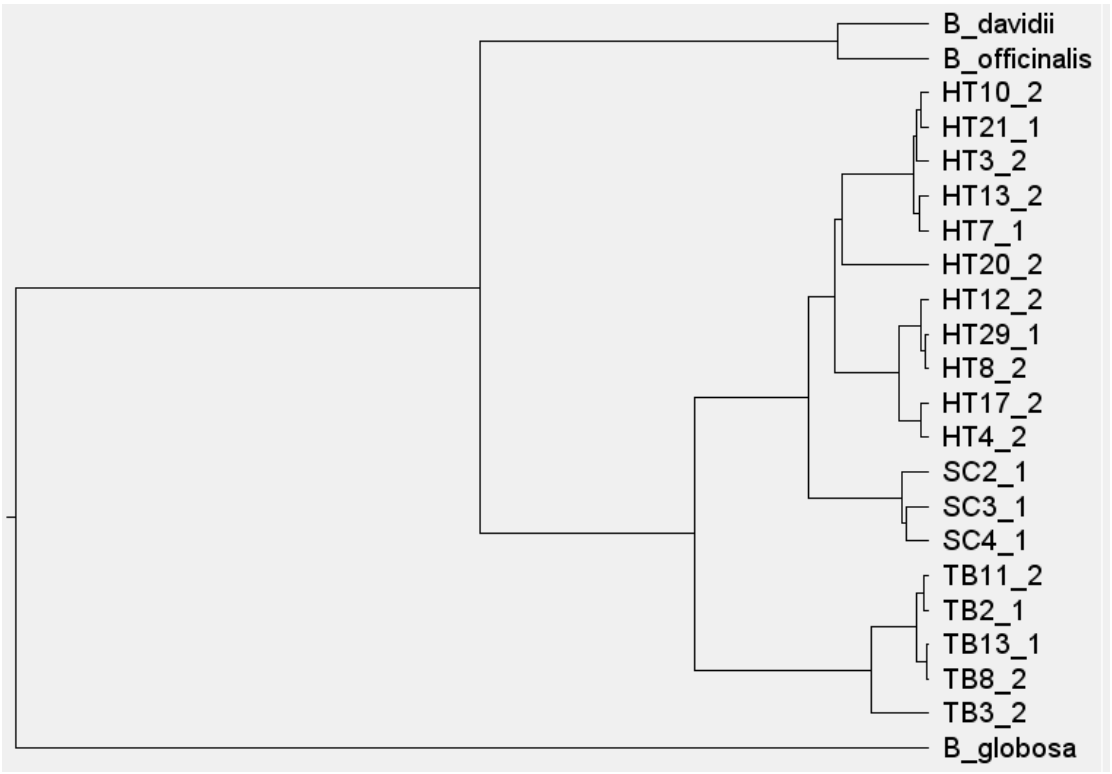

Fig. S4. The phylogenomic tree used for time assignment of divergence for ancestral area reconstruction using representative samples of *B. alternifolia* and three species in the genus are currently available with resequencing data, i.e., *B. globose, B. officinalis* and *B. davidii*.

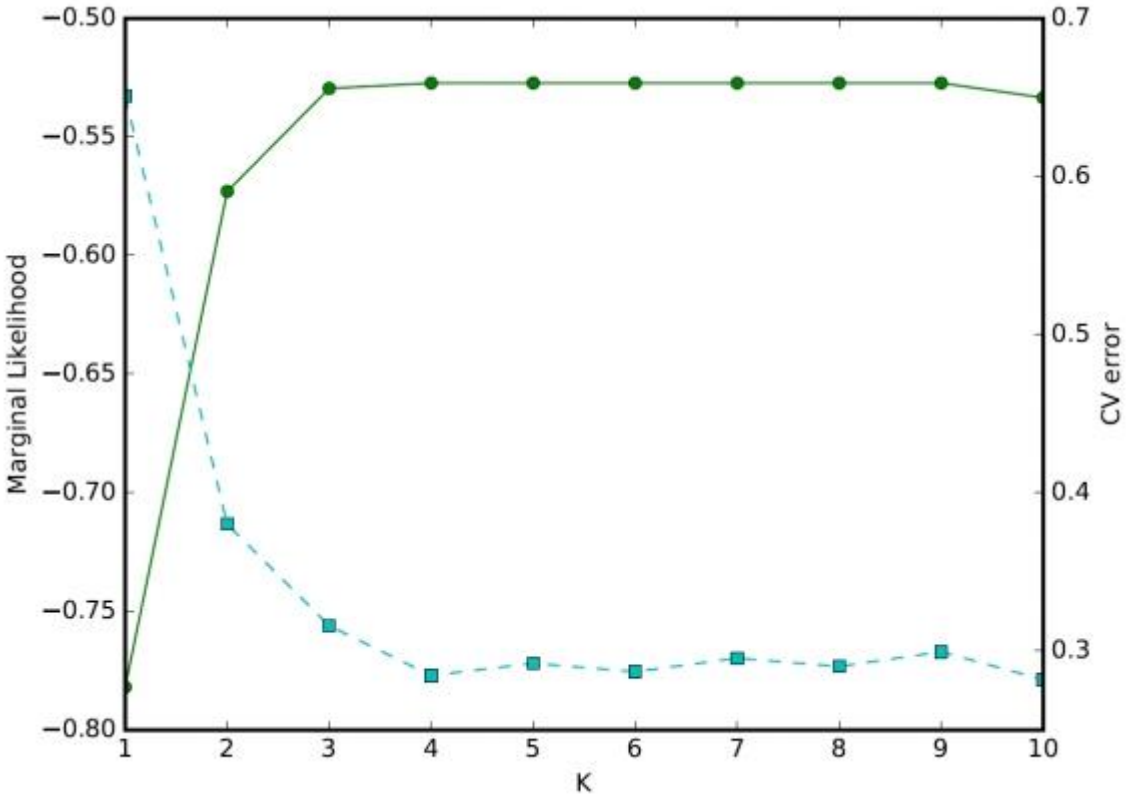

Fig. S5. Cross validation (CV) error (green line) and marginal likelihood (dashed blue line) values for different model *K*.

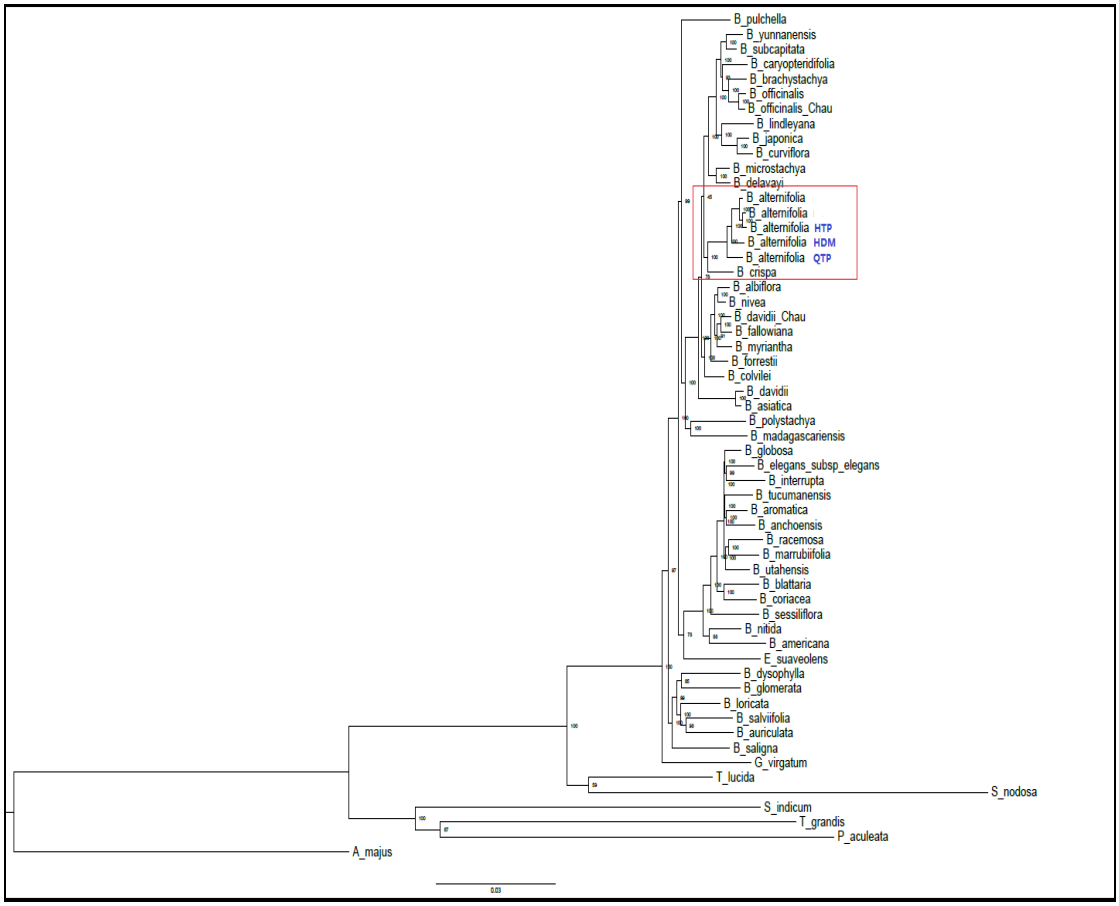

Fig. S6. Reconstructing the phylogenomic relationships for 46 species of *Buddlejas* using single copy genes. The red box indicates the phylogenetic positions of *B. alternifolia* with different distribution areas and its close relative, *B. crispa*.

Supporting Information Methods

Methods S1 Site ancestral state estimation

We firstly mapped sequences obtained from *B.davidii* [SRR7121509], *B. officinali s*[SRR7121979], *B. globosa* [SRR6331521], *T. grandis* [SRR7984127], *S. indicum* [SRR105519(012)] and *A.majus* [ERR2744404], to the *B. alternifolia* reference genome using BWA-MEM with the default parameters. Genotypes were called for each site by Freebayes with same parameters as for above SNP callings, except allowing for monomorphic genotypes output. The ancestral state of *Buddleja* for each chromosome was estimated by employing an empirical Bayesian method using IQ-TREE (Nguyen *et al*., 2014). In total, we identified with high confidence the ancestral state for 755,788,848 sites (88.5% of the total) in the assembled *B. alternifolia* genome; this was used to estimate unfolded SFS and polarize SNPs.

Methods S2 Estimating mutation rate of *B. alternifolia*

We firstly used OrthoFinder2, to find a total of 150 single copy genes among the 17 genomes (the same species as **in identification of orthologous genes and phylogenetic tree construction part**). Then MAFFT (Katoh & Standle, 2013) was used to perform multiple alignment of protein sequences for each set of single-copy orthologous genes. PAL2NAL（Suyama *et al*., 2006) was used to convert protein sequence alignments into the corresponding codon alignments andthen were trimmed using trimAI v1.2 (trimal -gt 0.8). In the end, a total of 241623nucleotides were retained to build the maximum likelihood phylogenetic tree using IQTree. The

divergence time was estimated with r8s v1.81 and calibrated against the divergence timing of Monocotyledoneae and Eudicotyledoneae (synchronously 135-130 million years), of Pentapetalae (126-121 Ma), and Rosidae (123-115 Ma). We estimated the mutation rate of *B. alternifolia* by the mutation rate of *A. thaliana**(*B.alternifolia* branch length/divergence time)/ (*A. thaliana* branch length/divergence time) 】*2 year = $7e^{-9}$*【0.1128/54.1714/0.4153/107.4197】*2=1.08*$7e^{-9}$.

Methods S3 Reconstructing the phylogenomic relationships for 46 species of *Buddlejas* using single copy genes

We firstly download raw reads from sequencing after targeted sequence capture for 45 species of *Buddleja* which were available in the NCBI Sequence Read Archive, BioProject ID PRJNA419999, SRA SRP125765 (https://www.ncbi.nlm.nih.gov/sra/?term=SRP125765). Following the pipeline HybPiper which were detailed described by Chau *et al*. (2017), target reads were selected and assembled respectively. Any samples with assembled coding sequences for less than 50% of target sequences were excluded from further analyses. Additionally, loci with missing data > 50% or with paralogs was filtered out. For each sequence set, sequences were aligned with MAFFT (Katoh & Standle, 2013) using default parameters meanwhile further removing sites with > 5% gaps using Trimal. The concatenated alignment based on remaining single copy genes were used to construct phylogeny tree using IQTREE**.**

References

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology Evolution* **32:** 268-274.

Katoh K, Standley DM. 2013. **MAFFT multiple sequence alignment software version 7:** improvements in performance and usability. *Molecular Biology Evolution* 30**:** 772-780.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34:** W609-W612.

Chau JH, O'Leary N, Sun W-B, Olmstead RG. 2017. Phylogenetic relationships in tribe *Buddlejeae* (Scrophulariaceae) based on multiple nuclear and plastid markers. *Botanical Journal of the Linnean Society* **184:** 137-166.