Available online at www.sciencedirect.com
SCIENCE DIRECT
Molecular Phylogenetics and Evolution 39 (2006) 706–721
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Accounting for variation of substitution rates through time in Bayesian phylogeny reconstruction of Sapotoideae (Sapotaceae)

Jenny E.E. Smedmark *, Ulf Swenson, Arne A. Anderberg

Department of Phanerogamic Botany, Swedish Museum of Natural History, P.O. Box 50007, SE-104 05 Stockholm, Sweden

Received 9 September 2005; revised 4 January 2006; accepted 12 January 2006
Available online 21 February 2006

Abstract

We used Bayesian phylogenetic analysis of 5 kb of chloroplast DNA data from 68 Sapotaceae species to clarify phylogenetic relationships within Sapotoideae, one of the two major clades within Sapotaceae. Variation in substitution rates through time was shown to be a very important aspect of molecular evolution for this data set. Relative rates tests indicated that changes in overall rate have taken place in several lineages during the history of the group and Bayes factors strongly supported a covarion model, which allows the rate of a site to vary over time, over commonly used models that only allow rates to vary across sites. Rate variation over time was actually found to be a more important model component than rate variation across sites. The covarion model was originally developed for coding gene sequences and has so far only been tested for this type of data. The fact that it performed so well with the present data set, consisting mainly of data from noncoding spacer regions, suggests that it deserves a wider consideration in model based phylogenetic inference. Repeatability of phylogenetic results was very difficult to obtain with the more parameter rich models, and analyses with identical settings often supported different topologies. Overparameterization may be the reason why the MCMC did not sample from the posterior distribution in these cases. The problem could, however, be overcome by using less parameter rich evolutionary models, and adjusting the MCMC settings. The phylogenetic results showed that two taxa, previously thought to belong in Sapotoideae, are not part of this group. *Eberhardtia aurata* is the sister of the two major Sapotaceae clades, Chrysophylloideae and Sapotoideae, and *Neohemsleya usambarensis* belongs in Chrysophylloideae. Within Sapotoideae two clades, Sideroxyleae and Sapoteae, were strongly supported. Bayesian analysis of the character history of some floral morphological traits showed that the ancestral type of flower in Sapotoideae may have been characterized by floral parts (sepals, petals, stamens, and staminodes) in single whorls of five, entire corolla lobes, and seeds with an adaxial hilum.

Keywords: Bayesian inference; Bayes factors; Character evolution; Covarion model; Phylogeny; Rate variation; Sapotaceae

1. Introduction

Sapotoideae (Swenson and Anderberg, 2005), one of the two major clades in the Sapodilla family (Sapotaceae), is almost completely restricted to tropical and subtropical areas of the world and contains 543 known species (Govaerts et al., 2001). The majority of the species (ca. 300) occur in the Indo-Pacific region (sensu Cox, 2001). Africa is about half as species rich as the Indo-Pacific region, while the Americas are again about half as species rich as Africa. So far, no morphological character has been found that is strictly unique for Sapotoideae (Swenson and Anderberg, 2005). The group consists of shrubs or trees, often becoming very large. Giants are found for example in *Palaquium* and can grow to be 60 m tall.

Relationships within Sapotoideae were difficult to resolve with the coding *ndhF* sequence data alone (Anderberg and Swenson, 2003) and also in combination with morphological data (Swenson and Anderberg, 2005). In this extended study of Sapotoideae phylogeny we use Bayesian inference to analyze a data set consisting of four noncoding intergenic spacers; *trnH–psbA*, *trnC–petN*, *petN–psbM*, and *psbM–trnD*, and part of a coding gene; the

* Corresponding author. Fax: +46 0 8 5195 4221.
E-mail address: jenny.smedmark@nrm.se (J.E.E. Smedmark).

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.01.018

3′-end of *ndhF*. By adding another 3.3 kb of cpDNA sequence data, and also 28 additional Sapotoideae species, we hope to provide a more complete picture of the phylogeny of the group. The study was designed to investigate two major issues relating to the evolutionary history of Sapotoideae. First, we wanted to identify the major evolutionary lineages within Sapotoideae, and determine what their interrelationships are. A previous study (Swenson and Anderberg, 2005) indicated that, of the larger groups currently recognized based on morphology (Pennington, 1991), Sideroxyleae is a clade (with the exception of *Sarcosperma* and *Diploön* Cronquist), while Isonandreae and Sapoteae (Mimusopeae, Pennington, 1991) both seem to be polyphyletic, although this has not been determined with sufficient support. A group of particular interest is *Sideroxylon*, which Pennington (1991) gave a broader circumscription compared to earlier workers (e.g., Aubréville, 1964; Baehni, 1965; Dubard, 1912; Lam, 1939). Phylogenetic analysis (Swenson and Anderberg, 2005) has clearly indicated that *Argania* and *Nesoluma* are nested within *Sideroxylon* s.l. and in this study we increase the sample from Sideroxyleae in order to better decide whether to include these two genera in *Sideroxylon*, or if any of the earlier classifications better reflect phylogenetic relationships within this group. Second, we wanted to explore the evolution of some morphological characters within Sapotoideae. When it was shown that *Sarcosperma* is the sister of the rest of Sapotaceae (Anderberg et al., 2002), Anderberg and Swenson (2003) proposed that characters present in *Sarcosperma*, such as a simple calyx, undivided corolla lobes, an equal number of sepals, petals, staminodes, and stamens, and seeds with a basal hilum, are likely to be plesiomorphic in Sapotaceae. Thus, the common ancestor of Sapotoideae may be hypothesized to have been characterized by such simple flowers, and the double calyx, segmented corolla lobes, increased numbers of petals, staminodes, and stamens to have evolved later, within Sapotoideae. We used a Bayesian approach to explore the evolution of these characters.

Preliminary phylogenetic analyses of the present data set showed that internal branches, separating the first diverging branches in Sapotoideae, are very short. Previous phylogenetic analyses of the group have used parsimony (Anderberg and Swenson, 2003; Swenson and Anderberg, 2005) and have not been able to resolve these branches with sufficient support. Simulation studies demonstrate that Bayesian inference has a greater ability to correctly resolve short internodes in phylogenetic trees compared to parsimony bootstrap analysis (Alfaro et al., 2003). We were consequently interested to see if Bayesian inference with its allegedly higher sensitivity to the signal in the data set will find any well supported branches that are not identified by a parsimony bootstrap analysis. We also wanted to test whether Bayesian phylogenetic analysis of this expanded cpDNA data corroborates resolution previously obtained by parsimony analysis of combined *ndhF* and morphological data (Swenson and Anderberg, 2005). In this paper, we explore the occurrence of rate variation within Sapotoideae and compare the performance of models that only account for rate variation across sites with covarion models (Fitch and Markowitz, 1970; Fitch, 1971; Tuffley and Steel, 1998) that allow the rate of evolution of a site to vary through time. We also compare the performance of complex evolutionary models, allowing different DNA regions to have their own set of model parameters, with simple models, where the entire data set is assumed to evolve homogenously.

2. Materials and methods

2.1. Plant material and taxon sampling

A total of 66 species classified in Sideroxyleae, Mimusopeae, and Isonandreae by Pennington (1991), and assumed to belong to the clade Sapotoideae (Swenson and Anderberg, 2005) were included in the study. It should be noted that the tribe including the type species of Sapotaceae, *Manilkara zapota*, should be named Sapoteae, since the name Sapotaceae has been conserved (Swenson and Anderberg, 2005). Thus, based on our current understanding of phylogenetic relationships within the group, the correct name of Mimusopeae is Sapoteae, and we will use that name in this paper. Sapotoideae contains 29 genera after the exclusion of *Diploön* and *Sarcosperma*, and the inclusion of *Capurodendron* in this group (Anderberg and Swenson, 2003). Representatives from as many of the recognized genera as possible were included. Of the 182 species placed in Sapoteae (Mimusopeae) by Pennington (1991), 32 were included in the study. Three of the genera that Pennington recognized in this group, *Baillonella*, *Labourdonnaisia*, and *Gluema*, were not available to us. Of the 280 species classified in Isonandreae (Pennington, 1991), 11 were included, representing all but two genera, *Aulandra* and *Isonandra* (Pennington, 1991). Of the 81 species classified in Sideroxyleae (Pennington, 1991), 23 were included. From each of the genera *Madhuca*, *Payena*, *Mimusops*, *Vitellariopsis*, *Labramia*, *Manilkara*, and *Sideroxylon*, more than one species was included to give an indication of whether the genus is monophyletic or not. Since *Sarcosperma* has been shown to be the sister to all other Sapotaceae (Anderberg et al., 2002), *S. laurinum* was included for rooting purposes. One species, *Xantolis siamensis*, representing the other major clade in Sapotaceae, Chrysophylloideae (Anderberg and Swenson, 2003), was also included in the study. Leaf material sampled from these 68 species was either silica gel dried or, more often, from herbarium specimens. Voucher specimens used for study are listed in Table 1.

2.2. Extraction, amplification, and sequencing

DNA extractions were performed with DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) following the instructions provided by the manufacturer. Leaf material was ground in AP1 buffer, using a Mini-beadbeater with 2.5 mm

Table 1
Voucher specimens used for molecular analyses

Species	Tribe	Origin and voucher	Character						EMBL accession				
			1	2	3	4	5	6	*ndhF*	*trnC/petN*	*petN/psbM*	*psbM/trnD*	*trnH/psbA*
Argania spinosa (L.) Skeels	Sid.	Morocco: Nordenstam 9325 (S)	0	0	0	0	0	0	AY230664	AM179647	AM179519	AM179580	AM179712
Autranella congolensis (de Wild.) A. Chev.	Sap.	Congo: Bokdam 4401 (WAG)	3	1	0	0	0	0	AY230666	AM179648	—	AM179581	AM179710
Burckella macropoda (K. Krause) H.J. Lam	Iso.	Indonesia, Bogor: Chase 1359 (K)	1	0	1	2	3	1	AY603773	AM179649	AM179520	AM179582	AM179713
Capurodendron androyense Aubrév.	Chr.	Madagascar: Humbert 28855 (B)	0	0	0	0	0	1	AM179777	AM179650	AM179521	AM179583	AM179714
Capurodendron sp.	Chr.	Madagascar: Schatz and Miller 2477 (S)	0	0	0	0	0	1	AY230668	AM179651	AM179522	AM179584	AM179715
Diploknema butyracea (Roxb.) H.J. Lam	Iso.	Nepal: Polunin, Sykes and Williams 3975 (UPS)	0	0	1	2	3	1	AY230681	AM179652	AM179523	AM179585	AM179716
Diploknema oligomera H.J. Lam	Iso.	Indonesia, Bogor: Chase 1360 (K)	0	0	1	1	3	1	AY603775	AM179653	AM179524	AM179586	AM179717
Eberhardtia aurata (Pierre ex Dubard) Lecomte	Sap.	Cult. South Cina Bot. Gard.: G. Hao 534 (S)	0	1	0	0	0	1	AM179778	AM179654	AM179525	AM179587	AM179718
Faucherea parvifolia Lecomte	Sap.	Madagascar: Birkinshaw et al., 357(P)	2	0	0	0	0	1	AY230687	AM179655	AM179526	AM179588	AM179719
Inhambanella henriquesii (Engl. and Warb.) Dubard	Sap.	Natal: de Winter and Vahrmeijer 8536 (S)	0	1	0	0	0	1	AY230688	AM179656	AM179527	AM179589	AM179720
Labramia costata (Hartog ex Baill.) Aubrév.	Sap.	Madagascar: Randrianmanalinarivo 577 (UPS)	2	1	0	0	0	1	AY230689	AM179657	AM179528	AM179590	AM179721
Labramia mayottensis Labat, Pignal and Pascal	Sap.	Comores Islands, Mayotte: Labat et al., 3309 (P)	2	1	0	0	0	1	AY230690	—	—	—	AM179722
Lecomtedoxa klaineana (Pierre ex Engl.) Pierre ex Dubard	Sap.	Cult. Holland: Veldhuizen 1509 (WAG)	0	1	0	0	0	1	AY230691	AM179658	AM179529	AM179591	AM179723
Letestua durissima (A.Chev.) Lecomte	Sap.	Congo: Normand s.n. (P)	2	1	2	2	3	0	AM179779	—	—	AM179592	AM179724
Madhuca hainanensis Chun and F.C.How	Iso.	Cult. South Cina Bot. Gard.: G. Hao 530 (S)	1	0	1	2	3	1	AM179780	AM179659	AM179530	AM179593	AM179725
Madhuca longifolia (J.König ex L.) J.F.Macbr.	Iso.	Cult. South Cina Bot. Gard.: G. Hao 531 (S)	1	0	1	2	3	1	AM179781	AM179660	AM179531	AM179594	AM179726
Madhuca microphylla (Hook) Alston	Iso.	Sri Lanka: Fagerlind 4790 (S)	1	0	1	2	3	1	AF421064	AM179661	AM179532	AM179595	AM179727
Manilkara butugi Chiov.	Sap.	Ethiopia: Friis et al. 4097 (B)	2	1	0	0	0	0	AM179782	AM179662	AM179533	AM179596	AM179728
Manilkara chicle (Pittier) Gilly	Sap.	Guatemala: Castillo et al., 2083 (B)	2	0	0	0	0	0	AM179783	AM179663	AM179534	AM179597	AM179729
Manilkara concolor (Harv.) Gerstner	Sap.	South Africa: Swensonand Karis 635 (S)	2	1	0	0	0	0	AM179784	AM179664	AM179535	AM179598	AM179730
Manilkara discolor (Sond.) J.H.Hemsl.	Sap.	South Africa: Swenson and Karis 632 (S)	2	0	0	1	3	0	AM179785	AM179665	AM179536	AM179599	AM179731
Manilkara hexandra (Roxb.) Dubard	Sap.	Thailand: Chantaranothai 2340 (Khon Kaen Univ. Herbarium)	2	1	0	0	0	0	AY230694	AM179666	AM179537	AM179600	AM179732
Manilkara huberi (Ducke) Standl.	Sap.	Fr. Guiana: Mori 25803	2	1	0	0	0	0	AM179786	AM179667	AM179538	AM179601	AM179733
Manilkara kauki (L.) Dubard	Sap.	Thailand: Chantaranothai 2341 (Khon Kaen Univ. Herbarium)	2	1	0	0	0	0	AY230695	AM179668	AM179539	AM179602	AM179734
Manilkara mayarensis (Ekman ex Urb.) Cronquist	Sap.	Cuba: Ekman 9971 (LD)	2	1	0	0	0	0	AM179787	AM179669	—	AM179603	AM179735
Manilkara obovata (Sabine and G.Don) J.H.Hemsl.	Sap.	Burkina Faso, Provinz Tapoa: Küppers 937 (FR)	2	1	0	0	0	0	AM179789	AM179671	AM179541	AM179605	AM179737
Manilkara zapota (L.) P.Royen	Sap.	Thailand: Chantaranothai 2378 (Khon Kaen Univ. Herbarium)	2	0	0	0	0	0	AY230696	AM179672	AM179542	AM179606	AM179738
Mimusops caffra E.Mey ex A.DC.	Sap.	South Africa: Swenson and Karis 636 (S)	3	1	0	0	0	0	AM179790	AM179673	AM179543	AM179607	AM179739

Mimusops comorensis Engl.	Sap.	Comores Islands: Pignal and Ginguette 1065 (P)	3	1	0	0	0	0	AY230700	AM179674	—	AM179608	AM179740
Mimusops elengi L.	Sap.	Thailand: Chantaranothai 2305 (Khon Kaen Univ. Herbarium)	3	1	0	0	0	0	AY230701	AM179675	AM179544	AM179609	AM179741
Mimusops obovata Sond.	Sap.	South Africa: Swenson and Karis 633 (S)	3	1	0	0	0	0	AM179791	AM179676	AM179545	AM179610	AM179742
Mimusops zeyheri Sond.	Sap.	South Africa: Dahlstrand 6386 (GB)	3	1	0	0	0	0	AY230702	AM179677	AM179546	AM179611	AM179743
Neohemsleya usambarensis T.D.Penn.	Sid.	Tanzania: Borhidi et al. 84905 (UPS)	0	0	0	0	0	1	AM179792	AM179678	AM179547	AM179612	AM179744
Neolemonniera clitandrifolia (A.Chev.) Heine	Sap.	Ghana: Jongkind, Schmidt & Abbiw 1777 (MO)	0	1	0	0	0	1	AY230703	AM179679	AM179548	AM179613	AM179745
Nesoluma polynesicum (Hillebr.) Baill.	Sid.	Hawaii: Degener 20770 (S)	0	0	1	1	4	1	AY230704	AM179680	AM179549	AM179614	AM179746
Northia seychellana Hook.f.	Sap.	Seychelles: Chong-Seng s. n. (S)	2	1	0	0	0	1	AY230706	AM179681	AM179550	AM179615	AM179747
Palaquium formosanum Hayata	Iso.	Taiwan: Chung & Anderberg 1421 (HAST)	2	0	0	1	3	1	AF421068	AM179682	AM179551	AM179616	AM179748
Payena acuminata (Blume) Pierre	Iso.	Indonesia, Bogor: Chase 1368 (K)	1	0	1	2	3	1	AY603779	AM179683	AM179552	AM179617	AM179749
Payena lucida A. DC.	Iso.	Borneo: Ambri et al., AA1604 (L)	1	0	1	2	3	1	AY230709	AM179684	AM179553	AM179618	AM179750
Sarcosperma laurinum (Benth.) Hook. f.	Sid.	Hong Kong: Saunders 2000 (S)	0	0	0	0	0	0	AF421080	AM179646	AM179518	AM179579	AM179711
Sideroxylon americanum (Mill.) T.D.Penn.	Sid.	Bahamas: Gillis 11576 (B)	0	1	0	0	0	0	AM179793	AM179685	AM179554	AM179619	AM179751
Sideroxylon betsimisarakum Lecomte	Sid.	Madagascar: Schönenberger, von Balthazar and Smedmark A-102 (UPS)	0	0	0	0	0	0	AY230729	AM179686	AM179555	AM179620	AM179752
Sideroxylon cubense (Griseb.) T.D.Penn.	Sid.	Dominican Republic: Beurton and Mory 927 (B)	0	1	0	0	0	0	AM179794	AM179687	AM179556	AM179621	AM179753
Sideroxylon foetidissimum Jacq.	Sid.	Cuba: Lundin 638 (S)	0	0	0	0	0	0	AY230730	AM179688	AM179557	AM179622	AM179754
Sideroxylon horridum (Griseb.) T.D.Penn.	Sid.	Cuba: Gutiérrez and Nilsson 5(S)	0	1	0	0	0	0	AY230731	AM179689	AM179558	AM179623	AM179755
Sideroxylon inerme L.	Sid.	Cult. Denmark: Nielsen s.n. (S)	0	0	0	0	0	0	AY230732	AM179690	AM179559	AM179624	AM179756
Sideroxylon lanuginosum Michx.	Sid.	Texas: Correll and Ogden 28456 (S)	0	1	0	0	0	0	AY230733	AM179691	AM179560	AM179625	AM179757
Sideroxylon lycioides L.	Sid.	South Carolina, USA: Radford et al. 11453 (B)	0	1	0	0	0	0	AM179795	AM179692	AM179561	AM179626	AM179758
Sideroxylon majus (C.F.Gaertn.) Baehni	Sid.	Reunion: Capuron 28185_SF (B)	0	0	0	0	0	0	AM179796	AM179693	AM179562	AM179627	AM179759
Sideroxylon marginatum (Decne. Ex Webb)	Sid.	Cape Verde: Leyens CV-96-672 (B)	0	0	0	0	0	0	AM179797	AM179694	AM179563	AM179628	AM179760
Sideroxylon marmulano Banks ex Lowe	Sid.	Canary Islands: Swenson and Fernandez 581 (S)	0	0	0	0	0	0	AY603783	AM179695	AM179564	AM179629	AM179761
Sideroxylon mascatense (A. DC.) T.D. Penn.	Sid.	Yemen: Thulin, Beier and Hussein 9774 (UPS)	0	0	0	0	0	0	AF421066	AM179696	AM179565	AM179630	AM179762
Sideroxylon obtusifolium (Roem. and Schult.) T.D.Penn.	Sid.	Mexico: Alvarez et al., 28772 (B)	0	1	0	0	0	0	AM179798	AM179697	AM179566	AM179631	AM179763
Sideroxylon occidentale (Hemsl.) T.D.Penn.	Sid.	Mexico: Carter and Sharsmith 4268 (B)	0	1	0	0	0	0	AM179799	AM179698	AM179567	AM179632	AM179764
Sideroxylon reclinatum Michx.	Sid.	USA: Traverse 592 (GB)	0	1	0	0	0	0	AY230734	AM179699	AM179568	AM179633	AM179765
Sideroxylon repens (Urb. and Ekman) T.D.Penn	Sid.	Dominican Republic: Greuter and Rankin 24954 (B)	0	1	0	0	0	0	AM179800	AM179700	AM179569	AM179634	AM179766
Sideroxylon salicifolium (L.) Lam	Sid.	Cuba: Gutiérrez and Nilsson 14 (S)	0	1	0	0	0	0	AY230735	AM179701	AM179570	AM179635	AM179767
Sideroxylon saxorum Lecomte	Sid.	Madagascar: Jongkind 3500 (WAG)	0	0	0	0	0	0	AY230736	AM179702	AM179571	AM179636	AM179768

(*continued on next page*)

Table 1 (*continued*)

Species	Tribe	Origin and voucher	Character						EMBL accession				
			1	2	3	4	5	6	*ndhF*	*trnC/petN*	*petN/psbM*	*psbM/trnD*	*trnH/psbA*
Sideroxylon tenax L.	Sid.	South Carolina: Radford & Leonard 11519 (B)	0	1	0	0	0	0	AM179801	AM179703	AM179572	AM179637	AM179769
Sideroxylon wightianum Hook. and Arn.	Sid.	Cult. South Cina Bot. Gard.: G. Hao 532 (S)	0	1	0	0	0	0	AM179802	AM179704	AM179573	AM179638	AM179770
Tieghemella heckelii (A.Chev) Pierre ex Dubard	Sap.	Ghana: Jongkind 3936 (WAG)	0	0	0	0	0	0	AY230742	AM179705	AM179574	AM179639	AM179771
Vitellaria paradoxa C.F.Gaertn.	Sap.	Benin: Neumann 1512 (FR)	3	0	0	0	0	1	AM179804	—	—	AM179644	AM179776
Vitellariopsis cuneata (Engl.) Aubrév.	Sap.	Tanzania: Thomas 3662 (WAG)	3	1	0	0	0	1	AY230743	AM179706	AM179575	AM179640	AM179772
Vitellariopsis dispar (N.E.Br.) Aubrév.	Sap.	South Africa: Pentz 2 (P)	3	1	0	0	0	1	AY603785	AM179707	AM179576	AM179641	AM179773
Vitellariopsis dispar (N.E.Br.) Aubrév.	Sap.	Mocambique: Balkwill and Balkwill (B)	3	1	0	0	0	1	AM179788	AM179670	AM179540	AM179604	AM179736
Vitellariopsis kirkii (Baker) Dubard	Sap.	Kenya: Robertson 4085 (WAG)	3	1	0	0	0	1	AY603786	AM179708	AM179577	AM179642	AM179774
Vitellariopsis marginata (N.E.Br.) Aubrév.	Sap.	South Africa: Chase 1122 (K)	3	1	0	0	0	1	AM179803	AM179709	AM179578	AM179643	AM179775
Xantolis siamensis (Fletcher) P. Royen	Chr.	Thailand: Smitairi 1 (L)	0	0	0	0	0	1	AY230744	—	—	AM179645	DQ344151

Tribes according to Pennington (1991) are given for each species; Isonandreae (Iso.), Sapoteae (Sap., Mimusopeae according to Pennington, 1991), Sideroxyleae (Sid.), and Chrysophylleae (Chr.). Scores for six morphological characters that were mapped onto the molecular trees. Character 1, calyx a single whorl of five quincuncial sepals (0); two whorls of sepals; each with two (1), three (2) or four sepals (3). Character 2, corolla lobes entire (0) or subdivided into three segments (1). Character 3, Corolla isomerous (0), dimerous (1) or polymerous (2) as compared to calyx. Character 4, Stamens isomerous (0); dimerous (1) or polymerous (2) as compared to calyx. Character 5, Staminodes isomerous (0); absent (1) or anisomerous (2) as compared to calyx. Character 6, position of hilum on seed basal to basiventral (0) or adaxial oblique (1).

zirconia/silica beads (Biospec Products, Bartlesville, OK, USA). Polymerase chain reaction (PCR) was carried out using puReTaq Ready-To-Go beads (Amersham Biosciences, Buckinghamshire, England) in a DNA Thermal Cycler 480 (Perkin-Elmer, Wellesley, MA, USA). The *trnH–psbA* spacer was amplified using the primers described by Hamilton (1999). The *trnC–trnD* region (consisting of the *trnC–petN* spacer, the *petN* gene, the *petN–psbM* spacer, the *psbM* gene, and the *psbM–trnD* spacer) was amplified in two segments; the *trnC–psbM* region with the trnC (Demesure et al., 1995) and psbM2R (Lee and Wen, 2004) primers, and the *psbM–trnD* spacer with the psbM1 (Lee and Wen, 2004) and trnD (Demesure et al., 1995) primers. The 3′-end of *ndhF* was amplified with primers 972 and 2110R (Olmstead and Swere, 1994), which are often called 5 and 10. Thirty-three of the included *ndhF* sequences were published in a previous study by Anderberg and Swenson (2003). PCR products were cleaned using QIAquick PCR Purification Kit (Qiagen, Valencia, CA, USA). In addition to the primers mentioned above, petN1, petN2, psbM2, and petN2R (Lee and Wen, 2004) were used for sequencing of the *trnC–trnD* region, and 1260 (Eldenäs et al., 1999) and 1750R (Anderberg and Swenson, 2003) for *ndhF*. Sequencing reactions were performed with Big Dye Terminator Cycle Sequencing Kit v. 3.1 (Applied Biosystems, Foster City, CA, USA) in a Gene Amp PCR System 9700 (PE Applied Biosystems, Warrington) and cleaned using DyeEx 96 plates (Qiagen, Valencia, CA, USA). Sequencing reactions were read in a 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA).

2.3. Sequence assembly and alignment

DNA sequences were assembled and edited using the phred (Green and Ewing, 2002) and phrap (Green, 1999) modules in Pregap4 and Gap4 (Staden et al., 1998). All new sequences have been submitted to EMBL. Accession numbers are presented in Table 1. Sequence alignment was performed by eye, in the sequence alignment editor Se-Al (Rambaut, 1996). Exons of the *trnC–trnD* region were removed from the data set since primers were located in these regions and therefore data was missing for some taxa, but also to make modelling of sequence evolution more rational. Indels in unambiguously aligned regions were coded for presence and absence. The 29 coded indels begin in the following positions of each matrix: *ndhF*, 939; *trnH–psbA*, 327, 328, 580, 585, 682; *trnC–petN*, 265, 303, 567, 597, 712; *petN–psbM*, 13, 300, 519, 549, 618, 638, 816, 1118, 1207, and 1391; *psbM–trnD*, 447, 603, 859, 1075, 1158, 1220, and 1226. To avoid making uncertain assumptions, only nonoverlapping indels were coded. This binary data set was combined with the molecular data in the parsimony analyses. Since it is not possible to assume unequal rates for a character to arise and disappear in MrBayes (ver. 3B4, Ronquist and Huelsenbeck, 2003), which would be a reasonable assumption for indel data, this data set was not included in the Bayesian analyses. The combined matrix and trees are available at TreeBASE (study ID # S1433).

2.4. Partitioning of the data set and model testing

In initial phylogenetic analyses, three different ways of partitioning the molecular data were compared. In the simplest case, the entire data set was analyzed in a single partition, evolving according to the same evolutionary model. A more complex model, with three partitions, was achieved by treating the *trnC–trnD* region as one large partition, and *ndhF* and *trnH–psbA* as two separate partitions. In the most complex case tested, each of the three spacers of the *trnC–trnD* region (*trnC–petN*, *petN–psbM*, and *psbM–trnD*) was allowed to have different model parameters, resulting in a total of five partitions. For each of the three partitioning schemes, the best performing evolutionary models for each partition were identified under three different model selection criteria, the hierarchical likelihood ratio test (hLRT), the Akaike information criterion (AIC; Akaike, 1973), and the Bayesian information criterion (BIC; Schwartz, 1978). Hierarchical likelihood ratio tests were performed with MrModeltest (ver. 2.0, Nylander, 2004c) and AIC and BIC were calculated with MrAIC (ver. 1.1, Nylander, 2004b). Both programs test the substitution models that can be applied in MrBayes (ver. 3B4, Ronquist and Huelsenbeck, 2003), in combination with different ways of modelling rate variation among sites, namely equal substitution rates, γ-distributed rate variation among sites (+Γ), and some sites being invariable (+I). When the entire data set was treated as a single partition the General time reversible (GTR; Tavaré, 1986) substitution model +I+Γ was selected by all three selection criteria. Six partitioned models were attempted, three combinations of substitution models for three partitions, and three for five partitions. However, due to the failure of the Markov chain Monte Carlo (MCMC) (Larget and Simon, 1999) to sample properly from the posterior distribution under any mixed model, they had to be abandoned (see below).

For all data partitions tested, the three model selection criteria, hLRT, AIC, and BIC, favored models assuming that sites evolve at different rates. These types of mathematical models, for example γ-distribution of rates across sites, require that each site maintain its characteristic rate over the whole evolutionary period. There are, however, models of molecular evolution that allow the rate of evolution of a site to vary through time (Fitch and Markowitz, 1970; Fitch, 1971; Tuffley and Steel, 1998), viz. covarion models (for concomitantly variable codon). To explore whether this type of rate variation is an important aspect of molecular evolution within this plant group, a covarion model was also applied to the simplest partitioning scheme (GTR+Γ+COV).

2.5. Bayesian model selection and phylogenetic analysis

A Bayesian approach to model selection is the use of Bayes factors (Kass and Raftery, 1995; Nylander et al., 2004a; Wasserman, 2000). To compare two models, MCMC is used to generate a representative sample of trees from the posterior distribution under each model of interest. The relative performance of one model compared to the other is measured as the ratio of their model likelihoods, that is, the likelihood of the data under the model (Kass and Raftery, 1995; Nylander et al., 2004a; Wasserman, 2000). The model likelihood may be approximated by the harmonic mean of likelihood values of evolutionary hypotheses sampled from the posterior distribution (Nylander et al., 2004a). The harmonic mean is calculated by the sump command in MrBayes (ver. 3B4, Ronquist and Huelsenbeck, 2003). Thus, with this approach, Bayesian phylogenetic analysis has to be performed under each model of interest in order to decide which one best explains the data.

In MrBayes (ver. 3B4, Ronquist and Huelsenbeck, 2003), the Markov chain was run for 3 million generations, sampling phylogenetic hypotheses every 100 generations. For the mixed models, the rate prior was set to variable to allow rates to vary across partitions, and parameters applying to more than one partition, like the shape of the gamma distribution of rate variation among sites, substitution rates, and state frequencies, were unlinked to allow values to differ among partitions. Convergence of the Markov chain was assumed to have been reached when plots of the overall likelihood, as well as individual parameters of the model, were fluctuating around stable values. Then the first 1,000,000 generations were discarded as burnin and the last 20,000 trees used to construct a majority rule consensus tree. To make sure that the Markov chain really had been sampling from the posterior distribution, three independent analyses, each starting from a random tree, were performed. If the topologies of the resulting majority rule consensus trees contained the same nodes with posterior probabilities above 0.95 and were free from supported incongruencies, this was considered to be the case.

For one of the simplest models, where the entire data set was assumed to evolve according to the GTR+I+Γ model, the temperature parameter was lowered to 0.05, eight parallel chains were run, and the Dirichlet parameter for the updating of state frequencies in each step of the MCMC was increased to 500, in order to get proper mixing of chains, reasonable acceptance rates, and reproducible results. Despite extensive efforts, we did not succeed in adjusting the settings of the MCMC such that reproducible results could be obtained with any of the mixed models. With these models, the resulting topologies and support values differed considerably between analyses with identical settings and therefore had to be abandoned (see below). With a single partition and a model allowing substitution rates to vary through time but not across sites (GTR+COV) the temperature was lowered to 0.05, eight chains were run, the Dirichlet parameter for state frequencies was increased to 800, and the sliding window size for covarion switch rates was increased to 1.4. When rates were allowed to vary across sites, as well as over time (GTR+Γ+COV), the temperature parameter was set to 0.06, eight parallel chains were run, the Dirichlet parameter

for state frequencies was increased to 900, the Dirichlet parameter for the six substitution rates of the GTR rate matrix was increased to 1.2, and the sliding window size of covarion switch rates was set to 2.0.

To confirm that none of the five cpDNA data sets strongly supported a topology that did not agree with any of the others, Bayesian analyses of the separate data sets were also performed (results not shown).

2.6. Pairwise relative rates test

Likelihood based pairwise relative rates tests were performed with HyPhy (ver 0.99B, Kosakovsky et al., 2005) to explore whether substitution rates have varied during the history of the group. To determine whether rates differ significantly between sister groups, rates of each species in one clade were compared to those of all species in the sister clade. Rather than using the same outgroup for all comparisons, an outgroup species was selected from the lineage most closely related to the two groups under analysis. This was done to get accurate estimates of branch lengths and avoid problems of homology assessment that may arise when using a distantly related outgroup. That of the species in the sister group of the two compared clades which had the shortest branch in the majority rule phylogram from the Bayesian analysis was selected as outgroup. Sequences were assumed to evolve according to a global GTR+I+Γmodel with four rate categories. Rate comparisons were made for 24 sister groups. Species for which data was missing for one or more of the sequenced regions (Table 1) were omitted from the analyses to avoid underestimation of rates for these taxa.

2.7. Parsimony bootstrap analysis

A bootstrap analysis with 100,000 replicates was performed with PAUP* (ver. 4.0b10, Swofford, 2002). Each replicate was analysed with heuristic search, creating a single start tree by random addition and improving this tree by TBR branch swapping. A single tree was saved in each replicate and used to construct a majority rule consensus tree and calculate bootstrap proportions (BPs).

2.8. Bayesian mapping of morphological characters

Scoring of character states for six morphological characters (Table 1) was based mainly on information from the literature (Pennington, 1991; Swenson and Anderberg, 2005). In cases where a genus is polymorphic for a character and it was unclear how to score the particular species, herbarium material was investigated. The first character studied was whether the calyx consists of a single whorl of five quincuncial sepals (0), or of two whorls, each with two (1), three (2) or four sepals (3). *Lecomtedoxa klaineana* and *Inhambanella henriquesii* usually have a single whorl of five sepals, but rarely two whorls of either two or three sepals are encountered (Pennington, 1991). We chose to code these taxa as having a uniseriate calyx, as the other condition seems to be exceptional. The second character regards whether the corolla lobes are entire (0) or subdivided into three segments (1). Further division of the lateral corolla lobes occur in *Manilkara* and *Mimusops*, but this was not scored as an extra state. Three characters concern the number of floral parts, corolla lobes, stamens, and staminodes, in relation to the number of sepals. In the case of corolla lobes and stamens, these are either the same number (0), twice as many (1) or more numerous (2) than the number of sepals. We chose to score *Letestua durissima*, in which both the number of corolla lobes and stamens vary and may be either twice or three times the number of sepals, as polymerous (2), since this seemed to be the more derived state, based on preliminary phylogenetic analyses. For the staminode character, the relevant states were isomerous (0) no staminodes at all (1) and staminodes of a number which is not an even multiple of the number of sepals (2). The final morphological character studied was the position of the hilum on the seed, which is either basal to basiventral (0) or adaxial oblique (1).

To explore the evolution of these six characters, they were mapped onto all trees sampled from the posterior distribution under the best model of evolution and the character histories are summarized on the majority rule consensus tree. The mapping was done using the program SIMMAP (ver. 1.0, Bollback, 2004) implementing the Bayesian method posterior mapping (Nielsen, 2002), originally developed for mapping DNA data, but extended to the mapping of morphological characters (Huelsenbeck et al., 2003) under the M_k model (Lewis, 2002). All characters were mapped onto the last 2000 trees sampled from the posterior distribution under the best model. One hundred realisations were performed per tree. This produces an approximate sample of the posterior distribution of character histories that can be used to summarize features of a character history. Based on this sample, the average number of transformations between any two states were calculated. Posterior probabilities of ancestral states for internal nodes in the majority rule consensus tree were also calculated.

2.9. Covarion test

The covarion test by Ané et al. (2005) tests the null hypothesis that rates only vary across sites against the alternate hypothesis that rates also vary according to the Tuffley and Steel (1998) model of covarion evolution. It was performed on the combined data matrix, as well as the separate data sets (*trnH*, *ndhF*, and the *trnC–trnD* region). Since the calculation of the test statistic (W) may be affected by indels, any sites that contain gaps or missing data are removed prior to the analysis. Therefore, all taxa for which data is missing for one or more regions (Table 1) were excluded from the analysis. Two thousand replicates of parametric bootstrapping were performed using a GTR+Γ model with four rate categories. The tree

from the Bayesian analysis (Fig. 1) was used in the analysis. The tree was divided into bipartitions at two nodes where the relative rates tests indicated a substantial change in rates (marked with asterisks in Fig. 1), to measure the independence of the substitution process between the groups.

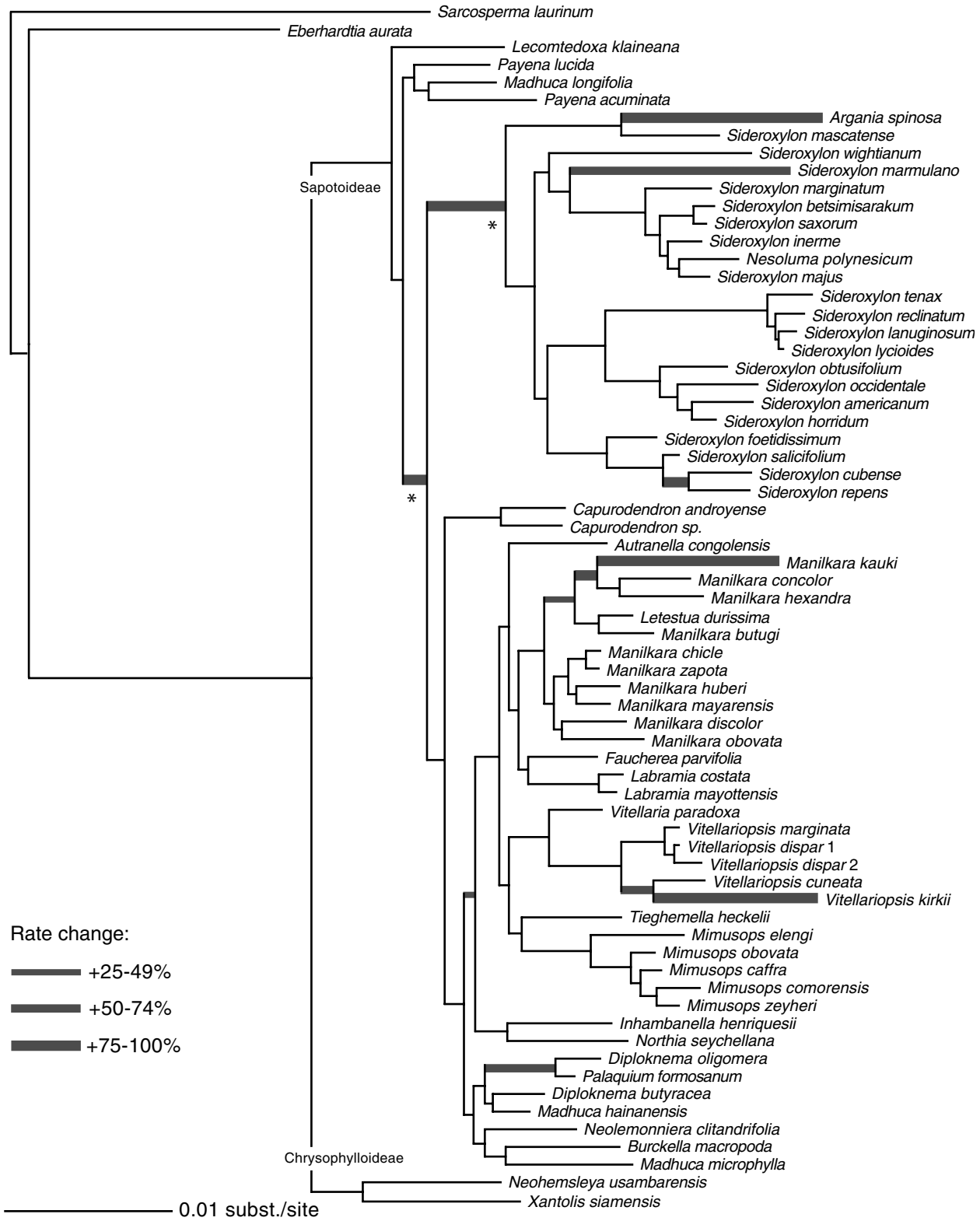

Fig. 1. Majority rule consensus tree from Bayesian analysis under the GTR+Γ+COV model, displaying all compatible partitions. Branch lengths are drawn proportional to the amount of change. Branches of clades where more than 25% of pairwise species comparisons show a significantly higher substitution rate than those of the sister clade are drawn with thicker lines. Asterisks indicate nodes where the tree was divided into bipartitions to measure the independence of the substitution process between the groups in the Covarion test (Ané et al., 2005).

3. Results

3.1. Data

The partial *ndhF* gene sequences ranged in length from 797 to 1064 basepairs (bp), the *trnH–psbA* intron sequences from 249 to 572 bp, the *trnC–petN* intron sequences from 683 to 788 bp, the *petN–psbM* intron sequences from 690 to 1256 bp, and the *psbM–trnD* intron sequences from 417 to 1178 bp. EMBL accession numbers for DNA sequences are shown in Table 1. The combined molecular matrix consisted of 5274 aligned DNA characters and 68 taxa and included 6.7% missing data. Of the 919 characters that were variable in the combined data set, 352 were parsimony informative. The *ndhF* matrix contained 180 variable characters (of which 71 were parsimony informative), *trnH–psbA* 141 (45 parsimony informative), *trnC–petN* 127 (45 parsimony informative), *petN–psbM* 216 (89 parsimony informative), and *psbM–trnD* 218 (73 parsimony informative).

3.2. Difficulties with the MCMC with parameter rich models

We experienced severe problems when running Bayesian phylogenetic analysis of our data set with all of the mixed models. Repeating an analysis with identical settings usually resulted in contradicting topologies of majority rule consensus trees and strongly deviating support for nodes. The likelihood of the data appeared to have reached the same stable value, and tree lengths of the sampled trees were virtually indistinguishable, but still there were well supported (PPs > 0.95) differences in topology between analyses. This variability shows that convergence had not been reached and that the MCMC was not sampling from the posterior distribution. Not even analyses where the number of generations had been increased to 20 million could overcome this problem. Reproducible results could not be obtained without adjusting MCMC settings (see below) and were never reached for any of the mixed models which did not account for rate variation through time, despite extensive analyses.

Four parallel chains are run by default in MrBayes (ver. 3B4, Ronquist and Huelsenbeck, 2003), one cold and three heated. Heated chains can more easily traverse parameter space and more effectively explore any isolated peaks in the posterior distribution. For effective exploration of parameter space, chains also have to change states frequently to avoid that the cold chain gets stuck in one region, failing to visit other regions of potentially higher posterior probability. With our data set, the proportion of successful exchanges between chains was invariably <10% with the default settings in MrBayes (ver. 3B4, Ronquist and Huelsenbeck, 2003), indicating that the different chains did not mix properly (Ronquist and Huelsenbeck, 2004). This problem was more pronounced with more parameter rich models. One option that we used to achieve appropriate mixing was to lower the temperature parameter successively from 0.2 to 0.05 in some cases. The decreased temperature difference between chains does, however, lead to a poorer capability of the MCMC to traverse parameter space. To compensate for this, up to nine heated chains were used in our analyses.

With all tested models, there were some parameters for which acceptance rates for proposed parameter changes in the cold chain were too low (<10%) or too high (>70%) (Ronquist and Huelsenbeck, 2004). By altering proposal parameters that regulate the updating of a specific parameter in the MCMC, the magnitude of proposed steps were either decreased or increased, making them more likely or less likely to be accepted. For example, because the acceptance rate for state frequencies was much too low (often close to zero with the simpler models), the proposal rate was increased by using a higher variance parameter of the Dirichlet distribution.

Extreme acceptance rates also led to difficulties in partitioning the data. In mixed models, a parameter that was unlinked between partitions often had highly variable acceptance rates for the different partitions. Quite often, values were too high for one partition and too low for another. Since the proposal mechanism for a parameter in the MCMC cannot be altered for one specific partition in MrBayes (ver. 3B4, Ronquist and Huelsenbeck, 2003), this problem could not be adjusted. If proposal parameters are altered to increase the acceptance rate, it will increase for all partitions. Therefore, all partitioned models were very difficult to use.

Especially for more complex models, some parameters failed to reach stable values, which was obvious from messy parameter plots and large variances of parameter means. This indicates that, although the overall likelihood of the data had stabilized, the MCMC was not sampling from the posterior distribution. Running more generations usually did not help to overcome this instability. In some cases, linking parameters across partitions, or removing troublesome parameters, could improve the situation.

With models where the data set was divided into three partitions, levels of mixing between chains and acceptance rates could be brought within the suggested span (10–70%, Ronquist and Huelsenbeck, 2004) by adjusting MCMC settings, and parameter values converged on stable values, but analyses still did not produce sufficiently similar topologies. With five partitions, alteration of one proposal parameter always affected acceptance rates of other parameters, as well as the mixing of chains, and stable results could never be obtained. When substitution rates were assumed to be constant through time we were only confident that the MCMC was sampling from the posterior distribution when the data set was treated as a single partition.

3.3. Model choice

Model likelihoods of evaluated models are shown in Table 2. The GTR+Γ+COV model has the highest model likelihood, followed by GTR+COV and GTR+I+Γ. The Bayes factors in favor of GTR+Γ+COV over the other two

Table 2
Number of free parameters and estimated model likelihood for the three evaluated models

Model	Free parameters	Model likelihood	BF over 1	BF over 2
1. GTR+I+G	10	−16,052	—	—
2. GTR+COV	10	−15,999	53	—
3. GTR+G+COV	11	−15,977	75	22

For models two and three, the Bayes factor in favor of each model over the model/models with lower likelihood is given.

models are 53 and 75, respectively (Table 2). According to the interpretation by Kass and Raftery (1995), this indicates strong support for this model, compared with the other two. They suggest that a Bayes factor between 20 and 150 indicates strong evidence against the competing hypothesis (Kass and Raftery, 1995). The Bayes factor in favor of GTR+COV over GTR+I+Γ is 22, which also falls within this span.

3.4. Phylogeny

The majority rule consensus tree from Bayesian analysis under the GTR+Γ+COV model, displaying all compatible partitions, is shown with branch lengths in Fig. 1. Posterior probabilities (PPs) and parsimony BPs of clades, are shown in Fig. 2. Sapotoideae is strongly supported (Fig. 2, node B) but does not include *Eberhardtia aurata* or *Neohemsleya usambarensis*, two taxa included by Pennington (1991) in Sapoteae (Mimusopeae) and Sideroxyleae, respectively. *Eberhardtia* is the sister of the large clade comprising Chrysophylloideae and Sapotoideae (Fig. 2 node A) while *Neohemsleya* groups with *X. siamensis*, which belongs in Chrysophylloideae (Fig. 2, node C). Relationships among lineages in the early history of Sapotoideae, especially taxa classified in Mimusopeae (Sapoteae)-Glueminae or Isonandreae by Pennington (1991), are to a large part not resolved with strong support. An exception is Sideroxyleae (Fig. 2, node D), which is strongly supported and relationships within this clade are also well resolved. It includes *Argania*, *Nesoluma*, and *Sideroxylon*. As sister to the rest of Sideroxyleae we find a clade consisting of *Argania* and *S. mascatense* (Fig. 2, node E), both spiny with a distribution in Africa (and South West Asia for *S. mascatensis*). The remainder of Sideroxyleae is divided into an Old World (Fig. 2, node F) and a New World clade (Fig. 2, node I), neither of which are strongly supported. In the Old World clade, *S. wightianum* from South East Asia is the sister of a nearly completely African clade (Fig. 2, node G). Nested within the otherwise African group of species is the Pacific *Nesoluma polynesicum* (Fig. 2, node H). Outside Sideroxyleae, there are some pairs of species which group together with strong support, for example the two species from *Capurodendron* (Fig. 2, node M) and *Palaquium formosanum* and *Diploknema oligomera* (Fig. 2, node U). Sapoteae s.str. is strongly supported in the Bayesian analysis (Fig. 2, node N) and includes species classified in Mimusopinae and Manilkarinae (Pennington, 1991). It is notable that this group, with a PP of 1.00, does not get a BP above 50 in the parsimony analysis. The phylogenetic position of *Northia*, which has been classified in Sapoteae-Manilkarinae (Mimusopeae–Manilkarinae, Pennington, 1991), is uncertain. It is sister to *Inhambanella* (Fig. 2, node Y), in the Bayesian analyses. *Mimusops* (Fig. 2, node T),*Vitellariopsis* (Fig. 2, node S), *Labramia* (Fig. 2, node R) are all indicated to be monophyletic with strong support. *Manilkara* (Fig. 2, node P), which is indicated to include *Letestua* (Fig. 2, node Q), is strongly supported in the Bayesian analysis, but receives weak support by parsimony. Relationships among these groups are, however, not resolved with convincing support.

3.5. Pairwise relative rates tests

The percent of pairwise species comparisons where rates differ significantly between sister clades was found to be above 75% for six nodes, between 50 and 74% for four nodes, and below 50% for two nodes (Fig. 1). For these nodes, differences were consistent, that is species in one of the clades are always faster than those in the other clade. To avoid assigning a change in rate to the wrong node, cases where significant differences were only found for minor clades, nested within the actual clades under comparison, were left out.

3.6. Character mapping

The average number of transformations between any two states for each of the six morphological characters (Table 1) are shown in Table 3. For each character, a mapping on the majority rule consensus tree from the Bayesian analysis, consistent with the highest posterior probabilities of ancestral states for internal nodes, is shown in Figs. 3A–F.

3.7. Covarion test

The covarion test did not reject ($P = 1$) a rates across sites model in favor of a covarion model for any of the tested data sets, with or without rate variation across sites.

4. Discussion

4.1. Phylogeny

Our analyses indicate with high support that, apart from *Sarcosperma*, another, presumably small, Indo-Pacific lineage, here represented by *Eberhardtia aurata*, diverged before the two major clades in Sapotaceae; Sapotoideae (Fig. 2, node B) and Chrysophylloideae (Fig. 2, node C). Pennington (1991) placed *Eberhardtia* in Glueminae T. D. Penn., together with *Gluema*, *Lecomtedoxa*, *Neolemonniera*, and *Inhambanella*. Of these five genera, four were included in the present study, none of which

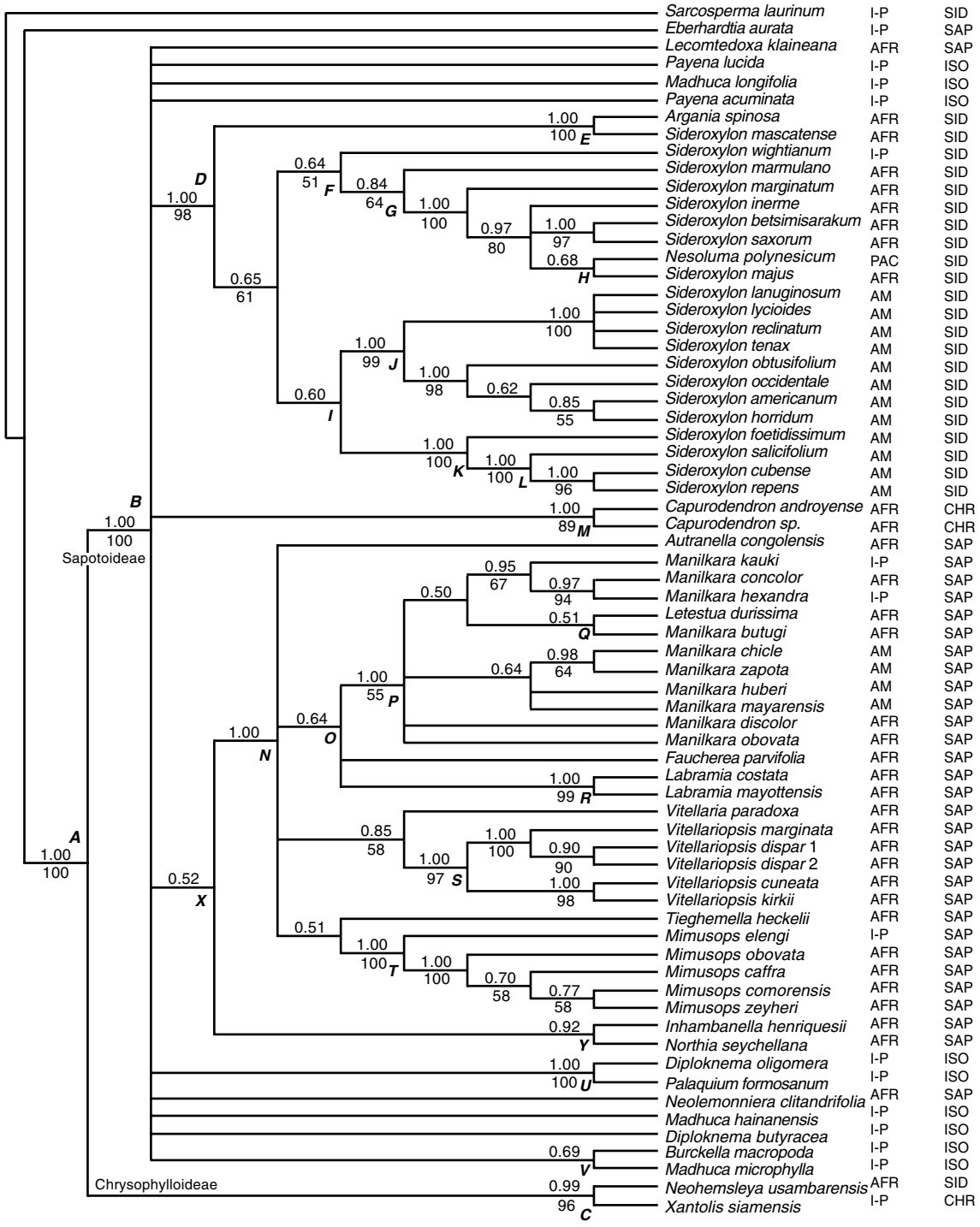

Fig. 2. Majority rule consensus tree from Bayesian analysis under the GTR+Γ+COV model with posterior probabilities of clades marked above branches and parsimony bootstrap proportions below. Letters refer to clades mentioned in the text.

group together (Fig. 2), although in previous phylogenetic analyses (Anderberg and Swenson, 2003; Swenson and Anderberg, 2005) *Lecomtedoxa* and *Neolemonniera* did form a group, albeit with moderate support. While the vast majority of species in Sapotaceae have fleshy fruits, four of the Glueminae genera, *Gluema*, *Lecomtedoxa*, *Neolemonniera*, and *Eberhardtia*, have dry loculicidal capsules, which is probably the reason they were grouped together. In our analyses, *Eberhardtia* is separated from *Lecomtedoxa* and *Neolemonniera* by two strongly supported nodes, indicating a minimum of two origins of dry fruits in Sapotoideae. The capsules in *Eberhardtia* differ from the others by being several-seeded, while the other genera have one-seeded capsules. Also, the three species in *Eberhardtia* all occur in South East Asia, while the other four genera are all African.

Table 3
Number of character states, estimated average number of transformations between states, and between any two specific states for the six morphological characters

Character	States	Trans.	0<=>1	0<=>2	0<=>3	0<=>4	1<=>2	1<=>3	1<=>4	2<=>3	2<=>4	3<=>4
1	4	8.6595	3.8611	2.1085	0.8008	—	0.2613	0.1790	—	1.4489	—	—
2	2	11.2598	11.2598	—	—	—	—	—	—	—	—	—
3	3	6.8818	5.7625	1.0810	—	—	0.0384	—	—	—	—	—
4	3	7.8808	2.3715	4.7470	—	—	0.7623	—	—	—	—	—
5	5	6.8304	0.0299	0.0298	5.6950	1.0175	0.0003	0.0160	0.0035	0.0158	0.0036	0.0191
6	2	6.6489	6.6489	—	—	—	—	—	—	—	—	—

The monotypic *Neohemsleya*, from East Africa, is in our analysis grouping with *Xantolis* with strong support (Fig. 2, node C), indicating that it belongs to the other major clade of Sapotaceae, Chrysophylloideae. *Neohemsleya* was described by Pennington (1991) who considered it to be of doubtful relationships, but placed in Sideroxyleae due to its uniseriate calyx and 5-merous flowers with simple corolla lobes. These characters were hypothesized to be plesiomorphic in Sapotaceae by Anderberg and Swenson (2003) and are also present in many species in Chrysophylloideae. According to Pennington (1991), *Neohemsleya* differs from *Sideroxylon* in its short anther filaments, corolla with almost completely included anthers, and seeds with adaxial hilum, characters that he noted are also present in a few species of *Pouteria* (e.g., *P. kaernbachiana* and *P. anteridifera*). We conclude that *Neohemsleya* is misplaced and leave it to subsequent studies to determine the closest relatives of *Neohemsleya* within Chrysophylloideae.

As noted by Swenson and Anderberg (2005), none of the three tribes recognized by Pennington (1991), Sideroxyleae, Isonandreae, and Sapoteae (Mimusopeae, Pennington, 1991), seem to be monophyletic. However, following the exclusion of *Sarcosperma*, *Diploön*, and *Neohemsleya* from Sapotoideae, one tribe, Sideroxyleae, now corresponds to a monophyletic group. Sideroxyleae is strongly supported in the present study (Fig. 2, node D) and includes *Sideroxylon*, in the wide sense of Pennington (1991), *Argania*, and *Nesoluma*, of which the latter two are nested within the former. Either these two genera have to be included in *Sideroxylon*, or *Sideroxylon* has to be circumscribed more narrowly and segregates of this genus recognized. Our results indicate that Aubréville's (1964) delimitation of genera corresponds well with phylogenetic relationships within the Sideroxyleae clade. Clade G (Fig. 2) includes African species that Aubréville, and most later workers, placed in *Sideroxylon*, including the generic type, *S. inerme*. *Sideroxylon*, as circumscribed by Aubréville, includes species with horizontal embryos, in contrast to the vertical embryos of other members of Sideroxyleae. Nested within this otherwise African group of species is the Pacific *N. polynesicum* (Fig. 2, node H). *Nesoluma* is diagnosed by its large number of corolla lobes and stamens, and lack of staminodes. Pennington (1991) noted that *Nesoluma* is similar to African species of *Sideroxylon*, an observation which is supported by our present results. The Chinese *Sideroxylon wightianum*, which is the type species of *Sinosideroxylon*, one of the genera accepted by Aubréville, is the sister of *Sideroxylon* s.str. (Fig. 2, node F). A second African lineage, according to our results, is the sister group of the remainder of Sideroxyleae (Fig. 2, node E) where we find Aubréville's *Monotheca* (=*S. mascatense*) and the likewise monotypic *Argania*. Future study, also including species of the third of Aubréville's spiny African genera, *Spiniluma*, may show whether these three taxa really constitute distinct lineages or not. It seems likely that *Spiniluma* also will turn out to belong to this clade. The three American genera recognized by Cronquist (1945), and later by Aubréville (1964), *Mastichodendron* Cronquist (Fig. 2, node K, type species: *S. foetidissimum*), *Dipholis* A.DC. (Fig. 2, node L, type species: *S. salicifolium*), and *Bumelia* Sw. (Fig. 2, node J, type species: *S. americanum*), seem to constitute distinct lineages within the same clade, although only the type species was included of the first of these genera. *Bumelia* and *Dipholis* differ from the rest of Sideroxyleae in having segmented corolla lobes (Fig. 3B) and are distinguished by seeds with (*Dipholis*) or without (*Bumelia*) endosperm.

Although our sample from Isonandreae is scarce, with only nine out of 280 species represented, from five of Pennington's (1991) seven genera, this tribe is clearly paraphyletic. The reason for the poorly supported resolution in this part of the tree is probably the short internal branches (Fig. 1), and an increased taxon sample from for example *Palaquium* and *Madhuca*, as well as more data is needed to resolve these branches with better support.

Synapomorphies for Sapoteae s.str. (Fig. 2, node N), which is strongly supported by Bayesian analysis but not by parsimony, are for example the presence of two calyx whorls, each with three or four sepals and the median segment of each corolla lobe which is usually clasping a stamen. Manilkarinae, which is characterized by two whorls of three sepals, is monophyletic if *Northia* is excluded, although the support for this clade is very low (Fig. 2, node O). There are some characters where *Northia* differs from *Manilkara* but is similar to *Inhambanella*, with which it groups in our analyses (Fig. 2, node Y), viz. the exendospermous seed, the large adaxial seed scar, and embryo with plano-convex cotyledons. In spite of its indehiscent fleshy fruits *Inhambanella* was placed in Glueminae (Pennington, 1991) because the calyx generally consists of a single whorl of sepals. Calyces with two distinct whorls, the kind which is present in *Northia*, do, however, occur also in *Inhambanella*. The monotypic

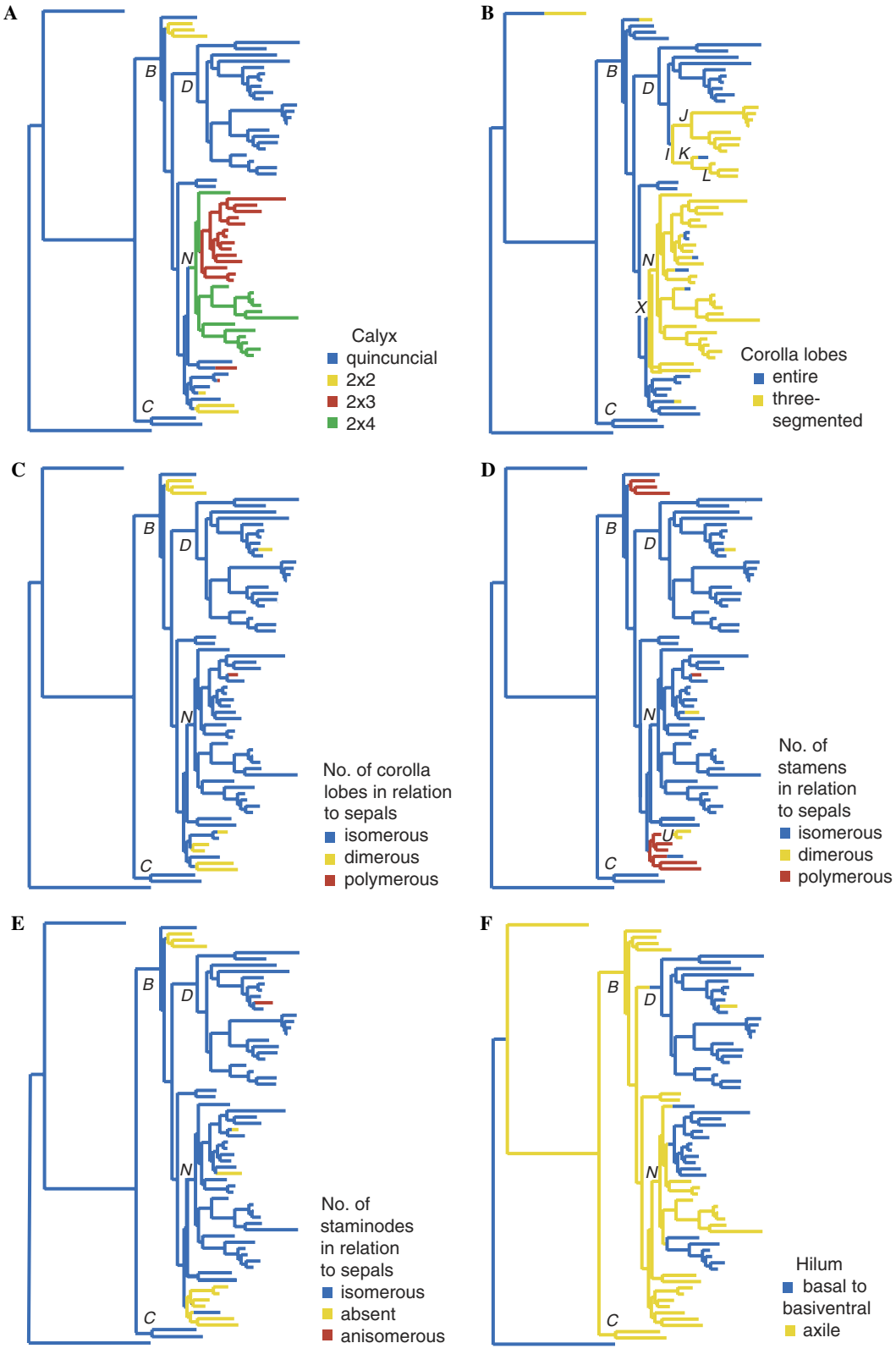

Fig. 3. Six morphological characters mapped onto the majority rule consensus tree from the Bayesian analysis. Letters refer to clades mentioned in the text.

Letestua is here found to be nested within *Manilkara* (Fig. 2, node Q). It should be noted, however, that the material of *L. durissima* used in this study was sterile, hence its identity could not be established with certainty and its position as indicated here must be regarded as tentative. *Letestua* is distinguished by its high number of corolla lobes, stamens and ovary loculi, but otherwise similar to *Manilkara*.

Sapoteae, as circumscribed by Swenson and Anderberg (2005), is not supported in the present analyses. With their suggested circumscription, Sapoteae includes *Mimusops*, *Tieghemella*, *Vitellariopsis*, *Autranella*, *Manilkara*, *Labramia*, *Faucherea*, *Northia*, *Madhuca*, *Payena*, and *Palaquium*. This is not consistent with the present results, since these taxa do not form a clade in either Bayesian or parsimony analyses (Fig. 2). Based on our current understanding of phylogenetic relationships within Sapotoideae there are two strongly supported clades, Sideroxyleae (Fig. 2, node D) and Sapoteae (Fig. 2, node N) and a number of taxa with unresolved relationships. Future phylogenetic studies, with a more extensive taxon sample and more data, will hopefully provide more information about the relationships among taxa outside these two clades.

4.2. Character evolution

Our results show that the ancestral type of flower in Sapotoideae may have been characterized by floral parts (sepals, petals, stamens, and staminodes) in single whorls of five (Figs. 3A, C, D, and E, node B, blue). The fact that our sample from Chrysophylloideae (Fig. 3, node C) is scarce should not affect character optimisations at this node, since multiplications of the number of floral parts are rare in Chrysophylloideae and have only occurred in a few small taxa nested within this clade (Swenson and Anderberg, 2005). The different types of biseriate calyces, with two (Fig. 3a, yellow), three (Fig. 3A, red) or four sepals (Fig. 3A, green) in each whorl, do not seem to be reliable characters for diagnosing groups within Sapotoideae. Over all trees from the posterior, there is an average of 6.80 transformations from a uniseriate to a biseriate calyx (Table 3), showing that this character is homoplasious.

Another homoplasious character state is presence of segmented corolla lobes (Fig 3B). This character undergoes on average 11.30 transformations between entire and segmented lobes (Table 3) when the character is mapped onto all trees from the posterior. Entire corolla lobes is indicated to be the plesiomorphic condition in Sapotoideae (Fig. 3B, blue), despite the fact that those of *Eberhardtia* are segmented. Within Sapotoideae, four separate origins of segmentation (Fig. 3B, yellow) are inferred. One of these is in the American lineage of *Sideroxylon* (Fig. 3B, node I). Segmented corolla lobes are present in two lineages within the clade, one includes species that have been placed in *Bumelia* (Fig. 3B, node J) and the other species that have been placed in *Dipholis* (Fig. 3B, node L), see above. In one species, *Sideroxylon foetidissimum* (Fig. 3B, node K), which is the type of *Mastichodendron*, a reversal to entire lobes has taken place. Thus, there seems to be some phylogenetic information in this character that may be useful for the purposes of diagnosing clades. Segmented corolla lobes are also inferred to have evolved in one large and poorly supported clade (Fig. 3B, node X), within which several reversals to entire lobes are indicated to have taken place. The strongly supported Sapoteae s.str. is nested within this clade (Fig. 3, node N). A synapomorphy for Sapoteae s.str. is the erect median corolla lobe segment, which clasps a stamen. In part of this group (*Mimusops*, *Tieghemella*, *Vitellaria*, *Vitellariopsis*, and *Autranella*), the lateral segments are situated on the dorsal side of the median segment. This type of corolla has been interpreted as consisting of two whorls, of which the outer has lobes divided into two segments, and the inner entire, erect lobes (Swenson and Anderberg, 2005). Since phylogenetic analysis indicated that these taxa are nested within a clade with three-segmented corolla lobes (Swenson and Anderberg, 2005), we chose to follow Pennington's interpretation (1991) in this study. Further division of the lateral segments, so that each corolla lobe has five segments, has taken place in some species of *Mimusops*, *Labramia*, and *Manilkara*.

A reduction in the number of staminodes (Fig. 3E, yellow) is clearly correlated with occurrence of more numerous stamens (Fig. 3D, yellow and red). All taxa that lack staminodes (Fig. 3E, yellow) also have di- or polymerous androecia (Fig. 3D, yellow and red). The plesiomorphic arrangement of stamens in Sapotaceae is opposite the petals, alternating with staminodes. In taxa where staminodes are absent, the additional stamens are found in positions where staminodes occur in other taxa. These patterns suggest that, in these species, staminodes may have been transformed to stamens.

Although a basal or basi-ventral hilum is present in *Sarcosperma* (Fig. 3F, blue), an adaxial hilum occurs in *Eberhardtia* and is ancestral in Sapotoideae (Fig. 3F, yellow), according to our results. Reversals to a basal hilum have taken place in *Sideroxylon* as well as within Sapoteae s.str.

There does not seem to be any correlation between clades with an increased rate of molecular evolution (Fig. 1) and branches where changes in morphological character states are inferred to have taken place (Fig. 3). There are only two instances where the two types of transitions are deduced to have taken place on the same branch (Fig. 3D, node U and Fig. 3F, node D).

4.3. Bayesian phylogenetic analysis

This study shows that variation in substitution rates through time is a very important aspect of molecular evolution in Sapotoideae. Bayes factors strongly support a covarion model, which models rate variation through time by allowing each site to switch between being either on or off in different parts of the tree, over models that only allow rates to vary across sites (Table 2). Accounting for rate variation over time is indicated to be a more important model component than rate variation across sites, since Bayes factors strongly support both GTR+COV and GTR+Γ+COV over GTR+I+Γ. The covarion test by Ané et al. (2005) did, however, not reject a rates across sites model in favor of a covarion model. The reason for this may be that the program removes any sites that contain gaps or missing data, prior to the analysis, since the calculation of the test statistic may be affected by indels. In a data set like this, consisting

mainly of noncoding sequences where variable regions often contain indels, too much of the most variable DNA regions may be excluded for the test to detect covarion evolution. Also, the test assumes a constant level of switches between sites that are on and off, over all lineages in the tree. It may therefore fail to detect covarion evolution in cases where the proportion of invariant sites changes drastically in one part of the tree.

Relative rates tests indicate that changes in overall rate have taken place in several lineages during the history of Sapotoideae (Fig. 1). Although it is a test of the molecular clock hypothesis, and does not test the covarion hypothesis, it shows that overall substitution rates are not constant over time in the group. This rate variation may or may not involve the same sites, but apparently it is better described by a covarion model than a model which does not allow rates to vary over time.

This study emphasizes the importance of the general recommendations to compare results from several independent MCMC runs starting from random trees, to monitor convergence of log likelihoods and parameter values, and to make sure that Markov chains mix properly (Huelsenbeck et al., 2002) and that acceptance rates of proposed steps in the cold chain remain at a reasonable level (Ronquist and Huelsenbeck, 2004). With our data set, the results from a single analysis with the default settings in MrBayes would have been severely misleading, both in terms of topology and node support. According to our experience, if repeated analyses with MrBayes (ver. 3B4, Ronquist and Huelsenbeck, 2003) do not support the same topology, it is best to start by making sure that the chains are mixing properly. If the chains are not swapping states often enough, other problems, for example with acceptance rates or convergence of parameter values, may be solved just by making sure that they do. Of the model parameters, topology and tree length are the most important in phylogenetic inference, and it is especially important to make sure that these parameters converge on stable values. If analyses are still not supporting the same topology there may be some assumptions of the model that are violated by the data, like in our case, and it may be best to modify the model.

It is also clear that only running a program like for example Modeltest (Posada and Crandall, 1998) to select a model, which is often the case in phylogenetic studies, is not enough in terms of exploring the evolutionary processes that have shaped the data set. Thanks to Bayesian inference using MCMC, it has become feasible to use more parameter rich models, and to develop models that are more realistic. This makes Bayes factors, which may be easily calculated from the output from a program like MrBayes (ver. 3B4, Ronquist and Huelsenbeck, 2003), an important tool to evaluate the performance of more advanced models that for example take heterogenous evolution of different types of data into account, or more complex patterns of substitution rate variation, such as individual rate classes for each codon position, or rates that vary over time.

We cannot identify the cause of the analytical problems encountered with some evolutionary models in this study with certainty. A likely explanation, however, is overparameterization, that is that the model contains more parameters than necessary to accurately describe the process that created the data. Complex models require that more parameters be estimated from the same amount of data and superfluous parameters increase the sampling variance, which may lead to errors in parameter estimation and have a negative effect on phylogenetic accuracy (Cunningham et al., 1998). Overparameterization has also been shown to slow the rate of convergence (Rannala, 2002). The difficulties that we experienced with mixing and convergence only occurred with more complex models, which suggests the additional parameters to be a likely cause. As pointed out by Rannala (2002), there is a potential danger in how easily additional paramaters may be added to a Bayesian model. Mixed models that seem reasonable a priori may contain parameters that are not justified and cause problems with the anaysis.

The performance of the covarion model has been examined for quite a large number of genes, and for the majority of these, it provided a better explanation of the data than models that assume substitution rates to be constant over time (Ané et al., 2005; Galtier, 2001; Huelsenbeck, 2002; Lockhart et al., 1998; Miyamoto and Fitch, 1995). The model was developed for coding DNA, and based on the assumption that selective constraints applying to a site vary over time between lineages (Fitch and Markowitz, 1970; Fitch, 1971; Galtier, 2001). In this study, Bayes factors show that the covarion model performs well also with data from noncoding DNA regions, suggesting that it is of general relevance in molecular evolution and deserves a wider consideration in model based phylogenetic inference.

Acknowledgments

We thank I. Bartish for letting us use a yet unpublished *trnH–psbA* sequence from *Xantolis siamensis*, G. Hao for material of *Eberhardtia aurata*, *Madhuca hainanensis*, *M. longifolia*, and *Sideroxylon wightianum*, S. Mori for material of *Manilkara huberi*, the curators of the herbaria B, FR, and UPS for samples for DNA study, J. Nylander for helpful discussion on Bayesian inference, T. Eriksson and N. Wikström for providing computers for analyses, P. Korall for inspiration and support, and M. Simmons and J. Thorne for valuable suggestions to improve the manuscript. This study was supported by a grant from the Swedish Research Council for Ericales phylogeny (to A.A.).

References

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Alfaro, M.E., Zoller, S., Lutzoni, F., 2003. Bayes or Bootstrap? A simulation study comparing the performance of Bayesian Markov chain

Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20, 255–266.
Ané, C., Burleigh, J.G., McMahon, M.M., Sanderson, M.J., 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol. Biol. Evol. 22, 914–924.
Anderberg, A.A., Rydin, C., Källersjö, M., 2002. Phylogenetic relationships in the order Ericales *s.l.*: analyses of molecular data from five genes from the plastid and mitochondrial genomes. Am. J. Bot. 89, 677–687.
Anderberg, A.A., Swenson, U., 2003. Evolutionary lineages in Sapotaceae (Ericales): a cladistic analysis based on *ndhF* sequence data. Int. J. Plant Sci. 164, 763–773.
Aubréville, A., 1964. Sapotacées. Adansonia, Mémoire no. 1, 1–157.
Baehni, C., 1965. Mémoire sur les Sapotacées. 3. Inventaire des genres. Boissiera 11, 1.
Bollback, J.P., 2004. SIMMAP: A Program for Stochastic Mapping of Molecular and Morphological Characters. <http://brahms.ucsd.edu/simmap.html>.
Cox, C.B., 2001. The biogeographic regions reconsidered. J. Biogeogr. 28, 511–523.
Cronquist, A., 1945. Studies in the Sapotaceae 3, *Dipholis* and *Bumelia*. J. Arnold Arbor. 26, 435.
Cunningham, C.W., Zhu, H., Hillis, D.M., 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution 52, 978–987.
Demesure, B., Comps, B., Petit, R.J., 1995. A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol. Ecol. 4, 129–131.
Dubard, M., 1912. Les Sapotacées de groupe des Sideroxylinées. Ann. Inst. Bot.-Géol. Colon. Marseille, Ser. 2 10, 1.
Eldenäs, P., Källersjö, M., Anderberg, A.A., 1999. Phylogenetic placement and circumscription of tribes Inuleae s. str. and Plucheae (Asteraceae): evidence from sequences of chloroplast gene *ndhF*. Mol. Phylogenet. Evol. 13, 50–58.
Fitch, W.M., Markowitz, E., 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4, 579–593.
Fitch, W.M., 1971. Rate of change of concomitantly variable codons. J. Mol. Evol. 1, 84–96.
Galtier, N., 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18, 866–873.
Govaerts, R., Frodin, D.G., Pennington, T.D., 2001. World Checklist and Bibliography of Sapotaceae. Royal Botanic gardens, Kew.
Green, P., 1999. Phrap, version 0.990329. <www.phrap.org>.
Green, P., Ewing, B., 2002. Phred, version 0.020425c. <www.phrap.org>.
Hamilton, M.B., 1999. Four primer pairs for the amplification of chloroplast intergenic regions with intraspecific variation. Mol. Ecol. 8, 521–523.
Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51, 673–688.
Huelsenbeck, J.P., 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. 19, 698–707.
Huelsenbeck, J.P., Nielsen, R., Bollback, J.P., 2003. Stochastic mapping of morphological characters. Syst. Biol. 52, 131–158.
Kass, E.R., Raftery, A.R., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.
Kosakovsky, P.S.L., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 2, 676–679.
Lam, H.J., 1939. On the system of the Sapotaceae, with some remarks on taxonomical methods. Recueil Trav. Bot. Néerl. 36, 509–525.
Larget, B., Simon, D.L., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16, 750–759.
Lee, C., Wen, J., 2004. Phylogeny of *Panax* using chloroplast *trnC–trnD* intergenic region and the utility of *trnC–trnD* in interspecific studies of plants. Mol. Phylogenet. Evol. 31, 894–903.
Lewis, P.O., 2002. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50, 913–925.
Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D.H., Charleston, M.A., Howe, C.J., 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15, 1183–1188.
Miyamoto, M.M., Fitch, W.M., 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. 12, 503–513.
Nielsen, R., 2002. Mapping mutations on phylogenies. Syst. Biol. 51, 729–739.
Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L., 2004a. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47–67.
Nylander, J.A.A., 2004b. MrAIC.pl Version 1.1. Program Distributed by the Author. Evolutionary Biology Centre, Uppsala University.
Nylander, J.A.A., 2004c. MrModeltest 2.0. Program Distributed by the Author. Evolutionary Biology Centre, Uppsala University.
Olmstead, R.G., Sweere, J.A., 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. Syst. Biol. 43, 467–481.
Pennington, T.D., 1991. The Genera of the Sapotaceae. Royal Botanic gardens, Kew.
Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.
Rambaut, A., 1996. Se-Al: Sequence Alignment Editor. Version 2.0a11. <http://evolve.zoo.ox.ac.uk/>.
Rannala, B., 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Syst. Biol. 51, 754–760.
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Ronquist, F., Huelsenbeck, J.P., 2004. MrBayes 3: A tutorial. <http://www.csit.fsu.edu/~ronquist/mrbayes/>.
Schwartz, G., 1978. Estimating the dimensions of a model. Annu. Stat. 6, 461–464.
Staden, R., Beal, K.F., Bonfield, J.K., 1998. The Staden package. In: Misener, S., Krawetz, S.A. (Eds.), Computer Methods in Molecular Biology. In: Bioinformatics Methods and Protocols, 132. The Humana Press, Totowa, pp. 115–130.
Swenson, U., Anderberg, A.A., 2005. Phylogeny, character evolution, and classification of the Sapotaceae (Ericales). Cladistics 21, 101–130.
Swofford, D.L., 2002. PAUP*: Phylogenetic Analysis using Parsimony (* and other methods), ver. 4.0b10. Sinauer, Sunderland, Massachusetts.
Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura, R.M. (Ed.), Some Mathematical Questions in Biology—DNA Sequence Analysis. Amer. Math. Soc., Providence, pp. 57–86.
Tuffley, C., Steel, M., 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147, 63–91.
Wasserman, L., 2000. Bayesian model selection and averaging. J. Math. Psychol. 44, 92–107.