A Phylogeny of Begoniaceae Bercht. & J.Presl.

A thesis submitted to the University of Glasgow
for the degree of Doctor of Philosophy

Laura Lowe Forrest

Division of Environmental and Evolutionary Biology
December 2000

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Declaration

I hereby declare that this thesis is composed of work carried out by myself unless otherwise acknowledged and cited and that this thesis is of my own composition. The research was carried out in the period of April 1997 to December 2000. This dissertation has not in whole or in part been previously presented for any other degree.

"I see a rose, that strange thing, and what's there
But a seeming coloured something on the air
With the transparancies that make up me,
Thickened to existence by my notice"

Norman McCaig, 'Ego'

Abstract

Begonia is one of the largest angiosperm genera, with 1400 species currently recognised. These were placed into 63 sections in the most recent taxonomic treatment. However, there is considerable uncertainty in both section inter-relationships and sectional composition, and there is no formalised phylogenetic hypothesis for the genus. Using the nuclear internal transcribed spacers (ITS) and partial large subunit (26S) sequences of ribosomal DNA, I have produced phylogenetic trees to form the basis of a cladistic framework for the interpretation of the evolution and sectional level systematics of *Begonia*.

Maximum parsimony, maximum likelihood and minimum evolution cladograms were produced for 35 *Begonia*, one *Symbegonia* and two *Datisca* species, for partial 26S, ITS and combined sequence data. The results of the analyses suggest that African taxa are basal in *Begonia*, but that there is not sufficient information to elucidate the precise relationships among these basal lineages. The genus *Symbegonia* is nested deeply within *Begonia*.

A far larger data set was constructed by sequencing the ITS region for 153 species. Different alignment methods (automated, elision and manual) were tested on these sequences, as were different search strategies. The topology which was taken to be the best estimate of the ITS phylogeny of Begoniaceae was constructed using manual alignment, culled of ambiguous regions, and adjusted to reflect the topologies of smaller, localised, compartment analyses. In the resulting tree, the African species of *Begonia* resolve as paraphyletic, with both Asian species (including Socotra) and American species (sister to southern African species) monophyletic.

Comparisons were made between the ITS sequence data and trees produced from the chloroplast *trnC* - *trnD* inter-genic spacer. Parsimony analyses of *trnC* - *trnD* sequences support African taxa as basal in *Begonia*; however, in contrast to the ITS data, *trnC* - *trnD* suggests polyphyly / paraphyly of American taxa, albeit with little bootstrap support.

A morphological data matrix (67 characters for 159 taxa) did not produce a phylogenetic hypothesis for *Begonia* that was congruent with any other

available data. Using a combined morphology - ITS analysis, the fit of individual morphological characters to a fundamental tree was examined. Some characters fitted the combined topology well, although some of the characters which have traditionally been considered important in *Begonia* taxonomy (e.g. the number of placental branches) proved misleading.

The ITS tree was used as a framework for reviewing chromosome evolution in the genus (604 published counts from 239 species). In contrast to sequence divergence, which was greatest among African species, chromosome number diversity was greatest among American species.

The correlation between phylogenetic relationships implied by the ITS tree and the geographical distributions of species was explored to obtain biogeographic hypotheses which may explain the present-day distribution of *Begonia*. As a general rule, related species are geographically proximal, suggesting limited dispersal of lineages. This finding contradicts observations made on morphology, where the close affinities of morphologically disparate (but geographically proximal) taxa were previously unsuspected.

Mechanisms responsible for the evolution of large genera were discussed, and Willis' 'age and area' hypothesis compared to the 'relict' hypothesis of Cronk. In *Begonia*, the morphological diversity of the genus, and most of the species, are encompassed among the putatively derived lineages, favouring the 'relict' hypothesis.

Acknowledgements

As must be the case with any work of this kind, I owe thanks to a large number of people.

Firstly, there are all the people who helped me organise my field work in south west China, 1997, and who assisted when I got there. Particular thanks are due to Mark Tebbitt, Zoe Badcock, Mark Watson, David Chamberlain, Pete Hollingsworth, Jim Dickson, Nick Turland and Rod Taylor. From Kunming, Yunnan, I would like to thank Guan Kaiun, and most specially Tian Dai'ke and Li Di-Xi (who both accompanied me down to the Vietnamese border and across to Nanning, Guangxi; I am particularly grateful to them for taking such care to find me vegetarian things to eat, to Tian for interpretation, to Driver Li for keeping the jeep on the goat-tracks which appear to pass for roads, and to both of them for spotting many *Begonia* that I would have trotted blithely past). Wei Yi Gang looked after me very well in Guilin, Guangxi and showed me a golden *Camilia* (which otherwise I would probably still not know existed). In Beijing, Fu De-Zhe, Qin Hai-Nin, Ban Qin, Song Shu-Yin, Sun Qi-Gao (who sorted out my paperwork, and made sure I left the country), Jin Xiao-Bai (who kindly provided me with living material of *B. morsei*, which survived just long enough for me to get some DNA out of it), Liu Zhao-Hua (who very kindly acted as interpreter, personal shopper and general dogsbody) and Jason Hilton (who gamely assisted in some spectacular hangovers) all deserve acknowledgement.

I am very grateful to Malcolm Wilkins and the MacIntyre *Begonia* Trust Trustees for funding my collecting trip.

People who have provided me with information about *Begonia* include Jan Doorenbos, Marc Sosef and Ferry Bouman in the Netherlands (I am particularly grateful to Ferry for taking me to visit the herbarium in Leiden, and for being good-natured when we found that the specimen we had gone to see was on loan to a 'L. Forrest' in Glasgow....), Martin Sands in Kew, Tracy McLellan, South Africa, and Susan Swensen in Ithaca, New York.

Friends and colleagues who have brought back living or silica dried material of *Begonia* from their travels include Peter Wilkie, Mary Mendum, George

Argent, Toby Pennington, Philip Thomas, Nick Turland, Colin Pendry, Bill Baker, Mark Hughes and Michael Möller.

There are several people who have been involved in keeping the living collections of *Begonia* going. They include John Stevensen, Euan Donaldson, David Menzies, Paul Mathews and many others at Glasgow Botanic Garden. Also Fiona Inches, Allister Reid, Fred Mobeck, Louise Galloway, Andrea Fowler, Steve Scott and Neil Watherston in Horticulture, and Becky Govier (for help with accessioning), at the Royal Botanic Garden Edinburgh.

Essential *Begonia* advice has been provided by Zoe Badcock, Mark Tebbitt, Mark Hughes and Vanessa Plana. The *Begonia* group meetings in the Marina have been a very valuable source of ideas.....

General advice has come from numerous sources, including Cymon Cox, Terry Hedderson, Ken Johnson, Richard Bateman, Quentin Cronk, James Richardson, Rod Page, Vince Smith, Diana Percy, Michael Möller, Kwiton Jong, Hans Sluiman, Pete Hollingsworth and Toby Pennington (for whom I summarise that chapter on species concepts which I spared everyone the ordeal of: Species are progress reports in the history of life (Eldridge, 1995). Fin.)

My attempts in the laboratory have been inevitably assisted by Caroline Guihal, Jill Preston, Michelle Hollingsworth, Alex Ponge and James Richardson. I am very grateful to Jill Harrison, Andy Hudson and other folk at the University of Edinburgh for lab space, chemicals and advice on cloning.

Other people who deserve a mention include the Library staff at the Royal Botanic Garden, Edinburgh. Furthermore, Steve Cafferty has raided the BM library for me on a few occasions, when I've had trouble getting hold of papers.

David Ingram and Richard Bateman kindly allowed me to move my studies across to the Royal Botanic Garden Edinburgh (and Steve Blackmore and Mary Gibby haven't objected yet). I would like to thank everyone at the Gardens (including Antonia) for providing a supportive, enthusiastic and friendly place to study.

Both my supervisors, Pete Hollingsworth and Jim Dickson, have provided support and encouragement where required. The MacIntyre *Begonia* Trust funded the bulk of this work. I hope that this thesis fullfills at least some of its expectations, and that it continues to support basic *Begonia* taxonomy and phylogenetic work. Molecular studies were also supported by a NERC Glasgow Taxonomy Initiative grant.

Jan Doorenbos kindly checked over the reports of chromosome numbers in the table on the CD-ROM. I owe particular thanks to Vanessa Plana and Mark Hughes for talking over ideas, Mark Hughes for manfully trudging his way through entire chapters of thesis, and Pete Hollingsworth for actually reading The Lot, and improving most of it. (Thanks are also due to Michelle Hollingsworth for her tolerance of house invasions and for pasta.) Without Pete's help and advice this thesis would have been significantly less interesting (which I am sure he will find hard to believe) and I am very grateful for all he has done (which I suspect he will find even harder to believe.....).

I must also thank my family for believing that a botanical PhD is some sort of worthwhile investment (and for the river of money).

I intended to put the following quote in somewhere as a reminder that, when talking on evolutionary timescales, unlikely events can happen. It seems, however, that there are people out there (Mark) who consider it equally relevant to the final completion of this thesis (and thanks to Silvia for making perfection unnecessary).

"the improbable is possible and ... the possible can occur"

Siddall and Kluge, 1997

Table of Contents

2. Using molecules to reconstruct evolutionary history

5. Establishing the backbone - ITS and 26S

11. Cytology

12. Evolution, Biogeography and the Begoniaceae

List of Figures

List of Tables

1. WHY ARE SOME GENERA LARGE?

1.1 Taxonomists have long been intrigued by very large and very small genera. The extreme variation which exists in genus size (which ranges from 1 to around 2000 species in vascular plants) prompts several questions. Is this size distribution indicative of some natural phenomenon / phenomena, or is it an artifact of the way we produce our classifications? Are we consistent in the way we perceive morphological discontinuities (and do we view some characters as more important than others in delimitation of taxonomic rank)? Are very large (or very small) genera useful to the consumers of taxonomic output? This thesis is an exploration into patterns of species diversity in large genera, focusing on *Begonia* L. (Begoniaceae Bercht. & J.Presl.) and is intended to address some of these issues.

It is first worth defining what a large genus is. Species number per genus is one way of measuring this, although it is not necessarily a measure of successfulness but presumably of morphological discontinuity, as the genus may be made up of many rare species. A large genus is a genus with a lot of species in it, but could equally be a genus with a lot of individuals in it. In this thesis, a large genus is arbitrarily defined as one including 400 or more species.

1.2 Genus size

Among the largest genera of vascular plants are *Euphorbia* L. (Euphorbiaceae Juss., c. 2000 species), *Piper* L. (Piperaceae Giseke, c. 2000), *Carex* L. (Cyperaceae Juss., c. 2000), *Astragalus* L. (Fabaceae Lindley, c. 1750), *Solanum* L. (Solanaceae Juss., c. 1700), *Senecio* L. (Asteraceae Martinov., c. 1250), *Psychotria* L. (Rubiaceae Juss., 800-1500), *Acacia* Miller (Fabaceae Lindley, c. 1200), *Pleurothallis* R.Br. (Orchidaceae Juss., c. 1120), *Bulbophyllum* Thouars (Orchidaceae, c. 1000), *Miconia* Ruiz & Pavon (Melastomataceae Juss., c. 1000) and *Syzygium* Gaertner (Myrtaceae Juss., c. 1000) (figures from Mabberly, 1997; for a more complete list of large genera (taken from Minelli, 1993) see Appendix 14.1 a, b). *Begonia* is estimated at 900 species (Mabberly, 1997), although the more reliable estimate of at least

1400 (Doorenbos et al., 1998) certainly places the genus well within the ten largest vascular plant genera.

Some families include more than one large genus, prompting the question, do they have some biological attributes which make them more liable to include large genera, or are they just large families? There are 71 genera with 400 or more species and they are contained in 42 families (see Appendix 14.2). Dividing the total number of species by the total number of genera for each family to get a mean species number per genus allows a very rough comparison with values given by Clayton (1974) of an average of 18 species per genus for the angiosperms overall[1]. Values for the large-genus-containing families (given in Appendix 14.2) are frequently higher than Clayton's figures, with 38 of the 42 families having a mean genus size of over 20 species. 23 of the 71 large genera are contained in only four families (Orchidaceae, Asteraceae, Fabaceae and Euphorbiaceae), suggesting a non-random distribution of large genera.

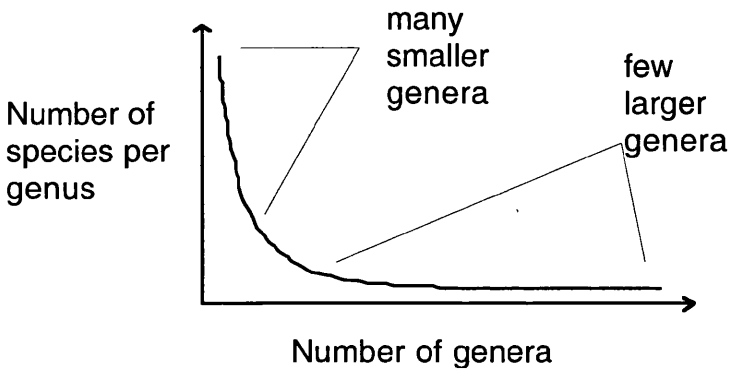
The number of large genera included in a family is positively correlated with the total species number for the family. (Obviously, small families cannot have many large genera - Aquifoliaceae A.Rich. has a total size of c. 420 species. Aquifoliaceae could include a maximum of 1 large genus, while Asteraceae could include a maximum of 56.)

1.3 The hollow curve

Several authors have explored the distribution of genus size within plant families. Plotting the number of species in a genus against the number of genera for a given family gives a characteristic 'hollow curve' distribution. The right-skewed shape of this curve is due to an excess of monotypic taxa and a dearth of larger taxa (Clayton, 1974) (see Figure 1.1).

[1] The figures are not directly comparable, as Clayton took his values from Shaw's Dictionary of the Flowering Plants and Ferns (1966), while mine come from Mabberly (1997).

Figure 1.1: The hollow curve distribution (number of species per genus for a family)



Explanations for this taxonomic pattern focus on either:

1. The behaviour of taxonomists (pragmatic decision making, folk history and chaining)
2. Natural phenomena

or a combination of these factors.

1.3.1 Behaviour of taxonomists

A. Genus size and the importance to man: Walters (1986) considers the size distributions of plant genera and families to be taxonomic artifacts. The average number of species per genus in the Poaceae Caruel is 15.5, while in the Cyperaceae it is 44.5. These families are similarly widespread in Europe and present similar problems in identification (with reduced complex flowers). Walters (1986) suggests that the difference in treatment between them reflects Linnaeus' formalisation of an extensive 'folk taxonomy' resulting from economic usage of grasses in Europe; this folk taxonomy is absent from the economically less important sedges.

B. Historical correlation: According to Walters (1986), looking for some natural law to explain the 'hollow curve' distribution overlooks the fact that many large genera are old historically (as opposed to biologically). Genus sizes follow the same pattern in any "reasonably large modern Angiosperm family": a few large genera and a lot of small genera (Walters, 1986, p. 535). The large

genera are "nearly always" in *Species Plantarum* (Linnaeus, 1753) (i.e. historically older), while the smaller ones tend to be nineteenth century creations (i.e. historically more recent) (Walters, 1986). Effectively his argument was that subsequent taxonomists have been more likely to add to existing Linnaean genera than to create new ones (this is described as 'chaining', which is people's tendency to add to taxa which already exist in preference to creating new ones). However, Cronk (1989) points out a flaw in Walters' (1986) argument: that the genera Linnaeus knew would most often be widespread therefore often large; monotypic genera would thus be expected to be found more recently.

C. Taxonomic pragmatism: Taxonomists deliberately try to keep large taxa small - probably as they try to keep a classification usable, while the creation of small taxa is due to taxonomists' "obvious predilection for the excision of solitary outliers" (Clayton, 1974, p. 278). Clayton believes the sizes of genera to have been influenced by convenience, in favour of simple circumscription and easy identification (Clayton, 1983). Cronk (1989) also finds that "[o]versized taxa and monotypic taxa make plant taxonomy irredeemably inefficient" (Cronk, 1989, p. 368). However, Cronk (1990) describes a general trend in classifications of the Fabaceae to, in practice, retain a similar median genus size over time (from Linnaeus in 1753 through to Hutchinson, 1964) despite range changes in genus size - large genera are growing bigger (through chaining) but very small genera are also being created.

Taxonomists may also consider the 'principle of ease of identification' (Backlund & Bremer, 1998) - big genera like *Begonia* and *Rhododendron* L. (Ericaceae Juss.) are currently very easy to identify to genus level; some of that utility may be lost by splitting.

D. Conservation and politics: A recent Nature paper (Myers et al., 2000) argued for targeted areas of the world for 'silver bullet' conservation money, based partly on consideration of the numbers of endemic genera in each region, with priority given to regions richest in such taxa. Counts of endemic genera per region would be inflated by splitting large genera. If this approach is adopted then there may be the temptation for taxonomists to bias

taxonomies towards the creation of endemic genera when they feel that a geographical region is in need of greater conservation recognition.

1.3.2 Natural phenomena: It is difficult to disagree with Minelli: "Raikow (1986) feels that only the vagaries of taxonomy explain the higher number of species belonging to the passerine birds, compared with the non-passerine birds..... I cannot agree with Raikow's views, as it is difficult to disprove the fact that there are more types of cats than there are types of elephants and more types of cone shells than there are types of nautilus. Despite the vagaries of systematics, there are, in a taxonomic sense, many dense clusters of biotic diversity" (Minelli, 1993, p. 185).

A. Age and Area: Willis (1922), who first described the hollow curve phenomenon, argued that it had a natural basis. He explained it by his 'Age and Area' hypothesis, namely, that younger taxa are less species rich and less widely distributed, while older taxa have had more time to diversify into species.

B. Relict Hypothesis: Cronk's Relict Hypothesis (1989) also provides a natural explanation for the observed pattern. Cronk (1989) explains the monotypic endemic genera of St Helena as ancient relicts. They are taxonomically and geographically isolated and he feels that such a pattern is better explained by widespread extinction than by some "special evolutionary syndrome" (Cronk, 1989, p. 359). He describes three factors to explain the phylogenetic patterns of species richness over time:

1. A bloom period, when groups diversify in species.
2. Evolutionary stasis.
3. Extinction, at more or less constant rate (approximating exponential decay).

The world's flora will always consist of groups at different stages in this progression; in general large genera are young and monotypic genera old; for example, on average the Magnoliidae have far lower numbers of species in their families than the dicotyledons as a whole, while the Asteridae have higher species' numbers per family (Cronk, 1989). Recent angiosperm phylogenies (e.g. Soltis, Soltis & Chase, 1999) certainly put the magnoliids as a more basal

clade than the asterids. Within the Fabaceae, subfamily Caesalpinoideae has a low median genus size (Cronk, 1990), so, following this argument, is an ancient relict group. However, interpreting legume phylogeny in this manner is extremely problematic, as phylogenetic study (Doyle et al., 1997) has revealed the Caesalpinoideae to be paraphyletic and so not suitable for consideration as a single evolutionary unit.

C. Partitioning of Biodiversity: The relict hypothesis affects the partitioning of taxonomic diversity for two reasons (Cronk, 1989):
1. Groups in 'bloom' phase may be recognised as a single taxon because diversity produces intermediates and discontinuities are not evident.
2. Groups depleted by extinction may have "large areas of empty phenetic space" (Cronk, 1989, p. 368) so taxonomic boundaries are clear.
Therefore new groups will be poorly divided and old groups, well divided: "Speciation tends to fill phenetic space, extinction to empty it. Although speciation produces clades, extinction produces taxa" (Cronk, 1989, p. 368).

D. Summary: Despite observations which attribute to psychological or historical factors the numbers of species in genera, and the observation (Cronk, 1990) that taxonomists are usually happier to describe new species than new genera, the role of evolutionary process in the generation of patterns of biological diversity cannot be ruled out; one or a few subtaxa often account for much of the diversity in higher taxa - most mammals are rodents, most birds are passarines and most insects are beetles (Heard, 1992).

1.3.3 Combination: Clayton (1974) thought hollow curve frequency distributions to be due to a combination of natural and psychological phenomena. Cronk (1989) also believes that both factors contribute; he found that he could explain about half the 'hollowness' of hollow curves by psychological and historical factors, while the other half is due to biological reality.

1.4 Phylogenetic inputs

1.4.1 The need for monophyly: In the light of a phylogeny, the question 'What is a genus?' is one of rank, not of monophyly (Wojciechowski, Sanderson & Hu, 1999). If a reliable phylogeny is not available, influences of taxonomic grouping can be misleading as recent molecular data have shown that some traditional genera are paraphyletic, or actually consist of phylogenetically disparate taxa. For instance, *Eupatorium* L. (Asteraceae) was once considered a very large genus with about 1200 species (Mabberly, 1998). A recent ITS phylogeny (Schmidt & Schilling, 2000) found that many of the species included in *Eupatorium* s.l. are scattered across several clades; the characters used to define the genus were sympleisiomorphies which occur throughout the tribe Eupatoriinae. Restricting *Eupatorium* to the 42 species which form a monophyletic assemblage allied to the type species allowed generic synapomorphies to be identified.

Guyer and Slowinski (1993) express concern that studies which count the distribution of units within larger taxa (e.g. species within genera) may indicate more about how taxonomists delimit these taxa (as discussed above) than about evolutionary pattern. Phylogenetic trees, which have now more or less replaced the use of taxonomic lists of genus or family size to extrapolate macroevolution, represent histories of the diversification of clades (Mooers & Heard, 1997) (i.e. real events) and negate the potential problems caused by arbitrary taxonomic decision-making.

1.4.2 The shape of phylogenetic trees: Before discussing inferences from phylogenetic trees it is worth discussing terminology and some theoretical issues regarding their interpretation. Any phylogeny can be separated into 3 distinct parts (Lapointe & Cucumel, 1997):

1. topology (tree shape)
2. branch length (difference in evolutionary change between clades)
3. label position (the phylogenetic relationships).

A. Terminology: Terminology can be confusing; the topology of trees which are not symmetric (or 'balanced') is varyingly referred to as comb-like (which may also be used to describe unresolved trees), an Hennigian comb,

unbalanced or pectinate. In a pectinate tree only one of the two descendant species continue to speciate after a splitting event; in a balanced tree, all extant lineages participate equally in cladogenesis (Kirkpatrick & Slatkin, 1993).

'Stemminess' is a measure of the relative amount of change, such as branch length differences within and between clades. A 'stemmy' tree is one where there is more opportunity for change before speciation events than after (Salisbury, 1999) (see Figures 1.2 - 1.5).

Fig. 1.2: Balanced, unstemmy tree

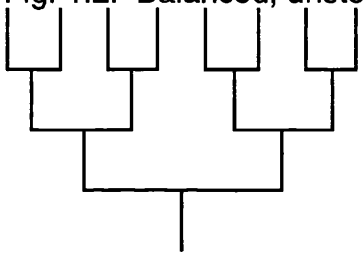


Fig. 1.3: Balanced, stemmy tree

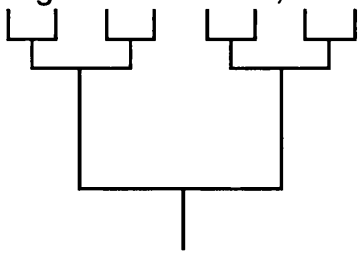


Fig. 1.4: Pectinate, unstemmy tree

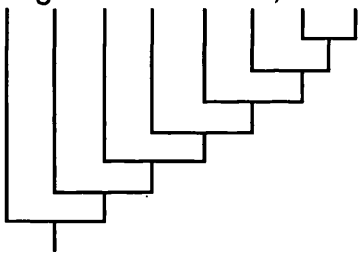
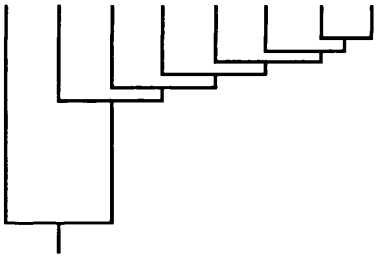
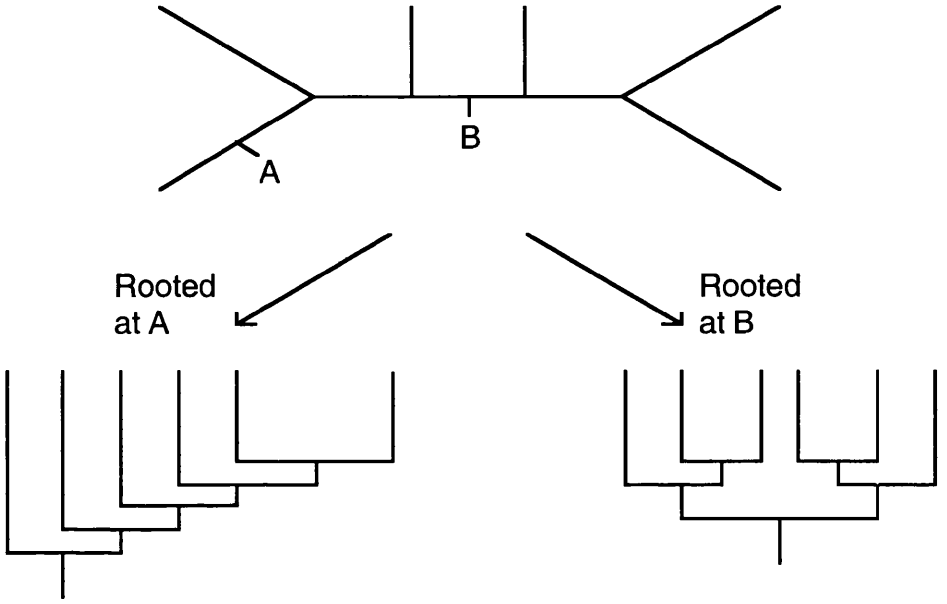


Fig. 1.5: Pectinate, stemmy tree



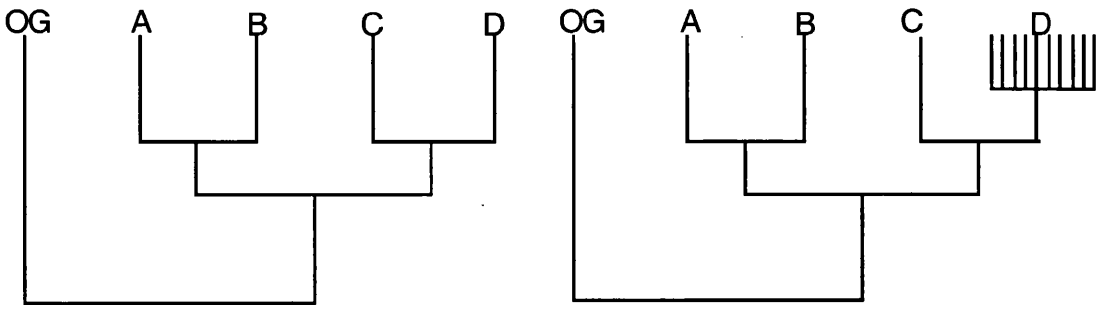
The relative symmetry (or lack of) is a function of a rooted tree; a fully pectinate rooted tree is produced from a symmetric network. The same network, with different rooting, can produce a balanced tree (see Figure 1.6).

Figure1.6: How rooting can affect tree symmetry



It is also vital to consider sampling strategy in any consideration of tree shape; frequently in phylogenetic analyses taxa serve as exemplars for other similar species. Often the tree shape as reconstructed is not the primary concern. For example a tree may appear perfectly balanced, but if, in biological reality, there is (for example) one taxon in clade A, one in clade B, 1 in clade C and 10 in clade D, the real situation is unbalanced (see Figure 1.7):

Figure 1.7: Symmetry and Tree Balance

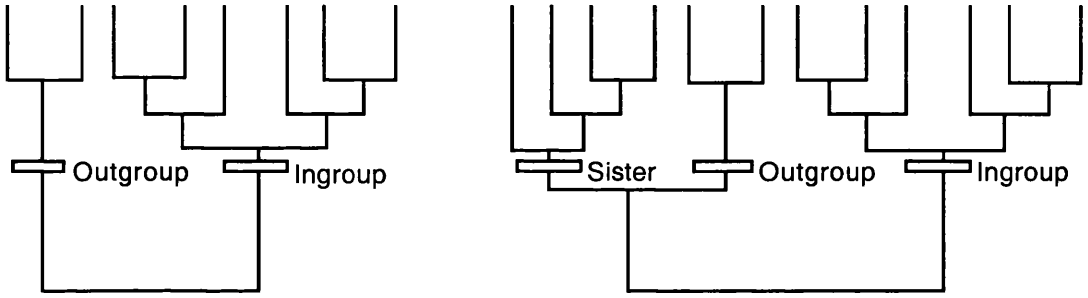


Tree reconstructed through sampling Real clade sizes

Therefore a balanced tree shape can represent an 'unbalanced' (pectinate) reality, where one lineage (that leading to D) is far more species rich than the others (those leading to A, B and to C). The outgroup should not be

considered in questions of balance because of differences in sampling strategy. Even where a small outgroup is used, and 100% of the taxa within it are sampled, it is important to ensure that it is the monophyletic closest sister to the ingroup. For example, Figure 1.8 shows a situation where sampling could lead to the supposition that the outgroup is considerably less species-rich than the ingroup, but when the entire monophyletic assemblage is considered it is evident that the ingroup is of comparable size to the outgroup and its sister group:

Figure 1.8: Outgroups and tree balance



Tree reconstructed through sampling Real clade sizes

This could be equivalent to rooting an analysis of Violaceae Batsch. (c. 800 species) on the smaller family, Turneraceae Kunth ex DC (c. 100 species); phylogenetic analysis (Savolainen et al., 2000; combined *atp*B-*rbc*L) shows that Passifloraceae Juss. ex DC is sister to Turneraceae; it contains approximately 575 species and so balances the larger Violaceae.

B. Evolutionary Scenarios:

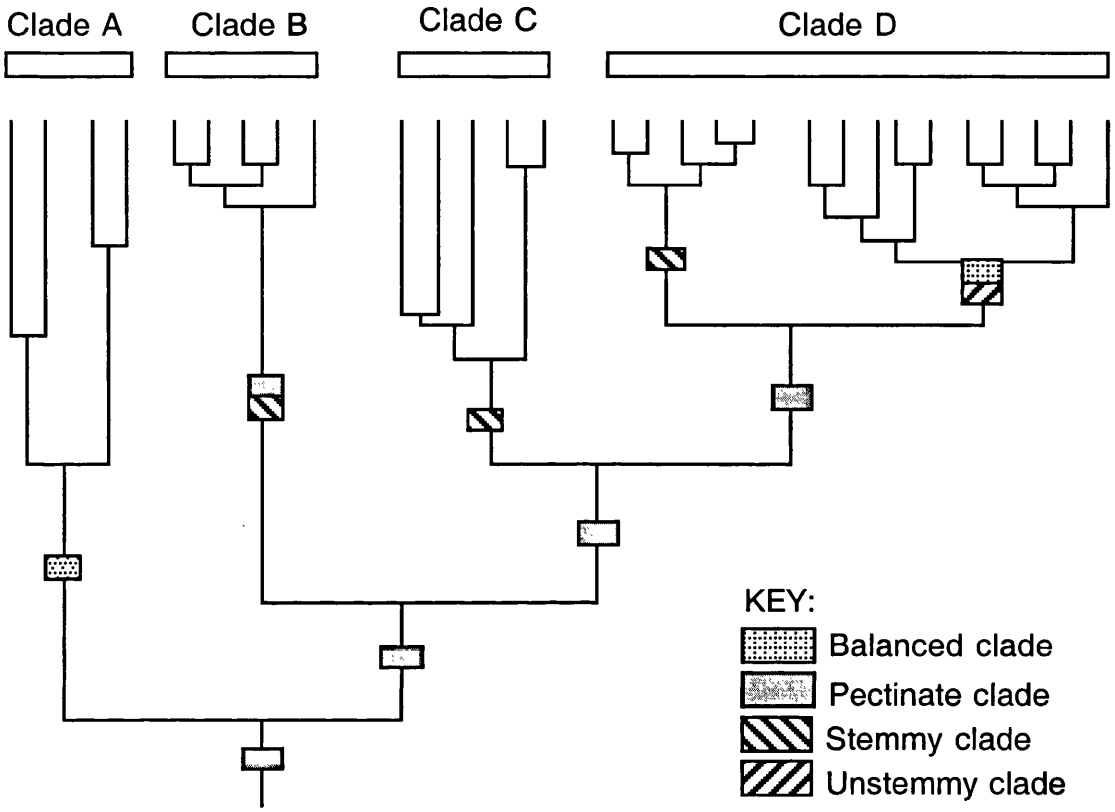
i. Balance: Balance correlates with relative diversification or extinction rates in different parts of the cladogram. In a balanced tree, either the rates are the same in all clades, or one clade may have a higher speciation rate AND a higher extinction rate (or vice versa) than the others. In a pectinate tree, rates of speciation and extinction can vary independently in different clades.

The balanced topology represented by the tree in Figure 1.2 is often regarded as the norm, and model systems have aimed to reconstruct this topology (e.g. Hillis et al., 1992). For any given scenario, however, it seems unlikely that every lineage should speciate as often as every other (producing a fully balanced tree); the reverse situation, where only one of each lineage pair speciates (producing a fully pectinate tree, as in Figures 1.4 and 1.5) is also improbable. Certainly some sort of favourable mutation (like a 'key innovation') along one branch of a tree could cause it to speciate more than other branches, thus causing some asymmetry. Kirkpatrick and Slatkin (1993) suggest one scenario which could lead to a symmetric tree: where there is "synchronous speciation caused by vicariance events that affect most of all of the species in a clade" (Kirkpatrick & Slatkin, 1993, p. 1179). However, published phylogenies are seldom either fully balanced or fully pectinate.

ii. Stemminess: Where trees contain more and less stemmy clades, the more stemmy clades have built up a series of unique characters, either through remaining undiversified for a long time (perhaps through a period of climatic or geological stability), or through the extinction of more basal members (the Relict Hypothesis).

iii. Hypothetical example:

Figure 1.9: Hypothetical phylogenetic tree



With a phylogenetic tree, we can make some inferences about the relative ages of clades. In Figure 1.9, clade A is old compared to clade D, and can be described as basal relative to it. Sister groups are always the same age - clade A is the same age as clade (B, C, D), clade B is the same age as (C, D) and clade C is the same age as clade D.

One could hypothesis that the smaller number of taxa in clade C (5) than in the sister group D (15) is due to some form of key innovation (e.g. the angiosperm flower) in the ancestor of clade D. A beneficial morphological innovation (such as the angiosperm flower) is liable to produce an unbalanced tree where the clade which possesses the innovation has a greater rate of speciation (or lower rate of extinction) than the clade which is without it.

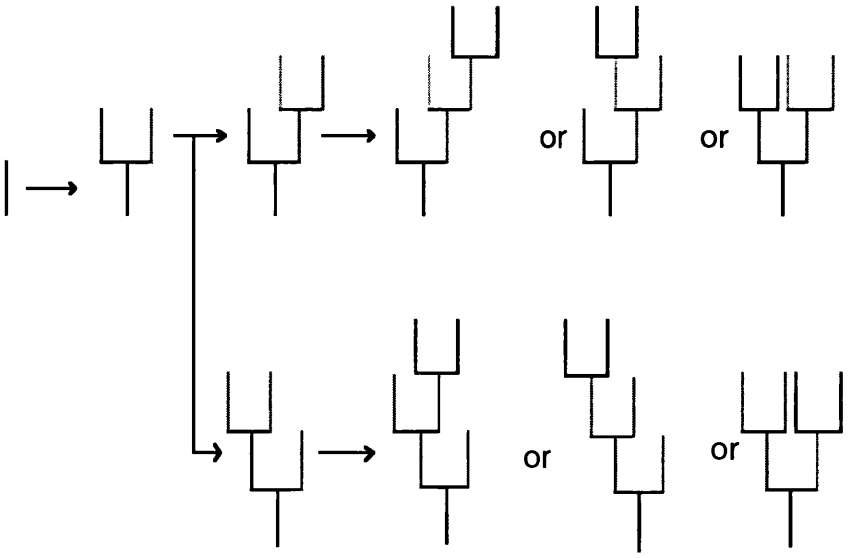
Phylogenetic trees which contain long, undivided branches interspersed with short branches (e.g. clade C) are notoriously difficult to reconstruct. Such

topologies exist when rates of evolution vary greatly among lineages, or where the timing of cladogenesis varies (Hillis & Bull, 1993), or where there is a lot of extinction in some lineages.

C. Caveats: There are several confounding variables which contribute to tree shape but tell us nothing about evolutionary process. Topology is due to a combination of three factors - noise (stochastic effects), bias (analytical method, taxon choice and definition) and signal (macroevolutionary process) (Mooes & Heard, 1997). We need to be able to isolate the effects of signal from those of noise and bias.

Guyer and Slowinski (1993) point out that, because phylogenetic trees must take some shape, species-rich clades can appear which require no adaptive explanation, and so a null model must be invoked to tell us whether our observed tree shape deviates from what is expected. The simplest null model of evolution is the Markovian model, which involves equal rates and random speciation (Heard, 1992). The growing branches of a phylogeny diverge at random. One species can progress to four terminal branches by six routes (Figure 1.10).

Figure 1.10: The Markovian model of evolution



Four of these routes produce a pectinate tree; two produce a balanced cladogram. Thus the null probability of a pectinate tree is 4/6 (0.67), and of a balanced tree, is 2/6 (0.33) (Guyer & Slowinski, 1993). The Markov model is only one of several which can be used; another is the Proportional-to-distinguishable-arrangements model, wherein each possible tree is assumed to be equally likely. Given four terminals (and allowing all combinations of taxa across the terminals), there are 15 possible arrangements. 12 of these are pectinate, which makes the null probability of a pectinate tree 12/15 (0.80) (Guyer & Slowinski, 1993).

Studies which have sought to recover trees from simulated data have revealed that phylogenetic methods can bias the shape of the recovered trees towards asymmetry, particularly in cases where the (simulated) evolutionary rates are high (Huelsenbeck & Kirkpatrick, 1996). Heard (1996) also found that, under most simulations and using different models of speciation rate variation, the amount of imbalance increases as speciation rate increases. However, various studies (reviewed in Mooes & Heard, 1997) indicate that although noise and bias do contribute to tree shape, macroevolution also plays a part and can be invoked in the explanation. Thus the prevalence of pectinate trees in the literature (Pearson, 1999) is not an entirely artificial phenomenon but reflects some natural pattern. Pearson (1999) explains it as "the result of the relative success of apomorphic taxa over their more pleisiomorphic sisters, where success is measured in terms of resistance to extinction or propensity for speciation" (Pearson, 1999, p. 405).

D. Summary: It is important to remember that sister clades the same size are no more likely than sister clades of very different sizes (in fact, under the two models discussed here, they are far less likely); a null model must be invoked in order to distinguish any significance in observed differences (Mooes & Heard, 1997). The simplest model is the Markovian model of equal rate random speciation (Heard, 1992); departure from this can be tested by evaluating whether each internal node in a tree is balanced or unbalanced (Bond & Opell, 1998).

1.4.3 Are big genera real?: One question which should be answerable given the relevant phylogenies is, 'Are big genera real?' If similar cladogram shapes and sequence divergence rates could be found for groups with very different hierarchical positions, we could conclude them to be a case of inconsistency in the application of taxonomic rank.

However, there are several variables which complicate such comparisons. One is that sequence divergence rates are known to differ in different groups. For example, in the monocots, *rbc*L is fastest in grasses and slowest in palms (Gaut, 1998). Another is that decisions about plant taxonomy are (or have traditionally been) made on the basis of plant morphological, not of molecular, distinction, and morphological and molecular evolution are uncoupled. Indeed, Bateman (1999) suggests that both track different facets of evolution, with "the vast majority of morphological character-state transitions occur[ing] during speciation events and the vast majority of molecular character-state transitions occur[ing] between them" (Bateman, 1999, p. 446).

Ideally, we would have a same-gene phylogeny for two related groups with similar life history characteristics, together with comparable morphological phylogenies to guard against differential rates of morphological evolution due, for example, to key innovations. This (perhaps rather unlikely) combination of data sets would allow comparison of the taxonomic treatment of two groups; for example, one could throw some light on whether the smaller average genus size of Poaceae (668 genera, 9500 species) over Cyperaceae (98 genera, 4350 species) **is** real or is an artifact of folk taxonomy, as has been suggested previously (figures from Mabberly, 1998).

1.4.4 Are big genera old or young?: The question of whether large genera are ancient (as expected from the Age and Area hypothesis), or the results of bloom phase groups (as expected by the Relict hypothesis) should be testable by two methods: one, by the physical evidence of when taxa appear in the fossil record, and the other, by evidence from phylogenetic trees.

A. Fossil Record: The Plant Fossil Record (http://ibs.uel.ac.uk/palaeo/pfr2/pfr.htm) only lists fossil records for 23 of the 71 large vascular plant genera listed in the Appendix (14.3). 9 genera have records which predate the Eocene (*Acacia*, Palaeocene; *Asplenium* L., Cretaceous; *Diospyros* L., Cretaceous; *Eucalyptus* L'Herit., Cretaceous; *Ficus* L., Palaeocene; *Ilex* L., Cretaceous; *Litsea* Lam., Cretaceous; *Quercus* L., Palaeocene, *Selaginella* Pal., Cretaceous). That no records were found for *Huperzia* Bernh. does not, however, indicate a lack of fossils; a thorough search would need to include any relevant form genera. There are two problems with interpreting this fossil evidence as evidence of the age of genera.

Firstly, it is noteworthy that all seven of the angiosperm genera listed above are predominantly trees. This is more likely to reflect a bias in the fossil record (such as: larger plants, producing huge quantities of pollen; woody tissue preserving better than, for example, succulent tissue) than any natural reality. Secondly, evidence about the age of a lineage is not the same as evidence about the age of the species which make up the genus today. A lineage may be very old, yet be made up entirely of young species. The fern *Asplenium*, despite having a fossil record, does not belong to a very old lineage compared to other ferns (M. Gibby, pers. comm., 2000; R.M. Bateman, pers. comm., 2000). *Selaginella* has a fossil record which dates back to the lower Carboniferous, but although the genus is old its component species are young (R.M. Bateman, pers. comm., 2000). The conifer genus *Araucaria* Juss. (Araucariaceae Henkel & Hochst) is old; there are good fossil records for the Jurassic. However, all 13 species on New Caledonian have near-identical *rbc*L sequences (pairwise differences between 0 and 0.5%, with 10 of the species having identical sequences) (Setoguchi et al., 1998), suggesting that they may be the results of comparatively recent speciation events. Setoguchi et al. (1998) suggest that rapid differentiation of *Araucaria* occurred after the Eocene, when a large part of New Caledonia, with predominately ultramafic soils, was formed. Therefore even in what are ancient lineages (*Selaginella* and *Araucaria*), there is evidence that the extant species are comparatively recent.

What is required to show the existence of ancient large genera is a large

genus with a fossil record which shows high taxon diversity over a relatively long time period.

B. Clade position: As more genus-level phylogenies are being produced, the relative ages of larger genera can be estimated from their positions on the phylogeny.

Begonia has a derived position (relative to the far smaller families Datiscaceae Bercht. & J.Presl., Corynocarpaceae Engl. and Coriariaceae DC) in the *rbc*L phylogeny of Wagstaff and Dawson (2000). In the combined *atp*B - *rbc*L trees of Savolainen et al. (2000), none of the larger genera included in the analyses take obviously basal positions. Also, in the 589 taxon *rbc*L phylogeny (Savolainen et al., 2000) many large genera including *Salix* L., *Passiflora* L., *Euphorbia*, *Viola* L., *Ficus* and *Rhododendron* all appear in derived positions. Likewise, *Astragalus* does not take a basal branch within an ITS phylogeny of temperate herbaceous legumes (Sanderson & Wojciechowski, 1996). Although broad generalisations are unlikely to apply to all large genera, the evidence from the position of these taxa in wider phylogenies suggests that many large genera are not very old.

If the diversity in a large genus is very young, one would expect the extant taxa to show little sequence divergence, while if the diversity is old, there would be high levels of divergence between extant taxa. Azuma et al. (2000) used *rbc*L to examine relationships between *Salix* species. Their phylogeny produced a polytomy with zero-length branches for 9 of the 19 *Salix* species they included; 9 other species were also on zero length branches. This result suggests that it is more likely that the species in this lineage are comparatively recent, and as such, disproves Willis' Age and Area hypothesis (Willis, 1922), that older taxa tend to be more speciose.

1.5 Biological factors

Before moving on, it is worth considering the biology which gives rise to the shapes of trees. The evolutionary patterns we reconstruct in our phylogenies are essentially the result of two processes, speciation (diversification) and extinction. Naturally large genera may be large because they have an above-average rate of speciation, or a below-average rate of extinction (or a combination of both). The most vital tool to test hypotheses about these patterns is phylogeny.

1.5.1 Diversification: The success of a large clade is often attributed to one or more 'key innovations' unique to that lineage (Bond & Opell, 1998). For example, changes in the jaw musculature of cichlid fish have led to their highly speciose lineage (Nee & Harvey, 1994). Dodd, Silvertown and Chase (1999) argue that if such 'key innovations' have opened up new adaptive zones and led to increased proliferation of species, it should be possible to identify certain clades which are more species rich than others and wherein this richness correlates with the presence of "traits that influence speciation and extinction" (Dodd, Silvertown & Chase, 1999, p. 732). A key innovation is not, however, a prerequisite to an adaptive radiation; Nee and Harvey (1994) point out that there was not necessarily anything special about the first finch to reach the Galapagos islands.

Sanderson and Wojciechowski (1996) used ITS to look at diversification rates in *Astragalus*, assessing whether increases in diversification rate within a wider group of Fabaceae coincided with the origin of *Astragalus* and if so, whether diversification could be tied in to identifiable 'key innovations'. *Astragalus* has several features which have been considered to promote diversification in angiosperms, namely geographic population structure consisting of local isolates with restricted gene flow; herbaceousness (correlating with reduced generation time); chromosomal variability; a tendency to parallelism and reversal associated with ecological specialisation; and lastly, morphological novelties (Sanderson & Wojciechowski, 1996). However, Sanderson and Wojciechowski (1996) found no significant difference in diversity between *Astragalus* and its sister groups; they did however find a significant increase in diversification in the branch which leads to this

Astragalean clade (*Astragalus* and its sister groups, which together comprise about 3000 to 3500 species of generally xerophytic, generally herbaceous perennials; often with high degrees of endemism). Thus there may well be no key innovations to be found for *Astragalus* alone; given the tree shape it seems more likely that the key innovations would belong to the Astragalean clade in its entirety, although no obvious morphological novelties coincide with the origin of the Astragalean clade or its increase in species diversity. In this clade, "similar morphological adaptations to extreme environmental conditions have evolved countless times in parallel ...[therefore]... species represent endless variation on an essentially constant ground plan" (Sanderson & Wojciechowski, 1996, p. 1499).

Using a molecular clock for ITS, Wojciechowski, Sanderson and Hu (1999) estimated an age of 11 Myr (t) for *Astragalus* (c. 2500 species (N)), giving an average lineage diversification rate of 0.71 spp/Myr (estimated as ln N/t). They argue that this rapid rate of lineage diversification (compared to an estimated median value of 0.12 spp/Myr for continental plant families - Erikkson & Bremer, 1992), combined with conventional (or low) rates of morphological evolution, is what has given rise to this large genus.

Dodd, Silvertown and Chase (1999) were also interested in key innovations and diversification. They found that species richness in the angiosperms correlates with evolutionary changes in pollination and in growth form, although not consistently with changes in dispersal mode. (The loss (not the gain) of biotic pollination is almost ubiquitous, leading to the suggestion that, as a significant number of these losses of biotic pollinators has been accompanied by a subsequent fall in diversification, this could be used as a positive test for the theory that it was the early evolution of biotic pollination which was responsible for the original diversification of the angiosperms.)

However, diversification may not be due to any intrinsic biological property of species. In geologically active areas, for example, higher speciation rates may be an incidental side effect of biogeography (habitat fragmentation leading to vicariant speciation) (Kirkpatrick & Slatkin, 1993). Many clades in plant phylogenies have geographical correlations; scenarios which involve

increased speciation or increased extinction in some areas (therefore in some clades), for example due to climatic change and / or differences in sea level, are not difficult to imagine.

1.5.2 Extinction: Species can either die gradually, as the population dwindles and fails to respond to selective pressures, or suddenly, with the extinction of all individuals due to an environmental crisis which is beyond prior experience (Niklas, 1997). Because mass extinctions allow for only 4% of all extinctions in the fossil record, the geologically sudden loss of formerly successful species does not seem to be the most frequent mode of extinction. Although well over 90% of all the species which ever lived are extinct, the morphological and taxonomic diversity present today represents a surplus of species birth and survival over death; the taxonomic composition of the Earth's flora has changed dramatically in 450 million years (Niklas, 1997). (The obvious taxonomic bias in the groups suffering extinction should deflate the otherwise depressing sampling problems for phylogeny reconstruction!)

Extinction is a very difficult hypothesis to prove from a cladogram alone; many factors can lead to taxa appearing isolated on long branches - Pfosser and Speta (1999) found such a pattern in Hyacinthaceae Batsch ex Borckh. (a high number of nucleotide changes before speciation occurred within the subfamilies, or 'stemminess'), which they explain as due to either a) a higher rate of substitutions during this initial radiation, b) sampling bias because of extinction, or c) primary radiation occurring very slowly (analogous to the Punctuated Equilibrium hypothesis of Eldredge and Gould (1972) and the Turnover-pulse hypothesis of Vrba (1985) which states that "[e]volution is normally conservative.....[t]hus most lineage turnover in the history of life has occurred in pulses....in predictable synchrony with changes in the physical environment" (p. 232)). When there is fossil evidence it becomes easier to choose between such alternatives.

The fossil record for the genus *Ilex* L. (Aquifoliaceae) goes back 90 million years and the genus appears to have been cosmopolitan long before the end of the Cretaceous (Cuenoud et al., 2000). When comparing mean rates of nucleotide substitution for *rbc*L for *Ilex* with *rbc*L substitution rates given by

other authors for other taxa, the rates for *Ilex* are low, and a divergence time for the outgroups (*Helwingia* Willd. and *Phyllonoma* Willd.) is estimated as at least 198 Mya. Cuenoud et al. (2000) felt this age to be excessive. They used the relative test of the rate of nucleotide substitution (Wu & Li, 1985) to test whether a) the rate of substitution in *Ilex* is low compared to other lineages or b) the extant species only represent part of the 90 My old lineage, i.e. whether divergence rates indicate that the common ancestor of all the extant species of *Ilex* is 90 million years old, or whether the common ancestor of the extant species is more recent. Because the Aquifoliaceae do not appear to have diverged more slowly than their near relatives *Helwingia* and *Phyllonoma,* Cuenoud et al. (2000) suggest instead that the basal branches of the lineage are extinct (although equally there could have been little or no speciation in the early lineage). Evidence from the fossil record suggests that much diversification in *Ilex* occurred in the Eocene; this may be the time of ancestry for all extant species and so a complete study of the *Ilex* lineage would have to include fossil evidence (Cuenoud et al., 2000) as there is no other way around the sampling bias (due to extinction) among extant taxa.

1.5.3 Distribution: Diversification and extinction deal with the distribution of species richness through time. Phylogeny can also be used to uncover patterns of taxonomic richness in space[2] . Large genera are often widespread. These patterns of taxon distribution are attributable to dispersal ("the movement of an organism from one area to another independent of other organisms and of earth history, which changes the natural distribution of the organism"; Humphries & Parenti, 1999, p. 172) and vicariance events (the splitting of a taxon or biota into two or more geographical subdivisions by the formation of a natural barrier such as mountain building, glaciation, stream capture", Humphries & Parenti, 1999, p. 174) and may be complicated by extinction. Both dispersal and vicariance may be followed by radiation.

[2] There is a danger in using a non-phylogenetic approach to such studies - monophyly is a necessary criterion. For example, phylogenetic analysis of Hyacinthaceae revealed that the American genus *Camassia* Lindley. belongs in its own unrelated family, Camassiaceae (Pfosser & Speta, 1999). Biogeographers no longer have an unusual disjunction between the north American endemic *Camassia*, and the monotypic Chilean *Oziroe* Rafin. to explain. In *Eupatorium*, redefining the genus according to phylogeny also removes a geographical disjunction from within the genus (Schmidt & Schilling, 2000).

Certainly the vicariance hypothesis is more applicable to older lineages, as it is most cited in instances of continental drift. Phylogeny can help chose between these explanations.

A. Dispersal: The phylogeny of *Astragalus* produced by Wojciechowski, Sanderson and Hu (1999) allowed them to reject a previous hypothesis of vicariance, that two disjunct groups of *Astragalus*, one New World and one Eurasian, had been separate since the Tertiary and undergone independent, parallel evolution. The Old World *Astragalus* are not monophyletic; New World taxa nest clearly within them. The genus appears to contain more recent (Pleistocene to late Pliocene) immigrants to North America via a Beringian land bridge.

In the phylogeny of *Aeonium* Webb & Berth. (Crassulaceae J.St-Hil.) the distribution pattern is explained entirely by dispersal rather than vicariance (Jorgensen & Frydenberg, 1999). This genus of succulent, rosulate species is considered a prime example of adaptive radiation in an island plant group. Most taxa from the same islands did not form monophyletic groups. This has led the authors to believe that colonisation of similar ecological zones on different islands followed by divergence has been important in the speciation of these plants.
Coriaria L. (Coriariaceae DC) has a strikingly disjunct geographical distribution; it occurs in the Mediterranean, continental and insular eastern Asia, from Papua New Guinea to New Zealand, and from northern Mexico to southern Chile (Yokoyama et al., 2000) . There have been many suggestions as to the cause, including: migration north from the southern hemisphere via the Pacific islands (dispersal); habitat disturbance by Tertiary glaciers (within hemispheres) and vicariance between hemispheres caused by continental drift; and rafting north from Gondwanaland on the Indian plate (vicariance).

Yokoyama et al. (2000) produced a phylogeny to test these hypotheses. They were able to rule out northern expansion via the Pacific Islands, and instead postulate an Eurasian or North American origin for the genus. They also found it unlikely that *Coriaria* migrated from South to Central America. Fossil evidence suggests that the genus was more widely distributed in the past, at

least in Eurasia. They estimated the divergence time for the main clades to be c. 60 My ago (early Tertiary), when Eurasia and North America were closer and the Arctic region was warm enough to support temperate species. The common ancestor for these clades could have expanded its range through this region. Thus the present disjunct range may well have been caused by the climatic changes associated with glaciation and drying out during the Cenozoic. One lineage may have migrated into North America after the land bridge from South America formed, while another dispersed to the Pacific Islands and radiated in Papua New Guinea. Central America and the Pacific Islands were not connected in the Cenozoic, leading Yokoyama et al. (2000) to postulate long distance dispersal.

B. Vicariance: Pfosser and Speta (1999) suggest a southern Gondwanic origin for the Hyacinthaceae, because South African, South American and Madagascan species occupy the basal branches in their phylogeny. Direct migration was possible between these landmasses and India until the mid-Cretaceous (c. 100 Mya). The distribution of species within one clade (Africa south of the Sahara and the Indian subcontinent) suggests that the initial diversification within the family occurred while India was still connected to southern Africa. Because another clade has members in the Mediterranean and in Eurasia but not in North America, they suggest that this clade diversified after North America separated from Eurasia, which may explain why no Hyacinthaceae are found in North America, despite its climate being suitable for the family (Pfosser & Speta, 1999).

However, the phylogeny of extant species cannot always provide biogeographical answers. When Cuenoud et al. (2000) considered the geographic distribution of *Ilex* they found that they could not distinguish between Asia or South America as its area of origin. North America has been colonised from East Asia, South America or both areas. Africa and Europe have been colonised relatively recently from East Asia. As they have dated the ancestry of the extant species as Tertiary, they point out that we cannot expect to find evidence for a Gondwanan origin from molecular data alone in this genus.

1.6 Summary

The recognition of the phenomenon of the 'hollow curve' predates widespread used of phylogeny reconstruction, and many of the hypotheses produced to explain it have become, if not redundant, at least resolvable in the light of phylogenetic treatment. That the answers to questions about the application of taxonomic rank are not apparent from the literature reflects a shift in focus towards subjects like the identification of promoters of diversification (e.g. 'key innovations').

Looking at the shapes of phylogenies and at evidence from the fossil record, it appears likely that most of the species in the larger genera are the results of (comparatively) recent radiations, which may, however, bear no correlation with the ages of the lineages (in this instance, genera) themselves. Shapes and branch lengths of phylogenies can be used to answer a number of biological questions about species richness and distribution - whether radiations have occurred, if key innovations can be identified, if extinction is a likely explanation for isolated clades, and whether dispersal or disjunction account for present-day ranges. However, the answers we receive can only be as good as the phylogenies we produce; it is important to consider whether non-biological factors or homoplasy have affected our hypotheses of relationships.

2. Using Molecules to Reconstruct Evolutionary History

2.1. Why morphology is not enough

It is becoming increasingly apparent that small changes in single genes can be responsible for major shifts in plant morphology. Significant reorganisations of genomes can have little to no effect on the appearance of organisms, while dramatic morphological changes can result from what appear to be minor genic or chromosomal alterations. Thus morphological cladistic analyses can give hypotheses which differ radically from those based on molecular data. This can be regarded as an example of 'mosaic evolution' - "the ability of different characters to evolve at different rates and in different directions" (Niklas, 1997, p. 350).

For example, the gene *cycloidea* encodes a protein which causes bilateral symmetry in flowers (by acting on only the upper parts as the flower develops). If this gene is inactivated by even a single nucleotide substitution, flowers become peloric (Citerne & Cronk, 1999). Furthermore, if characters used in morphological phylogeny and / or classification are subject to strong selection pressures, recurrent evolution and similar forms can lead to misleading inferences of affinity. A common example of this relates to the evolution of pollination syndromes. For instance, within *Ipomoea* L. (Convolvulaceae Juss.) (Miller, Rausher & Manos, 1999) there have been multiple shifts in pollination syndrome, from bees to birds (associated with gain of red pigment in the corollas). There are also numerous independent shifts from pigmented to white flowers. Clearly classification based on floral similarity in this case would not reflect evolutionary relationships. If minor mutations causing such major shifts prove to be a common pattern in plant evolution, and where such genes belong to gene families, mosaic evolution may be found to riddle morphology-based phylogenies. In contrast, molecular data gathering and analysis is becoming increasingly rapid and cost effective and appears to generate usable (and apparently predictive) phylogenies.

2.1.1. Contrast between molecular phylogenies and traditional classification: A number of authors have found that the results of their molecular phylogenetic analyses are incongruent with traditional classifications. In *Viola*, ITS sequences uncovered a relationship between an Hawaiian clade and an amphi-Beringian complex that was not evident from morphological or cytological data (Ballard, Sytsma & Kowal, 1998). The previous placement of the Hawaiian group, with two neotropical sections, was based on close morphological similarity; the shared traits (the branching pattern, the woody stems, the leaf shape, the short corolla spur and the simple style) map on the phylogeny as being homoplastic, and appear to be a remarkable example of convergence between montane plant groups.

Pfosser and Speta (1999) describe problems with morphological circumscription in Hyacinthaceae, because characters which are useful in other families are often highly variable among closely related species, e.g. the type of embryo sac varies within *Scilla* L. s.s. Their phylogeny, based on *trn*L and *trn*L-*trn*F sequence data, disagrees with most of the traditional morphological treatments. In their opinion, "[f]or no plant family is it more true than for Hyacinthaceae that the interpretation of single morphological characters resulted in highly erratic classifications when delineating tribal and subfamilial relationships. No character, from bulb morphology to pistils or seeds, or even karyological data, has proved reliable" (Pfosser & Speta, 1999, p. 865).

In *Ilex*, the phylogeny inferred from *atp*B-*rbc*L spacer chloroplast sequence data is not congruent with traditional systematics: the infrageneric classifications of Loesener (in Engler & Prantl, 1942) and used by Hu (who published in 1949, 1950 and 1967) show "only a few examples of agreement with the molecular phylogeny" (Cuenoud et al., 2000, p. 121). Molecular phylogenetic analysis of *Ipomoea* (Miller, Rausher & Manos, 1999) also failed to support any previous subgeneric classification.

Similar problems occur in other taxa. Ro, Keener and McPheron (1997) used 26S rDNA to estimate a phylogeny for the Ranunculaceae Juss. Although chromosome number and karyotype are consistent with this tree (as are

chloroplast restriction site data and sequence data from three other genes (*atp*B (cp), *rbc*L (cp) and 18S (nr))), fruit type, which has been considered critical in subfamilial classification within Ranunculaceae, is not. Fleshy fruits have evolved in two, and achenes have evolved in at least three, independent lineages.

2.1.2. Contrast between molecular and morphological phylogenies: Baker, Hedderson and Dransfield (2000) found that their molecular phylogenies are not very congruent with previous morphological phylogenies (Baker et al., 1999a). Their molecular phylogeny of subfamily Calamoideae (Arecaceae C.H.Schultz.), based on nr DNA ITS and cp DNA *rps*16 intron sequence data, supported an Asian clade which has "no conspicuous morphological basis" (Baker, Hedderson & Dransfield, 2000, p. 213).

Watson, Evans and Boluarte (2000) produced a molecular phylogeny, based on cpDNA *ndh*F sequences, for Anthemideae (Asteraceae), to compare with the morphological phylogeny produced by Bremer and Humphries (1993). Although the molecular data from their study are congruent with ITS and chloroplast DNA restriction site data, and with the biogeography of the taxa, the molecular phylogeny (as in the Calamoideae example) is 'in general' incongruent with the morphological phylogeny and with all previously proposed classifications for the tribe.

There are now numerous examples of disagreement between molecular data and morphological data. In some cases, reexamination of morphology in the light of a molecular phylogeny has allowed reciprocal illumination and the identification of new morphological characters. In addition, while individual characters can be homoplastic across a given data set, they may be locally informative and can be useful at lower taxonomic levels (Pennington, 1995). This being said, there are also increasing reports of cryptic clades, well supported by molecular data but with no identifiable morphological characters (e.g. Richardson et al., 2000). Assuming these molecular phylogenies represent species phylogeny (see later for a discussion of the potential complicating factors) this indicates that under some circumstances morphology can be misleading.

2.2. Molecular phylogenies

The use of molecular data in phylogenetic reconstruction is now commonplace, with direct DNA sequencing the most widely used character source (Soltis & Soltis, 1998). There are many DNA regions (coding and non-coding, transcribed and untranscribed) available for sequencing; areas can be selected with different evolutionary histories, from different genomes, and with different rates of change. In 1999, papers in Systematic Botany used the following regions for DNA sequencing:
Chloroplast: *mat*K (gene), *ndh*F (gene), *rbc*L (gene), *rpl*16 (intron), *trn*L intron, *trn*L-*trn*F (intergenic spacer).
Nuclear: *Adh* (gene, alcohol dehydrogenase, low copy number), *vicilin* (gene, seed storage proteins, low copy number), *waxy* (gene, starch synthase; single copy), 18S (nuclear ribosomal RNA gene; high copy number), ITS (nuclear ribosomal RNA gene transcribed spacer; high copy number).

2.2.1 Which gene for which question?: The three genomes of plants offer genes with differing characteristics and tempos of evolution. The slowest substitution rates of all eukaryotic genomes are found in plant mitochondrial DNA (Li, 1997); chloroplast DNA has a substitution rate of about four times that of mitochondrial DNA and nuclear DNA has a 10-fold increase. However, these are broad generalisations based on a narrow range of genes; it would perhaps be more informative to consider commonly used genes individually.

Perhaps the most widely used genes in plant phylogeny reconstruction (at least until the late 1990s) are *rbc*L and 18S. The chloroplast gene *rbc*L is alignable over wide phylogenetic distances and has been used to infer the evolutionary history of the angiosperms. Sequences of rDNA 18S have also been used for deep level phylogenetic reconstruction; 18S includes slightly less phylogenetic signal than *rbc*L, but is also widely alignable and has been used to provide corroboration of novel deep level angiosperm-wide clades from a different genome to *rbc*L (Hershkovitz, Zimmer & Hahn, 1999). These genes do not, however, evolve quickly enough to resolve relationships at lower taxonomic levels such as among the species of a genus, or related genera in a family. Instead, more rapidly evolving regions such as the internal transcribed spacers (ITS) of nr rDNA and the chloroplast intron / intergenic spacer of *trn*L

have been widely used as a source of characters at this level. Among very closely related species even these regions are not variable enough, and attention has recently turned to fragment analyses such as AFLPs or RAPDs, or sequence data from introns of low copy number protein encoding genes, in the search for variable characters.

It should be noted that the concept of speed of genes is not straightforward with regard to the most suitable level for their application. It is clear when a gene is too slow, as few or no variable characters are present. Determining when a gene is too fast for the question in hand is more difficult. In some cases, where the taxonomic distance is too large, the sequences may simply not be alignable; without a satisfactory alignment, any phylogenetic hypothesis is hard to justify. However, for those rapidly evolving genes such as *mat*K, performance over deep levels of evolutionary history can be better than one might have predicted. For instance, the APG (V. Savolainen pers. comm., 2000) have, by combining data sets for several genes, produced what they take to be the closest tree to a 'true' phylogeny for the angiosperms and used this topology to examine the relative contributions of the component genes. The fastest gene they sampled (*mat*K) produced the 'best' tree (it had most nodes in common with the 'true' tree). However, one must note that the number of nodes in common, on its own, requires some qualification. The majority of nodes on a tree are near the terminals, and so a tree from a fast gene could be predicted to most approximate a 'true' tree using this criterion. Using a slower gene however, one would expect to resolve the deeper branching patterns. In order to obtain good resolution at both the distal and the basal nodes, the ideal gene / region would include both comparatively conserved and divergent sequence.

However, Savolainen et al. (2000) point out that rate *per se* is not a reliable explanation of how well a gene or region will perform in phylogenetic reconstruction; a better explanation is 'decisiveness'. Although more rapid regions may have more homoplasy, they may also have more signal therefore be more 'decisive'. Homoplasy is only a problem in phylogenetic reconstruction if it covaries (Chase & Cox, 1998), otherwise it should be swamped by signal.

Of course, there are two ways of looking at the rate of a gene or region - one, simplistically, is the number of sites that vary along its length; the other is how many changes there are per variable site. While a gene or region may be described as 'fast' because it has a lot of variable sites (and conversely, 'slow' if it has only a few), a gene which has a lot of changes at each variable site may also be described as 'fast' (and vice versa). Each nucleotide in a sequence is not necessarily equally likely to undergo a substitution; for example, in genes or proteins, substitution rates tend to be highest in third codon positions, and lowest in second codon positions (Yang, 1996). This sort of pattern can be modelled using the shape parameter α of the gamma distribution of substitution rates at sites, where a low value for α indicates extreme rate variation among sites, and a high value indicates minor rate variation (when all sites have the same substitution rate, α is infinity) (Yang, 1996).

Where there are many changes at each variable site, multiple hits on the same base are more likely, which increases the potential for homoplasy to obscure phylogenetic signal. Thus a measure of the number of variable sites for a gene or region will not necessarily correlate with the potential levels of homoplasy, if it does not take some account of the number of changes per site.

Even where variable sites are very variable, and multiple hits are thought to be a problem, however, signal is not always obscured. Although previous authors have advised eliminating third codon positions from analyses because they tend to have more changes per site than other positions and therefore more potential for homoplasy (e.g. Kitching et al. (1998, p. 103) find it "rational to downweight or even ignore third position changes"), Kallersjo, Albert and Farris (1999) found that the third codon positions in *rbc*L, although rapidly evolving and highly homoplastic, contain most of the phylogenetic signal in a 2538-taxon green plant matrix. Excluding these regions cuts down the resolving power of the matrix. Likewise, Chase and Albert (1998, p. 495) found that eliminating third positions gives "less resolution and weaker measures of internal support".

In most studies people seek to resolve not only within-clade relationships, but also the deeper relationships of the clades to each other. A matrix wherein

sites evolve at different rates, although problematic for some phylogeny reconstruction algorithms, offers the potential for recovering clades at different hierarchical levels.

2.2.2 Evolutionary rates and molecular clocks: In reconstructing the phylogenetic relationships of organsisms, it is clearly desirable for the date, and not just the order, of branching patterns to be known. A cursory glance at most DNA sequences shows that the more phylogenetically disparate taxa are, the more divergent their sequences tend to be. This has led evolutionary biologists to explore the concept of the molecular clock - using sequence divergence to estimate times of separation. The molecular clock hypotheses are based around the Neutral Theory of Molecular Evolution (Kimura, 1968). The beauty of the neutral theory is the prediction that substitutional change over time is affected only by the mutation rate. Providing the mutation rate is constant across the lineages considered in the study group, there is an expectation that sequence divergence will be linearly related to time. One confounding variable is that of generation time; longer generation times are predicted to lead to slower divergence. This is attributable to fewer meiotic events per unit absolute time, which reduces the opportunity for mutation. However, in plants this is complicated as mutations may be fixed in vegetative meristems as well as reproductive cells, and seed banks and clonal reproduction may have a stabilising effect on the evolutionary rate in herbs (Baldwin et al., 1995).

In addition, in a less simplified extension to the Neutral Theory, the Nearly Neutral Theory predicts that many mutations will be slightly deleterious (Ohta 1973). If this is the case, then population size will also affect rates of change, due to the inefficiency of selection in small populations compared to large ones.

Various other confounding variables have also been postulated (Gaut, 1998) and in practice, estimates of relative nucleotide substitution rates among evolutionary lineages for *rbc*L and for ITS have shown that there are no time-calibrated clocks for these regions (Bousquet et al., 1992; Gaut et al., 1992; Baldwin et al., 1995).

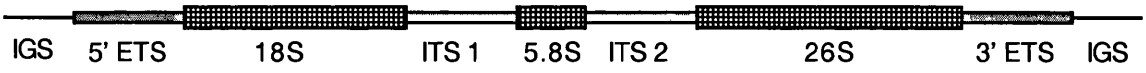
However, although the issue of whether percentage divergence between a pair of sequences can be related in some way to the time since the lineages split is still in many respects under debate, clocks calibrated by known events (e.g. fossil evidence or geological history) are appearing increasingly in the literature (e.g. Wagstaff & Dawson, 2000, dating by fossil records; Richardson, 1999, dating through island appearance). Methods to estimate lineage-specific evolutionary rates and / or divergence times are reviewed by Sanderson (1998).

2.3. Ribosomal DNA

As the current project is addressing patterns of diversification in a large genus, a region which offers resolution at the inter- and intra-sectional level was required. The region I have chosen is a combination of sections of nuclear ribosomal DNA. A summary of the characteristics and evolutionary dynamics of this region are given below.

The reasons why the rDNA cistron is so frequently used in phylogeny reconstruction have been comprehensively reviewed (Soltis & Soltis, 1998, and references therein). In eukaryotes the rDNA cistron encodes the 18S (SSU), 5.8S and 26S (LSU) rRNAs, which are separated by two internal transcribed spacers (ITS 1 and ITS 2) and flanked by the 5' and 3' external transcribed spacers (5' ETS and 3' ETS). There are thousands of copies of the cistron, which are each separated by the intergenic spacer (IGS) (see Figure 2.1).

Figure 2.1 The rDNA cistron



2.3.1 ITS: Many authors have found the internal transcribed spacer (ITS) of nuclear ribosomal DNA to provide useful characters for phylogenetic studies, particularly at lower phylogenetic levels. The utility of this region was first demonstrated by Baldwin (1992, 1993) in the Asteraceae and since then there has been a vast proliferation in studies using ITS (e.g. Yuan & Kupfer, 1997, *Gentiana* L., Gentiaceae Juss.; Li et al., 1999, Hamamelidaceae R.Br.). Baldwin et al. (1995) provide a comprehensive review of the use of the ITS region in angiosperm phylogeny reconstruction.

2.3.1.1 ITS Function: The two spacers, ITS 1 and 2, have different evolutionary histories. ITS 1 is homologous to the SSU-LSU spacer in non-eukaryote and organellar rDNA (Hershkovitz, Zimmer & Hahn, 1999), while ITS 2 is missing in prokaryotes (Clark et al., 1984). 5.8S is homologous to the 5' end of 23S rRNA in *E. coli* (Clark et al., 1984). ITS 1 is not only unrelated evolutionarily to ITS 2, but is also distinct structurally and functionally. However, evolutionary patterns in the two spacers (such as overall rates, and base composition biases) are usually parallel (Baldwin et al., 1995).

The ITS regions are thought to have a role in the maturation of nuclear rRNAs, bringing the large and small subunits close within a processing domain. Deletion of small parts of ITS 1 can inhibit the production of mature small and large subunit rRNAs in yeast, while some deletions or point mutations in parts of ITS 2 prevent or reduce the processing of large subunit rRNAs (references cited by Baldwin et al., 1995). Thus it seems as if there is some evolutionary constrain on the structure and sequence of ITS 1 and ITS 2. Baldwin et al. (1995) suggest that the similarity in G and C content between ITS 1 and ITS 2 reflects a degree of coevolution. The ITS regions are inherently G and C rich (the GC content of angiosperm ITS is almost always over 50% and can in some cases exceed 75% (Hershkovitz, Zimmer & Hahn, 1999)) and have some regions which are highly conserved across the angiosperms (Soltis & Soltis, 1998). Because of the presence of these conserved regions, Hershkovitz and Zimmer (1996) could align about 50% of the ITS 2 region above the family level in angiosperms.

2.3.1.2 Taxonomic level: ITS is considered to be best suited for "diagnosing relationships among closely related genera and infrageneric groups" (Hershkovitz, Zimmer & Hahn, 1999, p. 287). Divergence values between closely related species may be less than 1% (less than 5 substitutions); given that some of these may be autapomorphies, ITS may not offer much information about relationships at the species level (Hershkovitz, Zimmer & Hahn, 1999).

The most rapidly-evolving ITS regions are prone to length variation, which can cause problems with alignment. This means that increased divergence

between species is not always accompanied by an "equiproportional increase in the number of alignable informative sites" (Hershkovitz, Zimmer & Hahn, 1999, p. 287). There is approximately one indel per 2% sequence divergence; however, because the indel positions vary in different lineages, in many cases only a small minority of taxa have an indel at any one site (Hershkovitz, Zimmer & Hahn, 1999).

2.3.1.3 Secondary structure: A potential problem with analysis of ITS is non-independence of nucleotide sites. Estimated secondary structures of ITS 2 for species in the Asteridae (Baldwin et al., 1995) showed high levels of similarity, with a three-stem / loop structure. Mutations of positions along the stems may need compensatory mutations at the opposite sites to maintain structural integrity. This has led to authors suggesting that stem sites should be weighted over loop sites in analyses (Dixon & Hillis, 1993; Wheeler & Honeycutt, 1988; Baldwin et al., 1995), and that compensatory mutations should be downweighted (Hershkovitz & Zimmer, 1996), although such weighting schemes can be too constrictive. Soltis and Soltis (1998, p. 205) say: "most efforts to weight stem versus loops and transitions versus transversions, and even conserved versus variable domains, are probably not worth the effort or extra computer time required to conduct some of the analyses", and Kitching et al. (1998) suggest that any such weighting scheme may not be generalisable to all organisms and all molecules but should be investigated with reference to each individual study group and molecule.

Mai and Coleman (1997) produced secondary structures for ITS 2 for several green algae, and also for some angiosperms. They searched aligned sequence data for covariants (compensating base changes which change so as to maintain base pairing). Most of the ITS 2 region appears to be a self-contained folding complex, usually with four distinct hairpin loops. Highly conserved regions within ITS 2 (116 positions) were readily alignable across all of the Volvocales (algae). 85.3% of these positions fell into regions which base-pair. Despite considerable length heterogeneity, conserved structural elements consistently form. Mai and Coleman (1997) also found close structural correspondence between ITS 2 from the Rosaceae Juss. and Volvocales, although this was not due to similarities in nucleotide sequences.

Variations in distal portions of the hairpins, however, occur even among interbreeding organisms (Mai & Coleman, 1997).

Hershkovitz and Zimmer (1996) also looked for conserved regions of ITS 2; they identified six regions which are conserved across a wide range of angiosperms. "The combination of angiosperm-wide sequence conservation with species-level sequence variability renders ITS a unique window for examining the behaviour of a rapidly-evolving, homologous, non-coding DNA sequence through divergence times spanning relatively ancient (90-130 million years) to the most contemporary" (Hershkovitz & Zimmer, 1996, p. 2866).

Coleman et al. (1998) examined the secondary structure of ITS 1, and considered its use in primary sequence alignment. In comparison with their previous work on ITS 2 (Mai & Coleman, 1997) they found a significantly greater level of primary sequence divergence in Volvocalean ITS 1 sequences; they did not find any regions of conserved primary sequence across the family or order. The ITS 1 sequences were most useful at population and species levels although, in their more conserved portions, they contribute information up to the family level. ITS 2 provided information at higher taxonomic levels.

2.3.1.4 Intra-individual Polymorphism: In practice, the problem of polymorphism between the multiple copies of ITS is not whether it exists (it does, and can be demonstrated by cloning), but whether it can mislead phylogenetic analyses. Hershkovitz, Zimmer and Hahn (1999) suggest that it does not: ITS phylogenies are usually congruent with independent evidence. The levels of divergence are low between closely related species and paralogues will probably not support the incorrect tree as they are not differentiated enough; at greater taxonomic distances, homogenisation will fix the differences (Hershkovitz, Zimmer & Hahn, 1999).

2.3.2 5.8S: Much of the variation in the 5.8S gene is in a 24-base helix close to the 3' end. The overall variability of the gene is low, but Hershkovitz, Zimmer and Hahn (1999) suggest that it may be useful in augmenting 18S and / or 26S.

2.3.3 26S (LSU): The ribosomal large subunit gene is different lengths in different taxa, which means that the homologous region is somewhat confusingly known as 28S in animals, as 23S in prokaryotes and as 26S in plants; the region is also sometimes simply called the LSU.

2.3.3.1 26S Function: Most of the ribosomal large subunit is formed from 5.8S and 26S (Hershkovitz, Zimmer & Hahn, 1999).

2.3.3.2 Taxonomic level: Few studies have utilised the entire region, due in part to its large size (about 3.5kb, made up of around 2.5kb conserved sites and about 1kb variable regions (Hershkovitz, Zimmer & Hahn, 1999)). However, phylogenetic analyses of portions of this region have been used in collaboration with 18S (c. 1800 bp), producing similar topologies to 18S for termite and fungal taxa (references in Soltis & Soltis, 1998, p. 20) and also the same plant relationships as those revealed by 18S and *rbc*L sequences (Kuzoff et al., 1998). The conserved regions within 26S seem to be more conserved per unit length of sequence than 18S, while the variable regions have been thought to be too variable to be used at the same taxonomic (divergence) level as the conserved regions (Hershkovitz, Zimmer & Hahn, 1999). It appears that the entire 26S region evolves 1.6 to 2.2 times faster than 18S and at about half the rate of *rbc*L; as 26S is longer than 18S or *rbc*L, it provides two to three times as many informative characters as either region (Kuzoff et al., 1998). The c. 1 kb of variable regions are contained within expansion segments / divergent domains. As these regions are a source of many of the phylogenetically informative characters for this gene, their evolution is discussed in some detail below.

2.3.3.3 Expansion Segments / Divergent Domains

A. Description and definition: The first size differences between prokaryote and eukaryote LSU rRNA were due to a few inserted domains interspersed among a set of conserved regions. Long tracks of the LSU molecule have been strongly conserved during evolution; additional sequences in higher eukaryotes are clustered in a few highly divergent areas identified as D1 to D12 (mouse 28S rRNA sequence) (Hassouna et al., 1984) (12 'expansion segments' in the terminology of Clark et al., 1984). Outwith

these size-variable areas, the secondary structure between four eukaryotes and *E. coli* is almost identical. Variation in the D domains seems to be due to frequent inverted or direct repeats, possibly through DNA strand slippages during replication. Hassouna et al. (1984) found that the D domains of 28S rRNA in higher eukaryotes are closely related to the transcribed spacers of the ribosomal transcription unit; most if not all the transcripts of D domains are present in mature 28S rRNA of higher eukaryotes.

B. Cryptic Simplicity: Tautz et al. (1988) coined the term 'cryptic simplicity' for scrambled permutations of direct repetitive short motifs, which are not as obvious to the eye as tandem runs of a particular motif (pure simplicity). 5.8S and 18S rRNA genes and ITS 2 are not cryptically simple; slippage-like mechanisms of variation do not seem to occur to any great extent within them. The regions of high simplicity in the 28S rRNA gene correspond almost exactly to the expansion segments (or D domains). Despite this, it appears that the set of expansion segments is coevolving during interspecific divergence, suggesting that 28S rRNA alone of the rRNAs can remain functional in the presence of the repetitive and scrambled products of slippage-like events.

C. Compensatory slippage: 'Compensatory slippage' occurs when slippage products accumulate at sites within the DNA in a manner which conserves overall secondary structure. The main differences between the longer and shorter expansion segments of highly divergent organisms are the lengths of certain major secondary structural stems. There is an analogy to be made with compensatory point mutations. It is possible that slippage might be more frequent in sequences which have biased base composition therefore higher concentrations of repetitive motifs. The species which show most prominent accumulation of slippage-generated products in their expansion segments also have the expansion segments with the most biased base composition (Hancock & Dover, 1988).

D. Function: Because the expansion segments have higher rates of sequence variation than the rest of the LSU, it has been suggested that they may lack function (Hancock & Dover, 1988). However, Hancock and Dover (1988) point out that, despite suggestions that expansion segments are

functionless and tolerated only as they do not interfere with ribosome function, the interspecific conservation of gross secondary structure found by several authors suggests that these regions are subject to some sequence constraint. Expansion segments show sequence similarity patterns in mouse, rat, frog and human but not in slime mould, yeast, nematode and *E. coli*. Rice is less clear-cut, with lower similarities between expansion segments than mouse, rat, frog or human, and with less obvious regions of localised high simplicity. Sequence similarities and heightened simplicity could be due to the consistent bias of base composition of the expansion segments within any one species.

That the expansion segments are found to be coevolving also points to their having a degree of functional interaction (Hancock & Dover, 1988). Coevolution could occur by either slippages in short regions producing larger blocks of related sequence, or intragenic gene conversion. It has been suggested that expansion segments enlarge in the main by accretion of short tracts of simple sequence to the tips of secondary-structure stems; this is consistent with the observation by Hancock et al. (Hancock & Dover, 1988, cited in text) of conserved secondary structure in expansion segments even between species which show complete sequence divergence (similar to the results found by Mai and Coleman (1997) for ITS 2). As far as overall sequence goes, the high levels of similarities found within but not between species are suggestive of concerted evolution - the expansion segments of individual species appear to have diverged and evolved as a unit.

2.3.3.4 Secondary Structure and Weighting: Dixon and Hillis (1993) examined the secondary structure of the LSU; they found that the expansion segments contain significantly more paired bases than the rest of the gene. Stem base characters supported a conventional (morphology-based) hypothesis of vertebrate relationship, while loop characters supported unconventional trees. The best results, however, were obtained when the two data sets were combined. Although the secondary structure of rRNA reduces the evolutionary independence of paired nucleotides, weighting these paired bases by a half overcompensates; Dixon and Hillis (1993) suggest a value of 0.8 (although point out that different data sets may require different weighting schemes); weighting at 0.8 produces the same tree as equal weighting;

weighting at 0.5 produces an unconventional tree (consistent with one produced from loop data alone).

2.3.3.5 Practical applications to phylogeny reconstruction

A. Animals: Most of the cited studies have been on animal 28S. Several properties of animal 28S (and particularly expansion segments) have been cited as problematic for phylogenetic reconstruction (Kuzoff et al., 1998):

1. the expansion segments have a higher base substitution rate than the conserved areas.
2. the base composition is biased (high GC).
3. indels are frequent.
4. there is character non-independance, through compensatory mutations and sequence coevolution among remote domains. Cryptic sequence similarity also violates the assumption that characters at different sites evolve independently.

B. Plants: Comparing 7 full-length angiosperm 26S sequences, Bult, Sweere and Zimmer (1995) found that levels of GC are higher in expansion segments (65%) than in conserved core segments (52%). This high GC level may be a problem if the methods of phylogeny reconstruction used assume equal base frequencies (Kuzoff et al., 1998), but it is not a problem specific to the 26S region.

Overall sequence variance is much greater in the expansion segments than in the conserved core regions. Bult, Sweere and Zimmer (1995) found 42% of nucleotide positions in the expansion segments were variable, while 10% were variable in the core regions; rates found by Kuzoff et al. (1998), from 15 species of seed plants (basal and higher eudicots and monocots, and Gnetales), are slightly higher, with expansion segments evolving 6.4 to 10.2 times as fast as the conserved regions. Levels of internal sequence similarity (motif shuffling through repeated slippage events) within expansion segments, which can violate assumptions of character independence, are generally low in plants and are really most problematic for reconstructing deep divergences (Bult, Sweere & Zimmer, 1995). However, the fact that motif shuffling can occur, however infrequently, led Bult, Sweere and Zimmer to caution that 26S

can present difficulties in the basic assumptions of homology and independence among characters.

The levels of cryptic sequence similarity are considerably lower in plant 26S than in animal 28S, and are confined to the expansion segments (Kuzoff et al., 1998). Plant 26S has an average length of 3.4kb, while animal 28S has an average length of 4.5kb. Kuzoff et al. (1998) point out that there is a positive correlation between the length of expansion segments and the cryptic similarity in the sequence. In plant 26S there is less compensatory slippage and fewer length mutations; so there may be more phylogenetic signal at higher taxonomic levels in plants than in animals. This taxonomic correlation could explain Hancock and Dover's (1988) results, where expansion segments showed sequence similarity patterns in vertebrate 28S, less in rice 26S and none in nematode, yeast, slime mould and *E. coli* 23S. Further studies are needed to look for taxonomic correlations with LSU length and expansion segment sequence similarity.

i. **Deep level:** In a study across the angiosperms, sequence information from both the conserved core regions and the expansion segments produced greater internal support, more resolution, and greater congruence with studies based on other data than using the core regions alone. This has led Kuzoff et al. (1998) to suggest that the expansion segments have useful data to contribute to reconstructions of evolutionary events which occurred in the last 100 to 200 million years.

ii. **Family and generic phylogenies:** In a phylogeny of the Saxifragaceae Juss., expansion segments provide more signal than the core regions, and the exclusion of core region sequences did not affect the resolution of a reconstructed phylogeny (Kuzoff et al., 1998).

Oxelman and Liden (1995) used the ITS 2 region and about 800 bases from the 5' end of 26S to look at evolution in *Circaeaster* Maxim. (Circeasteraceae Hutch.). For the 26S sequence they recovered a single most parsimonious tree, but the ITS 2 sequences "could not be meaningfully aligned above family level" (Oxelman & Liden, 1995, p. 191).

Ro et al. (1997) sequenced a kilobase long portion of the 5' end of 26S to test phylogenetic relationships within the Ranunculaceae, as this region shows the highest sequence variability in the gene across several angiosperm taxa. A further aim of their study was to test the phylogenetic utility of partial 26S sequence data, comparing results with morphology and other molecular studies. They found that the phylogenies produced were highly congruent with chloroplast restriction site data and sequence data from other genes and with karyological characters.

2.4. Homology assessment in molecular data sets

One of the perceived advantages of molecular data is that there are homologous characters (4 nucleotides) which are comparable across the deepest branches of life, over phylogenetic distances where identification of homologous morphological characters is difficult or impossible. However, molecular data are not immune from problems of homology assessment. One initial step is the assumption that the genes or regions under consideration are orthologous. Gene duplication is a frequent event in plant evolution and the potential exists for paralogous copies of genes to be sequenced (Page & Holmes, 1998). For cpDNA genes, where order is relatively conserved, problems of paralogy are limited (Stoebe et al., 1999). Likewise, for nuclear rDNA (a multi-gene family), providing homogenisation is efficient, problems with paralogy can be reduced. The problem is most apparent for low copy number nuclear genes, when multiple copies of divergent paralogues are often documented within individuals. (For example, Sang and Zhang (1999) found two to three diverged types of sequence at each of the *Adh*1A and *Adh*2 loci, for each of 5 putatively hybrid-origin species of *Paeonia* L. (Paeoniaceae Raf.).) If mistakenly sequenced, these can confound estimates of phylogeny.

Even when orthologous regions are being analysed, there are further problems with homology assessment. Sequences of one gene or region for two or more taxa will not necessarily be homologous at every position, due to the presence of inserted (or absence of deleted) segments of sequence. When indel events have occurred in the evolution of the sequences being analysed, then the sequence data cannot be considered to be a row of characters (a character is

a proposed homologue; prior to alignment there is no hypothesis of homology for nucleotide positions in length-variable sequence, therefore no character set; the character states are observed prior to character definition). Position (defined by alignment) is the only useful homology criterion for characters which have identical ranges of states (Doyle & Davis, 1998). Failure to insert gaps correctly causes inaccurate associations of states with characters (analogous to "leaf pubescence a cyme") (Doyle & Davis, 1998, p. 113).

Morrison and Ellis (1997) seek to distinguish between 'gaps', which are spaces introduced into sequences during the process of alignment, and 'indels', which are the actual mutation events. For the purpose of phylogenetic reconstruction, we often have to hypothesise that gaps do in fact represent indels.

Choosing between explaining differences by point mutations and explaining them in terms of indels requires some form of cost assessment. Global alignment programs look for an optimal alignment which maximises (or minimises) some overall score over entire sequences. Because it is possible to align any two sequences so that there is no mismatch (by the addition of a gap wherever a mismatch would occur) the addition of gaps must be penalised more than the cost of the mismatch (Doyle & Davis, 1998). The most commonly used form of cost assessment is the gap penalty, which specifies the cost of a gap relative to a substitution (Page & Holmes, 1998); it is also possible to consider the cost of changing the length of gaps. These gap opening and gap extension penalties influence the number and length of gaps.

There are several algorithms which will search for the alignment with the lowest cost for specified penalties. Most algorithms use exact procedures to align pairs of sequences and then use heuristics to make the pairwise alignments into a multiple alignment. There are two reasons why this may not represent the 'true' alignment (Morrison & Ellis, 1997):

1. this procedure will find local optima, not necessarily the global optimum.
2. the procedure seeks to maximise similarity, not sequence homology.

Sequence similarity may be due to common ancestry (homology), convergence, parallelism or reversal (all homoplasies).

For a data set with several sequences, Clustal constructs a tree using distances computed from pairwise alignments of sequences, and then uses this tree to determine the order of sequence input into the multiple alignment (Thompson, Higgins & Gibson, 1997). A different method is used in Wheeler and Gladstein's package (MALIGN, 1992); parsimony is used rather than distance in the initial tree construction, because the best alignment is that which produces the most parsimonious cladogram for a given set of gap costs. Selecting the appropriate costs in MALIGN is simplified in that the minimum gap cost must be over one half the substitution cost (or a change from A to gap to G would cost less than a change from A to G), while at the upper end of the scale all data sets 'asymptote' - a point is reached where further alterations to the ratios do not alter the alignment(s) (Gatesy, DeSalle & Wheeler, 1993). Likewise, if the cost assigned to transversion-transition is less than 0.5, the cost of A to C to G (where C is not observed) will be less than the cost of A to G (Wheeler, 1995).

Morrison and Ellis (1997) tested 5 different multiple alignment algorithms (including Clustal and MALIGN). Each produced different alignments. From each alignment they produced neighbour-joining, maximum likelihood and maximum parsimony trees. Although they found that the same "underlying phylogenetic signal is present in all of the alignments, and ... the phylogeny ... is thus relatively robust to variation in the sequence alignment process" (Morrison & Ellis, 1997, p. 433), they got greater variation in the tree topologies due to their alignment than they did from the different tree-building methods.

It is unlikely that **any** set of gap costs or algorithm will produce a correct alignment, because the best estimate will only be best on average and not for every part of the sequence - the likelihood of mutation varies across a nucleotide sequence (Doyle & Davis, 1998). Thus Hershkovitz, Zimmer and Hahn (1999) favour treating computational alignments as heuristic solutions, subject to reevaluation in the light of further evidence.

Alignment of a matrix by eye, although more subjective, also involves some assessment of relative costs.

Liston et al. (1999) divide the ways of dealing with problematic alignments into four categories:

1. Culling all ambiguous sites (Swofford et al., 1996).
2. Elision (Wheeler, Gatesy & DeSalle, 1995).
3. Optimal alignment - comparing individual automated alignments using tree statistics (Bogler & Simpson, 1996).
4. Single manual alignment (this is the most common approach, but is best used on relatively unambiguous matrices).

The methods of culling and elision represent the extremes of the analytical procedure.

2.4.1. Culling: Gatesy, DeSalle and Wheeler (1993) were concerned that data is usually excluded from analyses on subjective grounds. *A priori* data exclusion is an "extreme form of character weighting" (Gatesy, DeSalle & Wheeler, 1993, p. 155) and should not be determined by the "whim" of individual researchers. They suggest a repeatable, objective protocol, whereby alignments are created over a wide range of gap: substitution cost ratios (they varied settings in MALIGN from 2/3:1 to 300:1, although admitted that this was extreme). Alignment-invariant nucleotide positions (constant across all alignments for all taxa) are identified and are used in phylogenetic analyses. They do point out a problem with this method, which is that, despite its greater repeatability and subjectivity, much information (contained in the alignment-ambiguous sites) can be lost.

Swofford et al. (1996) put forward an alternative viewpoint: that data are excluded from analyses "from the moment one chooses a particular gene, set of genes, or gene region to use in a systematic study" (Swofford et al., 1996, p. 500). In the same way that researchers avoid genes they know *a priori* to be evolving too fast in their study group, sequence data may also be culled after being gathered. They believe that "the benefits of excluding clearly unalignable

regions - however subjectively determined - outweigh the dangers."

Culling is the most conservative method, but it can lead to poor resolution of relationships within clades (Gatesy, DeSalle & Wheeler, 1993; Soltis, Johnson & Looney, 1996).

2.4.2. Elision: Eight months after submitting their paper on the use of the culling method (Gatesy, DeSalle & Wheeler, 1993), the same authors (Wheeler, Gatesy & DeSalle, 1995) revisited the topic, with the paper 'Elision: A Method for Accommodating Multiple Molecular Sequence Alignments with Alignment-Ambiguous Sites". They suggest, as a method for including all the information from a data set, the accumulation of various alignments created using different gap penalties into one large 'elision' set, thus downweighting positions which vary among alignments and applying a heavier weight to positions which are consistently aligned. 'Culling' (Gatsy, DeSalle & Wheeler, 1993) created robust but rather unresolved hypotheses of relationship, whereas this new method applies weights in a continuous fashion to nucleotide positions. While there may be a problem in homology assessments with data analysed using the elision method (individual bases must have individual histories, but using elision, each base contributes more than once as different characters) it allows phylogenetic analysis even of data sets with sequence alignment ambiguities (Wheeler, Gatesy & DeSalle, 1995). Of course, the number of different alignments which are added together in the elision matrix will affect the amount of weighting placed on consistently-aligned sequence positions; given a sufficient number of matrices, the effect will be similar to that of culling, with virtually no information from variable positions filtering though.

Swensen, Luthi and Rieseberg (1998) had difficulty aligning ITS sequences from the Datiscaceae Bercht. & J.Presl., Begoniaceae and Cucurbitaceae Juss. They used ten different alignments generated by ClustalX (Thompson, Higgins & Gibson, 1997) using different gap opening and gap extension penalties as inputs for phylogenetic analysis and also produced an elision data set of all ten alignments put together. This strategy was preferred over 'culling', which would have removed a large amount of data given that

sequences from the outgroup taxa were substantially divergent from the ingroup.

2.4.3. Optimal alignment: Bogler and Simpson (1996) used ITS to produce a phylogeny of the Agavaceae Dumort. Because they found simple manual alignment of the sequences difficult and subjective, they used homoplasy indices to evaluate different alignments (created by varying the gap penalty in a computer package.) They considered the alignment which produced phylograms with the lowest levels of homoplasy (measured using the CI, RI and RC) to be optimal. They found that nearly all the alignments they created produced trees with similar topologies.

Li et al. (1999) also had difficulty aligning ITS sequences, from 28 genera in the Hamamelidaceae R.Br., a highly morphologically diverse family. Li et al. (1999) therefore tested various alignments, selecting the one which created trees with the highest RC index as being optimal for both ITS 1 and ITS 2.

However, tree statistics are not necessarily the best way to find a 'true' tree. Morrison and Ellis (1997) tested Clustal alignments using 9 gap opening penalties and 8 gap extension penalties (giving 72 separate alignments). None of these alignments produced what they considered to be the 'true' tree (which they obtained on the basis of an alignment which included information from secondary structure, as they expected this to be most likely to have produced the multiple-sequence alignment closest to the 'true' alignment. Of course, the validity of these assumptions is not testable).

2.4.4. Using the entire data set: Wenzel and Siddall (1999) found that, where 20% of a data matrix was replaced by "noise" (random, signal-free data), or where a noise matrix the same size as the original matrix was added on to it, if the original cladogram was supported by one synapomorphy per node, the original signal was recovered by parsimony over 50% of the time. A pectinate topology was more stable than a balanced cladogram to this sort of manipulation. A higher proportion of the trees reported in the literature are pectinate than would be expected given a Markovian (equal rate random) branching process of speciation (Pearson, 1999); if most real trees are

pectinate the effects of noise may be less severe. Wenzel and Siddall (1999) ask: "[i]f including all of the data results in a tree that coincides with conventional wisdom, would proponents of data triage still advocate the downweighting or elimination of whole portions of data, even if doing so results in a radically *unconventional* hypothesis? ..... In the very worst case, truly saturated data will not necessarily be misinformative. They might be misinformative, uninformative, or even informative ..... If one knows in advance what the relationships should be, there is not much point in looking for them" (Wenzel & Siddall, 1999, p. 62).

2.4.5 Secondary structure: If there is some *a priori* model of sequence secondary structure, the alignment can be constrained by this model (Morrison & Ellis, 1997).

Hershkovitz, Zimmer and Hahn (1999) detail two ways of using secondary structural information in alignment, with the proviso that RNA secondary structure is apparently dynamic in vivo, and presumably also dynamic evolutionarily:

1. analysing substitution covariance, and using it as evidence of compensatory mutation, therefore of base pairing in secondary structure.
2. analysing the minimum free energy of folded rRNA, using heuristics (therefore obtaining estimates).

Hershkovitz and Zimmer (1996) used an heuristic package (MULFOLD) to produce a set of consensus features for the ITS 2 region. They found that multiple, radically different, secondary structures may have similar minimum free-energy values. Also, experimental evidence suggests that in *Chlamydomonas* and in yeast, the secondary structures have sub-minimal free-energy. Thus minimum free energy is not reliable as the sole criterion for secondary structure prediction. Backing up a secondary structure with evidence of compensatory mutations in related taxa gives added weight to the hypothesis (Mai & Coleman, 1997).

2.4.6 Treatment of gaps: The gaps inserted during alignment represent hypothetical evolutionary events; they are thus potential phylogenetic characters. Although gaps may only be inferred, while nucleotide substitutions are observed, nucleotides themselves only become characters after alignment (Doyle & Davis, 1998). Just as there are alternative ways of dealing with alignments, there are different ways of treating gaps once they have been inserted:

1. culling all sites with gaps;
2. as a 5th state (A, C, T, G, gap);
3. as missing data / uncertainly (which in most parsimony analyses will be assigned the most parsimonious solution);
4. coded in a separate matrix.

Many studies include gap matrices to utilise any phylogenetic information. However, it can be difficult to assess homologies for overlapping or length-variable gaps (Doyle & Davis, 1998).

Swensen, Luthi and Rieseberg (1998) treat gaps in 18S as a fifth state because they consider it likely that the single nucleotide gaps in their sequence are caused by single evolutionary events, while they treat gaps in ITS as missing data (as these multiple nucleotide gaps could have been generated by one or more events). Gaps (treated as missing data) can lead to "the generation of multiple equally most parsimonious cladograms, to spurious theories of character evolution, and to lack of resolution by masking the phylogenetic signal implied by the observed data" (Kitching et al., 1998, p. 80). However, they will not alter the topological relationship of taxa (Kitching et al., 1998).

2.5 Summary

Molecular data can be used to generate phylogenies rapidly and efficiently, while, in contrast, morphological data suffers problems with homoplasy which may often tie in with convergence (e.g. habitat in *Viola*, and pollination syndrome in *Ipomoea*).

Selecting the correct gene or region for a problem is one of the most difficult stages in phylogenetic analysis; there are a wide range to choose from. It is becoming increasingly apparent that faster genes offer more information at deeper phylogenetic levels than had previously been supposed and they are more likely to track rapid speciation events (although may be more difficult to align). The ribosomal DNA cistron is very frequently used for phylogenetic reconstruction; ITS and 26S were selected from it for this present study. The ITS region is made up of two transcribed spacers (ITS 1 and ITS 2) separated by a short gene (5.8S). It appears to have some function in the maturation of nuclear rRNAs, which imposes some evolutionary constraint on it (most notably on the secondary structure of ITS 2). Although there is intra-individual polymorphism in ITS, it does not appear to lead to inaccurate phylogenetic reconstructions.

The ribosomal large subunit (26S in plants) has been used to a lesser extent in phylogenetics. It is made up of a long, relatively conserved, region which is broken up by 12 highly divergent regions. These divergent regions are not present in prokaryotes, and are shorter in the examined plant taxa than in animal taxa. Particularly within these divergent regions, there are complex patterns of sequence evolution (cryptic similarity and compensatory slippage), although these appear less liable to bias phylogenetic analysis in plant taxa than in animals.

Homoplasy is not restricted to morphological data, and can occur at several levels in molecular sequence data. First is the issue of orthology / paralogy, second, that of the alignment of the orthologous sequences by the insertion of hypothesised indels. Workers have used a variety of means, both more and less subjective, to obtain their aligned matrices, and have used a variety of methods to deal with the indel events within their matrices.

3. Analysis of large data sets using parsimony

No efficient algorithm exists to find the optimal tree (using minimum evolution or maximum parsimony) for over c. 20 sequences; heuristic methods must be used (Page & Holmes, 1998). Despite recent improvements in the programs used for maximum likelihood analyses, there is an upper limit of 50-60 taxa on the size of data set which can be handled (Soltis & Soltis, 2000) so it is not (yet) practicable for truly large data sets.

A recent review by Soltis and Soltis (2000) summarises the current methodology for the analysis of large data sets (defined (arbitrarily) as having over 150 placeholders (leaves / terminals)).

There has been a lot of debate about how feasible large analyses are, given the size of treespace - for 10 taxa there are over 34 million possible rooted trees (Page & Holmes, 1998); for 20 taxa there are 8.87×10^{23} possible rooted trees (Soltis & Soltis, 2000); for 135 taxa there are 2.113×10^{267} different trees, exceeding the number of particles in the known universe (Page & Holmes, 1998). Recent analyses of large angiosperm data sets have, however, come up with strikingly similar topologies for different genes, despite searches not swapping to completion, suggesting that real patterns are being recovered (Soltis & Soltis, 2000).

3.1. Addition of data

It appears from analyses of the Angiosperm Data Set that adding more taxa and more characters not only increases the accuracy of tree estimation, but also reduces the length of time the computer requires to find a solution; this seems to be because addition of taxa breaks up long branches and disperses homoplasy (Soltis & Soltis, 2000). Adding more characters will not only make it less probably that large numbers of trees with different topologies but the same overall length will exist, but can also reduce the difference between the length of the starting tree(s) in parsimony analysis and the length of the shortest tree(s). Chase and Cox (1998) found, for a 141 taxon, 3 gene matrix (*rbc*L, *atp*B and 18S), that this length difference explained the decrease in analysis time for the combined gene matrix over the single-gene matrices. In

fact, they argue that genes or regions with high functional constraints will have more homoplasy (e.g. convergence), therefore the starting trees will be further from the shortest tree length than regions with lower constraints (Chase & Cox, 1998).

'Long branch attraction' is said to occur when there are large differences in the rates of evolution among sequences, or where the sequences are quite divergent. The length of branches *per se* is not the problem; the difficulty occurs when the same substitutions occur independently on two long branches (homoplasy). Intuitively, this is less of a problem if the long branches are widely separated phylogenetically - closer relatives probably had similarities to begin with which have been compounded (Page & Holmes, 1998). Adding taxa to regions where there are perceived to be difficulties is one way of dealing with this problem (up to a point; it is not always possible to add taxa, e.g. Richardson, 1999, Rhamnaceae Juss. Following a "relict hypothesis" (Cronk, 1989) one would expect there to be many cases where data addition was not an option due to extinction.) Maximum likelihood is said to avoid such problems with homoplasy and so the comparison of trees produced using both methods is often advocated (but not possible for large data sets). An alternative (and faster) test, when two taxa are thought to be exhibiting long branch attraction, is reanalysis of the data, each time including only one of the two taxa. If the positions of the solitary taxa are invariant, then long branch attraction can be ruled out as a factor (Siddall & Whiting, 1999).

Inconsistent trees can be made consistent by the addition of taxa (which shortens the average branch length). Graybeal (1998) looked at whether it is better to add characters or taxa; she found that it is "always preferable to add taxa rather than characters" (Graybeal, 1998, p. 13). Trees are reconstructed most accurately when taxa are added closest to the bases of long branches; adding taxa near the tips is least efficient. For many real data sets this can be a problem, because the 'difficult' parts of the tree are often those which contain the most isolated clades, with the least potential for the addition of taxa. Furthermore, when the sequence data give a polytomy, adding taxa will not resolve the relationships; only adding more sequence data can give more resolution.

Kim (1996) also gives recommendations for avoiding inconsistency problems in tree reconstruction, but comes to a different conclusion: to use regions with a low rate of change and to use fewer rather than more taxa (as larger trees are more likely to include inconsistent branches).

3.2. Adding taxa and tree confidence measures

Sanderson (1990) discusses the problem of hidden homoplasy (for example, homoplasy on the same branch or on the branches leading to two sister taxa). The only way that this can be discovered is by the addition of taxa to the phylogeny. Sanderson considers there to be a bias in hidden homoplasy levels - "lineages in which many taxa have been 'added' by evolution will tend to display a larger fraction of the actual homoplasy than depauperate lineages" (Sanderson, 1990, p. 387). Thus tree lengths for phylogenies created with too few taxa may be artificially low. Also, tree statistics could be influenced by the diversification rate of the taxa being examined.

So the addition of taxa to the ingroup can increase measures of homoplasy. A further effect is the breaking up of branches. While in many ways this is desirable, reducing analytical problems with long branch attraction and potentially anchoring inconsistent clades or taxa, our measures of tree support rely to a greater or lesser extent on absolute branch length. In fact, one of the most obvious measures **is** branch length; Bremer values are also strongly correlated to it. Resampling measures like Bootstrapping and Jackknifing are also less likely to recover shorter internal branches. The addition of taxa may give a truer tree; it may also reduce islands of equally parsimonious trees; however, it will not necessarily lead to improved values for tree confidence measures.

Adding characters can also lead to decreased confidence values. For example, with bootstrapping, the expected bootstrap frequency of a group *G* which has *r* uncontradicted characters is $1-p^r$, where *p* is the probability of any character being absent from the resampled matrix. If *n* is the number of characters, and *r* is constant, the bootstrap frequency of *G* is $1-(1-r/n)^n$, a value which decreases as *n* increases. Even the addition of autapomorphies to

groups irrelevant to *G* can therefore decrease the bootstrap frequency of *G* (Farris et al., 1996). This actually relates directly to the issue of adding taxa to sequence data matrices, because in many cases the addition of taxa will increase the number of characters in the matrix, simply by turning some constant characters into parsimony-uninformative characters (and likewise, some uninformative characters into informative characters)

With a smaller data set, Wojciechowski, Sanderson and Hu (1999) found high bootstrap support for a monophyletic clade of New World *Astragalus* species; for a far larger data set it became low. Bootstrap proportions are expected to decline with increased taxon sampling in a large clade, and eventually taxa will be sampled which, by chance, have reversals at the synapomorphy for the clade. Thus with the addition of taxa, it becomes more likely that homoplasy will 'knock out' a clade (Wojciechowski, Sanderson & Hu, 1999). Adding taxa may also break up internal branches, decreasing the levels of bootstrap support for the clades at the ends of those branches. Procedures have been suggested which give better estimates of data support. Using an iterated bootstrap procedure (Efron et al., 1996), Wojciechowski, Sanderson & Hu (1999) were able to get corrected values for their large analysis which are very close to those they received for the smaller study.

3.3. Rapid searching using confidence measures

Soltis and Soltis (2000) also suggest that, as well-supported clades appear early-on in long parsimony analyses, it may be more efficient to only resolve those groups with reasonable support. This can be done using parsimony jackknifing (Farris et al., 1996). Savolainen et al. (2000) also argue that "[t]he only relationships that we can be confident about are those that have high internal support, and performing a bootstrap analysis does not first require swapping to find the shortest tree".

3.4. Using better programs and methods

Advances in the programs available for phylogenetic analyses have also helped the analysis of large data sets. PAUP* 4.0b3a (Swofford, 2000) run on a G4 Macintosh is many times faster than running PAUP 3.1.1 (Swofford, 1993) on a Macintosh with a Quadra operating system.

Goloboff (1999) describes the difficulty caused by 'composite optima' in large data sets, which make it unlikely that any search using random taxon addition and TBR will find a global optimum. Data sets with over 40 to 50 taxa can exhibit local optima (or 'islands'); large trees are composed of many sectors (clades of over 40 to 50 taxa), each of which will have its own local optima. The globally optimum tree will have all of the sectors at their local optima; Goloboff estimates that, for the Angiosperm phylogeny data set from Chase et al. (1993), there are 10 sectors and if each has a 50% chance of hitting its optimum in any search, the probability of hitting the tree where all 10 sectors are at their optima is 0.5^{10}, or less than one in 1000 replicates. However, identification of this problem of composite optima has led to a number of analyses methods which are designed to solve it. These methods do not spend time searching for large numbers of equally parsimonious trees at different optima, but concentrate on finding the shortest possible trees quickly (Goloboff, 1999).

One of these is Nixon's Parsimony Ratchet method (1999, which can be implemented with the PC based packages DADA and NONA), which is able to sample many different tree islands. Nixon claims that, compared to previous search strategies, the parsimony ratchet is more likely to encounter shorter trees in any given time and collects a broader sample of trees of any given length. Nixon (1999) reanalysed the Chase et al. (1993) 500 taxon data set. Chase et al. spent one month TBR swapping on a single tree, using PAUP. Rice et al. (1997) reanalysed the matrix; they swapped with TBR for 11.6 months, finding trees 5 steps shorter. Using NONA, Nixon found trees the length of the Chase et al. trees in about 15 minutes, and the length of the Rice et al. trees in between 30 minutes and one hour (depending on parameters used). In between one and a quarter and two and a quarter hours, ratchet analysis found trees two steps shorter than the Rice et al. trees.

As long as the search for the shortest tree is a recognised goal of phylogenetic analysis (but see provisos in Savolainen et al., 2000), such software and hardware advances will dramatically cut analysis times.

3.5. Super trees

Most individual cladistic studies only sample a few taxa; thus our knowledge of the wider tree of life is fragmentary. However, topologies which share several taxa can be 'grafted' together (Sanderson, Purvis & Henze, 1998). A tree which is made up in such a way is termed a 'supertree', and may include trees from several different types of data set (different genes or morphology). A 'strict supertree' is a supertree which agrees with all the trees from which it was derived (Sanderson, Purvis & Henze, 1998). Algorithms are also available to calculate 'reduced supertrees', which can be constructed from source trees which are not completely compatible (Wilkinson & Thorley, 1998).

The supertree approach can be used after a large data set has been analysed phylogenetically to obtain an overall topology, to graft clades which have been subjected to more intense sampling onto the main tree (Soltis & Soltis, 2000).

3.6. Compartmentalization

This method involves partitioning the data to allow subset analysis (Mishler, 1994; Mishler et al., 1998). Known monophyletic groups can be represented in the analysis by an inferred hypothetical ancestor (with character states based on the group rather than an exemplar taxon). Alternatively, compartments can be analysed using constraints imposed from the topologies found by local analyses. Thus the method would be followed thus: 1. Global analysis to identify compartments. 2. Local analysis within compartments. 3. Global analysis, with compartments represented by hypothetical ancestors or as constraint trees.

Of course, as this technique requires a global analysis in the first instance, it can still require intensive computer time. However, one benefit is that the homology assessment within compartments will be much improved from the

global analysis; this technique is most likely to be useful in the analyses of large data sets across large phylogenetic distance (Soltis & Soltis, 2000) or in data sets which contain conserved and variable regions.

3.7 Summary

Although heuristic searches must be used to analyse large data sets, and it has previously been supposed that the vast size of treespace makes it impossible to find the best solutions, current literature seems to be converging on the view that the problem is not intractable. Adding more characters to a matrix seems to decrease analysis time by making the length of the starting tree closer to that of the most parsimonious tree, while adding taxa can negate problems of homoplasy ('long branch attraction'). It may not even be necessary to find the shortest tree for a data set; several-gene studies suggest that the clades which are rapidly recovered (e.g. by bootstrapping) are those which are most reliable overall, so we may not then need to spend a long time searching for further relationships in which we can have little confidence.

4. Begoniaceae Bercht. & J.Presl.

4.1 Size and distribution

The Begoniaceae includes the genera *Symbegonia* Warb. (c. 12 species, New Guinea), *Hillebrandia* Oliver (monotypic, Hawaiian archipelago) and *Begonia* L. *Begonia* is one of the largest genera of vascular plants, with around 1400 named species and certainly much undescribed material from less collected areas like Sulawesi and the Philippines. *Begonia* has a near-pantropical distribution; it is absent only from Australia and New Zealand and extends as far north as the Western Hills near Beijing. Species in the Begoniaceae are largely understorey herbs, although the family also includes epiphytes, shrubs and sub-trees. The monophyly of the family has never really been questioned; autapomorphies of the family like the asymmetric leaf, dry 3 winged fruit and bifid style are common to most of the species, while a ring of collar cells below the micropylar-hilar part of the seed is present in all species (Bouman & de Lange, 1983). *Hillebrandia* is distinguished by being the only member of the family to have a semi-inferior ovary, while *Symbegonia* is characterised by including the only Asian species which have complete fusion of all tepals in the female flowers into long tubes.

4.2 Taxonomic history

4.2.1 Begoniaceae: The most recent comprehensive monograph of the Begoniaceae was by Irmscher (1925); Smith et al. (1986) produced an illustrated key to the species of Begoniaceae - subsequent taxonomic changes and the publishing of many new species render this rather unwieldy work outdated.

Although the order in which Begoniaceae is placed varies, the families it has been considered to be allied to are usually consistent. For example, Begoniaceae has been placed in Passiflorales (Bentham & Hooker, 1862, with Samydaceae Vent. [= Flacourtiaceae Rich.], Loasaceae Juss. ex DC, Turneraceae Kunth ex DC, Passifloraceae Juss. ex DC, Cucurbitaceae Juss. and Datiscaceae Bercht. & J.Presl.), in Cucurbitales (Hutchinson, 1959, with

Cucurbitaceae, Datiscaceae and Caricaceae Dumort.) and in the Violales (by Richardson, 1993, who comments that Begoniaceae is an "homogeneous assemblage of no obvious affinities ..... usually placed in the Violales ..... probably most closely related to Datiscaceae" (p. 114), and by Mabberly, 1998, who also comments that Begoniaceae is considered to be allied to Datiscaceae).

Begoniaceae is placed within the Cucurbitales by recent large-scale molecular phylogenies e.g. Savolainen et al.'s *rbc*L phylogeny (2000), which includes it in a clade with Anisophyllaceae Ridley., Datiscaceae, Cucurbitaceae, Coriariaceae DC, Corynocarpaceae Engl. and Tetramelaceae (Warb.) Airy Shaw. Datiscaceae has a sister group relationship with Begoniaceae in phylogenies produced from *rbc*L sequence data (Chase et al., 1993; Swensen, Mullin & Chase, 1994; Swensen, 1996; Swensen, Luthi & Rieseberg, 1998) and 18S rDNA sequence data (Soltis et al., 1997, Swensen, Luthi & Rieseberg, 1998).

4.2.2 *Begonia*: A recent subgeneric treatment classifies *Begonia* species into 63 sections, each limited to one continent (Doorenbos, Sosef & de Wilde, 1998). There has been no published phylogeny of the genus (although Doorenbos, Sosef & de Wilde (1998) produced a phenogram of sectional similarities, based on some admittedly polyphyletic sections). The delimitation of most of the sections of *Begonia* date from 1855; Klotzsch created them as genera in his monograph of the Begoniaceae; A. de Candolle (1859) reduced most of these genera to sections within a more broadly defined *Begonia*. However, the subsequent discovery of many intermediate species means that several of the boundaries to these taxa are no longer distinct (Tebbitt, 1997). Doorenbos, Sosef & de Wilde (1998) have produced a complete revision of the sections of the genus, which also includes a complete list of currently accepted species.

4.3 Taxonomic problems within *Begonia*

4.3.1 Homoplasy: One of the major problems with the genus *Begonia* is that, while it is comparatively easy to assign specimens to the genus, working out where they belong within it is extremely problematic. Identification to section is sometimes possible, identification to species is virtually impossible without at least a good indication of the area of geographic origin of the plant and some prior knowledge of the plants. Furthermore, sectional delimitation is inconsistent, for example the only characters shared by all members of section *Knesbeckia* (Klotzsch) A.DC. are 3 locular fruit with bifid placentation (Doorenbos, Sosef & de Wilde, 1998). These reproductive characters could be considered quite reliable; however 3 locular fruit with bifid placentae are found in 33 other sections of *Begonia*, 7 of which are African, 9 Asian and 17 American.

4.3.2 Genus size: Large genera are unwieldy and can be unpractical to construct or use keys for. Many of the taxa currently recognised as sections in *Begonia* were originally described as genera, and it may seem appealing to try to reinstate some of these genera, to reduce the size of the genus back to something more manageable.

a. Morphological splits: Dividing a large genus into several smaller ones requires the identification of major phenetic discontinuities, so that the resulting new genera are identifiable. However, if more divergent lineages **were** moved out of *Begonia*, cutting along lines where there appears to be most phenetic discontinuity, the result would be the removal of many of the African species. Clear phenetic discontinuities between American and Asian sections are not obvious. Africa is species-depauperate compared to Asia and America, with only c. 150 species, so one would be left with a genus of c. 1250 species, still ranking among the larger vascular plant genera. Furthermore, the African flora is relatively well studied and monographed and most of the undiscovered species are liable to be found in regions like Sumatra, the Philippines, New Guinea, Thailand, Viet Nam and Laos; it is more probable that any morphological discontinuities between Asian sections will be filled in as we discover and describe new species than that intermediates will be found between the strikingly distinct morphologies of the African sections.

Doorenbos, Sosef and De Wilde (1998), in the most recent revision of *Begonia*, have provided reliable placement of the estimated 1400 species into 63 preexisting sections, but have also highlighted the many problem areas within the genus. A major difficulty in *Begonia* research has been establishing whether species which have superficial similarities are closely or distantly related. The taxonomy is confused by high levels of homoplasy in the morphological characters traditionally used to delimit sections; consequently a large proportion of these sections contain species which are not closely related.

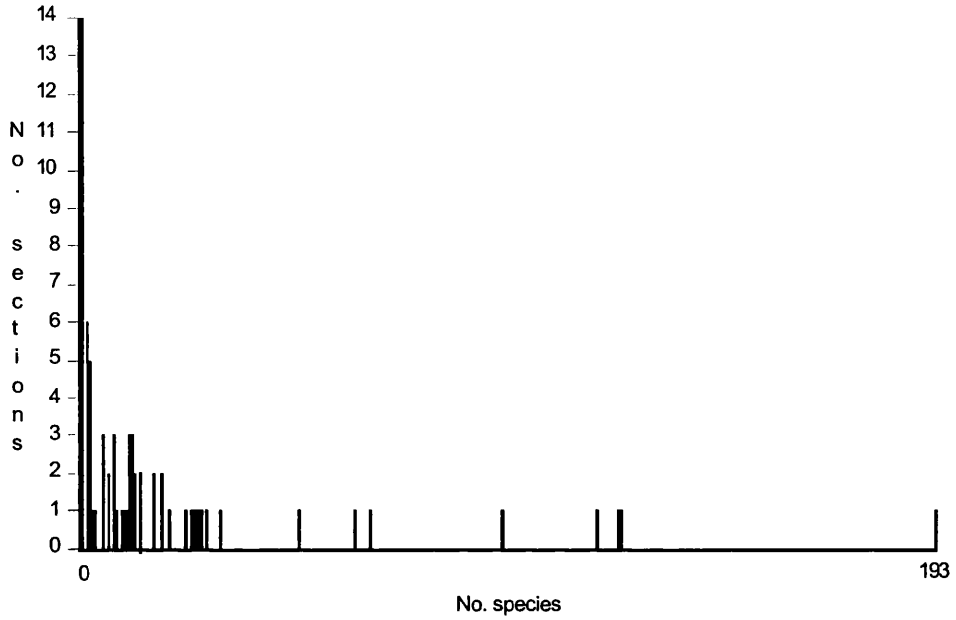
b. Molecular splits: Sequence divergence in ITS was reportedly very low (Brouillet, pers. comm. to Tebbitt, 1995), indicating that *Begonia* is a genus in which relatively rapid, recent speciation has occurred (Tebbitt, 1997). Tebbitt thus focused his research on cp-DNA, using restriction fragment length polymorphisms (RFLPs) of *nad*4 exon1 - *nad*4 exon2, *psb*C - *trn*S and *trn*C - *trn*D for cladistic analysis of 25 taxa. Badcock (1998) sequenced into the *trn*C - *trn*D intron (a non-transcribed non-coding region) from the tRNA genes, obtaining between 835 (*B. salaziensis* (Gaud.) Warb.) and 1621 (*B. rubella* Buch.-Ham ex D.Don) bases for 33 taxa. The data set contains a high number of indels, many of which are parsimony-informative. However, sampling for these phylogenies was concentrated on taxa from sections *Sphenanthera* (Hassk.) Warb. (Tebbitt, 1997) and *Knesbeckia* / *Diploclinium* (Lindl.) A.DC. (Badcock, 1998); no large-scale phylogeny for the whole genus has been constructed and no clear lines along which the genus could be split have been isolated by these studies.

4.4 Why are there so many species of *Begonia*?

Begonia contains several very small sections, and a few very large ones. The distribution of species per section (Figure 4.1) resembles the hollow curve described previously (see Figure 1.1), and prompts similar questions about taxonomy, such as whether the size of the genus *Begonia* reflects the behaviour of a plant group or the behaviour of taxonomists. However, attempting to answer such questions with the limited amount of phylogenetic knowledge we have is meaningless given that authors accept that many of

these sections are not evolutionary units but artificial constructs.

Figure 4.1: The number of species per section for *Begonia* (data from Doorenbos et al., 1998)



When addressing whether *Begonia* species richness is the product of recent or ancient events, it is worth considering published estimates for the age of the family. Wagstaff and Dawson (2000) use fossil evidence to infer a minimum age of 55 My for a Begoniaceae / Datiscaceae clade in their *rbc*L phylogeny. "The disparity in species-richness between families of the Cucurbitales [Cucurbitaceae and Begoniaceae are species-rich, while the other families are not] raises intriguing evolutionary questions" (p. 144): Has there been extinction in some lineages but not in Cucurbitaceae and Begoniaceae? Has the predominantly herbaceous nature of Begoniaceae and Cucurbitaceae allowed radiation into diverse ecological niches? The age of these families does not appear to relate to their species-richness, as, based on the fossil record, the Cucurbitaceae are far older than the Tetramelaceae, Coriariaceae and Corynocarpaceae, while Begoniaceae appear to be more recently derived.

Wagstaff and Dawson (2000) suggest that the disparity in species-richness between these groups may be a result of "imposing ranks under a traditional classification scheme" [thus artificial] and that a phylogenetic classification

assigning names to clades may better reflect patterns of diversification. (However, this appears to rely on the number of nodes from the terminal taxa down being the only consideration in rank assessment, without consideration of branch lengths across the tree. Because they have sampled all 5 *Corynocarpus* Forster & Forster f. species for *rbc*L, and only 6 of the c. 1400 *Begonia* species, better sampling in *Begonia* would add in a vast number of taxa on very short branches.)

4.5 Summary

Begonia is a remarkably species-rich genus and as such represents an useful model for understanding the processes responsible for the generation of biodiversity in the tropics. However, fundamental to any investigations seeking to understand such processes is a reliable estimate of phylogenetic relationship. Evolutionary hypotheses based on flawed estimates of relationship will be misleading. Thus it is imperative to produce a phylogeny for the genus before considering the evolutionary processes and patterns within it.

4.6 Aims

The aims of this thesis are thus:

To produce and compare ITS and 26S phylogenies for *Begonia*.
To produce an ITS sectional-level phylogeny for Begoniaceae.
To investigate the effects of changing alignment methods and the methods of analysis on tree topology.
To compare the ITS phylogeny with the existing data set for partial sequence for the chloroplast *trn*C - *trn*D region (Badcock, 1998).
To investigate mophological correlations with ITS clades, and morphological evolution in *Begonia*.
To investigate cytological evolution in *Begonia*.
To approach some understanding as to why *Begonia* is such a large genus.

5. Building a backbone

5.1 Introduction:

Obtaining Molecular-based Cladograms for Begoniaceae

Prior to evolutionary interpretation of molecular cladograms for *Begonia*, it is necessary to evaluate the diversity, support and congruence of differing topologies. Different topologies can stem from multiple most parsimonious solutions from a single analysis to explain a given data set. In addition, using different genes or parts of genes, different alignments and different search algorithms, alternate topologies may also be found. In this chapter I evaluate a range of cladograms obtained from Begoniaceae, to provide a framework for interpreting the evolutionary history of the family in subsequent chapters.

My strategy has been to obtain partial 26S and ITS sequences for 38 species. Initial results based on ITS alone presented alignment difficulties, particularly among African species and with the outgroup. Thus, to provide an alternative data set for the species which were difficult to align, the more slowly evolving 26S region was used. These two regions are physically proximal, maximising the probability of common gene history. The species chosen for this two gene approach were representative of the geographic range of the family, and showed maximal ITS divergence. More intensive sampling using just ITS is described in subsequent chapters.

5.2 Material and methods

5.2.1 Plant Material: The sources of plant material and vouchers used in this analysis are listed in Table 5.1, with sectional placements from Doorenbos, Sosef and de Wilde (1998). The choice of *Datisca* as outgroup was based on a sister group relationship in phylogenies produced from *rbcL* sequence data (Chase et al., 1993; Swensen, Mullin & Chase, 1994; Swensen, 1996; Swensen, Luthi & Rieseberg, 1998), 18S rDNA sequence data (Soltis et al., 1997; Swensen, Luthi & Rieseberg, 1998), and intuitive ideas about morphology (Lindley, 1846; Lawrence, 1951; Dahlgren, 1980; Takhtajan, 1980; Cronquist, 1981; Thorne, 1992; Bouman & de Lange, 1983; Boeswinkel, 1984). The monophyly of Begoniaceae was assumed due to the

synapomorphies of spirally arranged, asymmetric leaves and the ring of collar cells below the micropylar-hilar part of the seed; also the *Begonia* species sampled by previous molecular studies have been monophyletic with respect to Datiscaceae and Cucurbitaceae (Swensen, Luthi & Rieseberg, 1998).

Table 5.1: Taxa used in 26S and ITS analyses

SPECIES	SECTIONAL PLACEMENT	GEOGRAPHIC DISTRIBUTION	SOURCE AND ACCESSION No.
Begonia aequata	Petermannia	Asia: Philippines (Luzon)	E 1997 2515
Begonia angularis	Pritzelia	America: Brazil (Rio de Janeiro, Minas Gerias)	E 1969 1797
Begonia ankaranensis	Quadrilobaria	Africa: Madagascar	GL 001 064 97
Begonia annobonensis	Sexalaria	Africa: Cameroon, Principe, Sao Tome, Pagalu	GL 007 059 98
Begonia balansana	Ignota	Asia: IndoChina	GL 002 152 95
Begonia capillipes	Tetraphila	Africa: Cameroon, Equatorial Guinea, Gabon	GL 004 079 97
Begonia convolvulacea	Wageneria	America: Brazil (Ceara, Bahia, Rio de Janeiro)	GL 001 093 79
Begonia crassirostris (=longifolia)	Sphenanthera	Asia: China	GL 007 079 97
Begonia dewildei	Scutobegonia	Africa: Gabon	GL 001 041 97
Begonia engleri	Rostrobegonia	Africa: Tanzania	E 1998 2762
Begonia fallax = B. malabarica	Ignota	Asia: India, Sri Lanka	GL 002 018 96
Begonia floccifera	Reichenheimia	Asia: India	GL 030 099 89
Begonia francoisii	Quadrilobaria	Africa: Madagascar	GL 002 064 97
Begonia geranioides	Augustia	Africa: South Africa	GL 018 079 97
Begonia grandis var. holostyla	Diploclinium	Asia: China	E 1998 0035
Begonia holtonis	Ruizopavonia	America: Colombia, Ecuador	GL 011 129 84
Begonia incarnata	Knesbeckia	America: Mexico	GL 011 089 95
Begonia iucunda	Ignota	Africa: Congo, Dem. Rep. Congo	GL 022 079 97
Begonia lobata	Pritzelia	America: Brazil (Rio de Janeiro, Minas Gerias)	GL 020 167 95
Begonia luxurians	Scheidweilaria	America: Brazil (Sao Paulo to Minas Gerais)	E 1968 5494
Begonia madecassa	Nerviplacentaria	Africa: Madagascar	GL 003 064 97
Begonia masoniana	Coelocentrum	Asia: cult., Singapore	E 1998 0074
Begonia meyeri-johannis	Mezieria	Africa: East Africa	GL 002 041 97
Begonia molleri	Tetraphila	Africa: Sao Tome	GL 038 079 97
Begonia nossibea	Quadrilobaria	Africa: Madagascar	GL 007 064 97
Begonia obliqua	Begonia	America: Martinique	GL 005 105 91
Begonia palmata	Platycentrum	Asia: India, Nepal, Burma, China	E 1998 0059
Begonia poculifera	Squamibegonia	Africa: Nigeria to Tanzania & Angola	E 1992 3143
Begonia roxburghii	Sphenanthera	Asia: India, Nepal, Burma	GL 004 093 79
Begonia salaziensis	Mezieria	Africa: Reunion, Mauritius	K 1986 412
Begonia scapigera	Loasibegonia	Africa: Nigeria, Cameroon, Gabon, Congo	GL 002 057 96
Begonia socotrana	Peltaugustia	Socotra	E 1989 1081
Begonia sp. 'macG'	?	America	GL 1969 6248
Begonia thomeana	Cristasemen	Africa: Sao Tome, Gabon	GL 054 079 97
Begonia violifolia	Weilbachia	America: Mexico (Chiapas?)	GL 004 055 87
Datisca cannabina	N/A	Asia: S.W., Himalayas	E 1984 1126
Datisca glomerata	N/A	America: USA, California	Susan Swensen
Symbegonia sanguinea	N/A	Asia: Papua New Guinea	GL 003 127 93

5.2.2 Molecular methods

A. DNA extraction: DNA was extracted from fresh or silica gel-dried leaves using an hexadecyl-trimethyl-ammonium bromide (CTAB) method modified from Doyle and Doyle (1987), using one disc of fresh or silica dried material, ground (using a plastic pestle, in the eppendorf tube) directly in 400 µl preheated (65°c) 2x CTAB with 2 µl 2-mercaptoethanol, a pinch of polyvinylpolypyrrolidone (PVPP) and a pinch of acid-washed sand, and incubated for c. 1 hr at 65°c in a water bath. Protein extraction was performed with 500 µl 24:1 chloroform: isoamyl alcohol, gentle shaking, for c. 20 mins then 10 mins centrifugation, 13,000 revs per minute (rpm); the supernatant was removed and transferred to a clean eppendorf and this step was repeated; the DNA was then precipitated from the supernatant by adding 2/3 volume freezer-cold isopropanol, and leaving overnight in a freezer. DNA was pelleted by centrifugation (10 mins, 13,000 rpm) and left for at least 30 mins in wash buffer (76% ethanol, 10 mM sodium acetate). The pellet was then dried and dissolved in 50 µl tris-ethylenediaminetetraacetic acid (EDTA) (TE). See also Kopperud and Einset, 1995, for a *Begonia*-specific protocol.

DNA of *Datisca glomerata* was kindly supplied by Susan Swensen, Ithaca, N.Y.; DNA for *B. balansana* was supplied by Mark Tebbitt, Brooklyn, N.Y.

B. Sequence amplification and purification: For most taxa, ITS was amplified using primers p4 (White et al., 1990) and p6 (Sluiman, pers. comm., 1998). Where these did not result in a single clean amplification product, other primers were used (see Table 5.2 for sequences of primers, and Figure 5.1 for their placement):

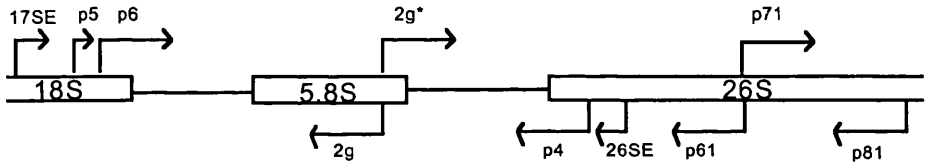
2g (Moeller & Cronk, 1997) and the reverse, 2g*.
p5 (White et al., 1990, modified by Moeller & Cronk, 1997, without the terminal 'G' cited in their paper).
p61 (Oxelman in Oxelman & Linden, 1995).
17SE and 26SE (Sun et al., 1994).

Part of the 26S region was amplified using either primers p71 and p81 (Oxelman & Linden, 1995), or, where this was problematic, 2g* and p81.

Table 5.2: Primer sequences for ITS and 26S (5' to 3')

PRIMER	SEQUENCE
17SE(F)	ACGAATTCATGGTCCGGTGAAGTGTTCG
p5 (F)	GGAAGGAGAAGTCGTAACAAG
p6 (F)	GTAGGTGAACCTGCAGAAGGA
2g* (F)	ACGTCTGCCTGGGTGTCAC
p71 (F)	ACGAGTCGGGTTGTTTGGGAATG
2g (R)	GTGACACCCAGGCAGACGT
p4 (R)	TCCTCCGCTTATTGATATGC
26SE (R)	TAGAATTCCCCGGTTCGCTCGCCGTTAC
p61 (R)	CATTCCCAAACAACCCGACT
p81 (R)	CCCGCTCAGGCATAGTTCACCAT

Figure 5.1: Primer positions



For *B. morsei*, where the DNA proved very problematic to amplify, ITS 1 and ITS 2 were amplified separately (p6 and 2g; 2g* and p4).

Polymerase chain reaction (PCR) was carried out in 50 µl reactions, using Biotaq DNA polymerase (0.2 µl Taq, 5 µl 10x reaction buffer, 5 µl deoxynucleoside triphosphates (dNTPs) at 2 mM, 2.5 µl $MgCl_2$ at 50 mM, 1.5 µl of each primer at 10 µM, 15-20 ng DNA, made up to 50 µl with water).

PCR products were electrophoresed in a 1.6% agarose gel, in 0.5 x Tris boric acid EDTA (TBE) buffer with 2 µl ethidium bromide, and visualised on an ultra violet light-box, to confirm that the PCR product was single banded.

Amplification products were purified using QIAquick PCR purification kits, following protocols supplied by the manufacturer. Some double-banded products were run out on agarose gels, cut out, and purified using QIAquick Gel Extraction kits.

PCR amplification of ITS involved: a preliminary denaturing step, 94°c for 3 minutes, (denaturing at 94°c for 1 minute; annealing at 55°c for 1 minute;

extension at 72°c for 1 minute 30 secs) for 28 - 30 cycles, a final extended extension period of 72°c for 5 minutes, then a holding stage at 4°c, using a Progene PCR machine.

PCR of 26S using p71 and p81 involved: a preliminary denaturing step, 95°c for 4 minutes, then 30 cycles (denaturing at 95°c for 30 seconds; annealing at 57°c for 1 minute; extension at 72°c for 2 minutes), then a final extended extension period of 72°c for 7 minutes, and a holding stage at 4°c, using a Progene PCR machine. PCR of the 26S region using 2g* and p81 was carried out using the ITS protocol.

A sequence for *Datisca cannabina* was obtained from Mark Tebbitt (Brooklyn, N.Y.) and Susan Swensen (Ithaca, N.Y.); the sequence from *Hillebrandia* was obtained from Susan Swensen (Ithaca, N.Y). Sequences for *B. dregei* and its varieties and *B. geranioides*, *B. socotrana*, *B. samhahensis*, *B. floccifera* and *B. dipetala* were obtained from Mark Hughes, RBGE.

C. Cloning reactions: Cloning was carried out for some ITS PCR products, to check whether there were different copies present and because heterozygosity for length mutations made some sequences unreadable from consensus sequences. Ligation of the PCR product into the vector was carried out using the protocol in the Promega pGEM-T Easy Vector kit, but halving the reaction quantities. Reactions were spread onto ampicillin and Luria-Bertani (LB) broth agar plates (with 20 ng/ml 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (XGal) and 30 µl 0.1 M isopropyl-β-thiogalactopyranoside (IPTG) for blue/white screening). Cells were cultured overnight, 37°c, in flasks with 5 ml LB broth and 0.1 g/ml ampicillin. DNA was isolated using QIAprep Spin kits and sequenced directly (3 µl product per 10 µl sequence reaction, 2 µl sequencing mix).

D. DNA sequencing: Sequencing was carried out using Amersham Thermosequenase II dye terminator cycle sequencing kit (2 µl Thermo Sequenase II reagent premix, 0.5 µl primer at 5 µM, 1-3 µl template, made up to 5 µl with water).

PCR amplification involved 25 cycles of denaturing at 96°c for 10 secs, annealing at 50°c for 5 secs, extension at 60°c for 4 minutes, then a 4°c holding stage, using a Perkin Elmer 9600 PCR machine.

Sequence gels were run by staff at the Royal Botanic Garden, Edinburgh.

Sequences were edited and assembled using Sequence Navigator (Applied Biosystems, Inc.) on a G4 Macintosh computer. All sequences will be submitted to GenBank.

5.2.3 Alignment

a. 26S: The data were aligned by eye. Sites which included gaps in more than 3 of the included species were removed from the data set prior to analysis, because their precise placement was open to interpretation. Excluded characters are 1-44, 68, 211-213, 335-337, 492, 499 and 587-595 from the matrix, see CD-ROM.

b. ITS: The data were aligned by eye; many sites were excluded because the alignment was variable (same exclusion matrix as used in next chapter). Excluded characters are 1-183, 188, 200, 204, 211-217, 223-225, 230-249, 255-256, 266, 274-329, 340-366, 378, 383-384, 406-407, 415, 419-421, 426-428, 435-437, 444, 449-451, 460, 466-469, 475-483, 493-497, 503-507, 513-514, 539, 571, 577, 603, 606, 615, 649, 686, 688, 693-856, 886-901, 930-931, 944, 957-966, 983-984, 992-993, 1013-1014, 1018, 1023, 1029-1035, 1041-1053, 1064-1093, 1110-1114, 1121-1122 and 1137-1154 from the matrix, see CD-ROM.

5.2.4 Analysis: Data matrix statistics (e.g. number of parsimony-informative characters, uncorrected pairwise differences (total number of differences/total number of available sites - Swofford et al., 1996)) were taken from PAUP* 4.0b2a (Swofford, 2000); the skewedness statistic g1 was estimated for a sample of 10,000 random trees, and the permutation tail probability test (PTP) was performed on the ingroup to assess the degree of cladistic covariation in the data set (Kitching et al., 1998) (MP heuristic search,

simple addition, saving no more than 100 MPTs for each replicate; 100 PTP replicates).

Three different types of analysis (MP, ME and ML) were run on each of three data sets, 26S, ITS and a combined 26S/ITS matrix. Maximum parsimony methods search for solutions which minimise the amount of evolutionary change required to explain the data, while maximum likelihood attempts to estimate the actual amount of evolutionary change according to a (specified) evolutionary model. Thus parsimony can underestimate 'true' change, due to unseen events (e.g. superimposed changes), while likelihood models can allow for such changes. Distance measures like minimum evolution can be used where the data set is too large for maximum likelihood studies to be feasible, although in simulation studies, likelihood methods have consistently outperformed distance methods in choosing the correct tree (Swofford et al., 1996).

5.2.4.1 Maximum parsimony (MP): Heuristic analyses were performed using maximum parsimony. No more than 100 MPTs were saved for each step, with TBR swapping to completion, zero length branches collapsed, for 1000 random addition replicates.

Bootstrapping was performed using the fast-heuristic search option, with 10,000 replicates. Bremer support was calculated using AutoDecay (Eriksson, 1998) (10 random additions, TBR swapping).

5.2.4.2 Maximum likelihood (ML): Using likelihood, the explanation which makes the observed data the most likely (i.e. probable) is preferred (Page & Holmes, 1998). An initial tree was calculated using the HKY85 model (which allows unequal base frequencies and for transversions (tv) and transitions (ts) to have different substitution rates) with gamma distribution shape parameter (α) and ti/tv ratios estimated using ML, no molecular clock assumed (discrete gamma approximation, 4 rate categories, average rate approximated by mean). This tree was used to estimate ti/tv rations and α and was then used as the starting tree for TBR swapping, using the HKY85 model.

Low values for the gamma distribution shape parameter α result from an L-shaped distribution whereby most sites have little variation while a few sites have very high rates of substitution, while when $\alpha > 1$ the distribution is bell-shaped (i.e. there is a small range of rates) (Page & Holmes, 1998).

5.2.4.3 Minimum evolution (ME): For an unrooted metric tree for n sequences, there are (2n-3) branches, each with their own length. The sum of these branch lengths is the length of the tree, L; the minimum evolution tree is that which minimises the value of L. Although this method is similar to parsimony, length is computed from pairwise differences rather than from the fit of characters to a tree (Page & Holmes, 1998). The LogDet/paralinear distance measure (which recovers an additive distance between sequences even when the base composition is variable - Page & Holmes, 1998) was used; a starting tree was calculated using neighbour joining then swapped with TBR. Zero length branches were not collapsed. (In additive distances, the distance between any two taxa is equal to the sum of the branches joining them - Swofford et al., 1996.)

5.3 Results

The data matrix of 26S sequences is presented in the Appendix, 14.5.

For each cladogram which is presented, the letters AF stand for Africa, S.AF, for southern Africa, MAD for Madagascar, SOC for Socotra, AM, America and AS, Asia.

5.3.1 26S:

5.3.1.1 Data set: The included data set comprises 439 constant characters, 34 parsimony-uninformative characters and 58 parsimony-informative characters.

Uncorrected pairwise distances (as given in PAUP 4*) are highest between *Datisca cannabina* and *B. dewildei* (0.084); within the ingroup, the highest values are between *B. ankaranensis* and *B. crassirostris* (0.065). The lowest value between the outgroup and ingroup is between *D. glomerata* and *B.*

engleri (0.047); the lowest value within the ingroup is 0.000, between *B. molleri* and *B. capillipes* and also between *B. nossibea* and *B. francoisii*. The two species of *Datisca* have a distance of 0.011.

Mean base frequencies for the data matrix are as follows:

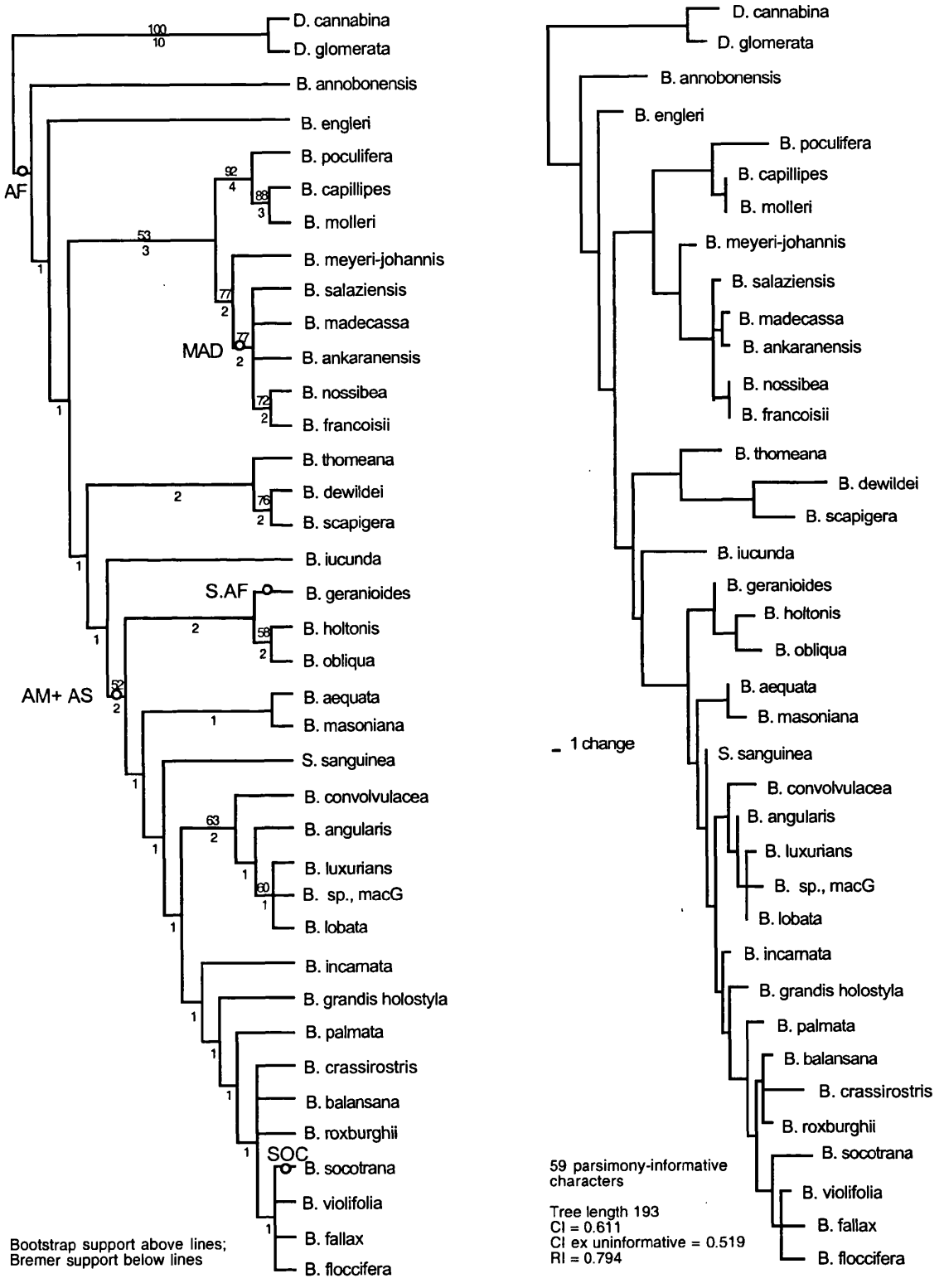
A = 0.214
C = 0.257
G = 0.352
T = 0.177
(GC = 0.609)

The skewedness statistic g_1 is -0.395. The probability for the PTP test is 0.001.

5.3.1.2 MP: 216 trees of length 185 were retained. To test whether there were more MPTs, these trees were used as starting trees on a second heuristic search and were swapped to completion. No shorter trees were found, nor any more equally parsimonious trees. The consistency index is 0.61 (excluding uninformative characters, 0.52); retention index is 0.79. 12 clades had over 50% bootstrap support. 29 nodes were resolved in the strict consensus tree. See Figure 5.2 for the strict consensus tree and one of the phylograms.

The support values for internal branches are generally low; the best supported clade consists of some African and Madagascan taxa. African taxa are resolved as basal; a derived clade with 52% bootstrap support includes all the Asian and American taxa as well as one Southern African species (*B. geranioides*) and one Socotran species (*B. socotrana*). The branch lengths in this derived clade are generally shorter than in the rest of the tree, with internal branch-lengths often in the region of 1 to 3 changes.

Figure 5.2: MP strict consensus of 18 MPTs and phylogram, 26S data set

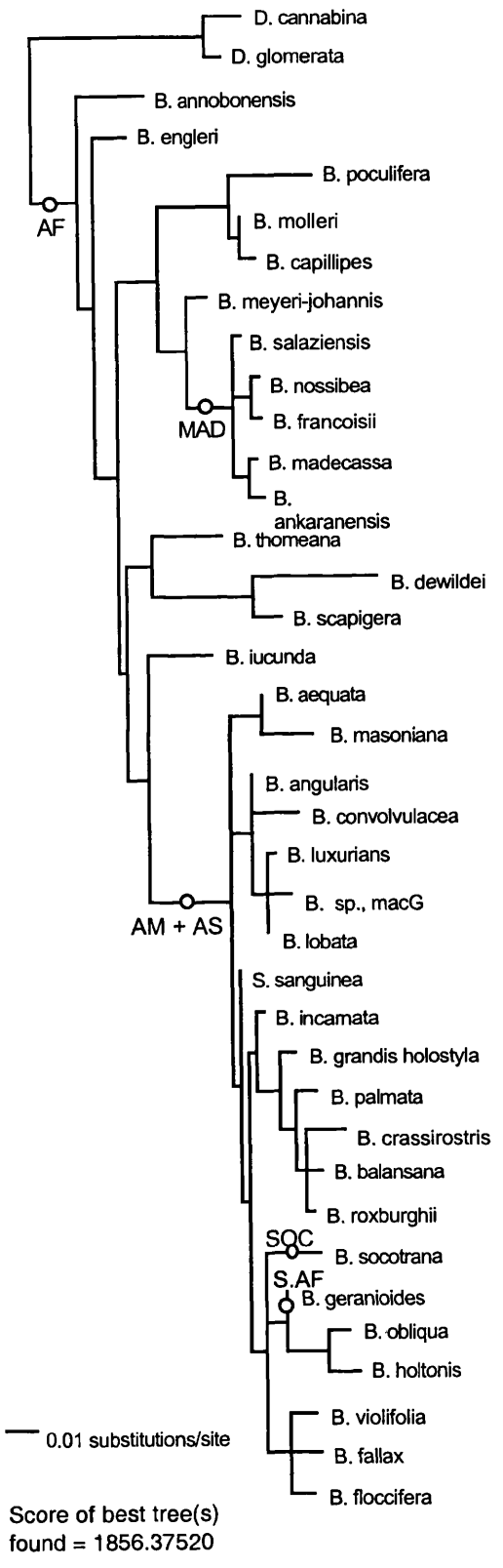


5.3.1.3 ML: The likelihood settings were as follows: Assumed nucleotide frequencies are the mean base frequencies for the data matrix; rates assumed to follow a gamma distribution with shape parameter α = 0.0906; transition/transversion ratio = 2.943 (κ = 5.982); number of distinct data patterns under this model = 129. -Ln likelihood of best tree found is 1856.375 (see Figure 5.3).

African taxa are resolved as basal, with Asian and American taxa in a derived clade with *B. geranioides* and *B. socotrana*. The tree is congruent with those derived by parsimony except for some of the placements of taxa within the Asian/American clade. All the clades with bootstrap support in the MP trees are found in the ML tree.

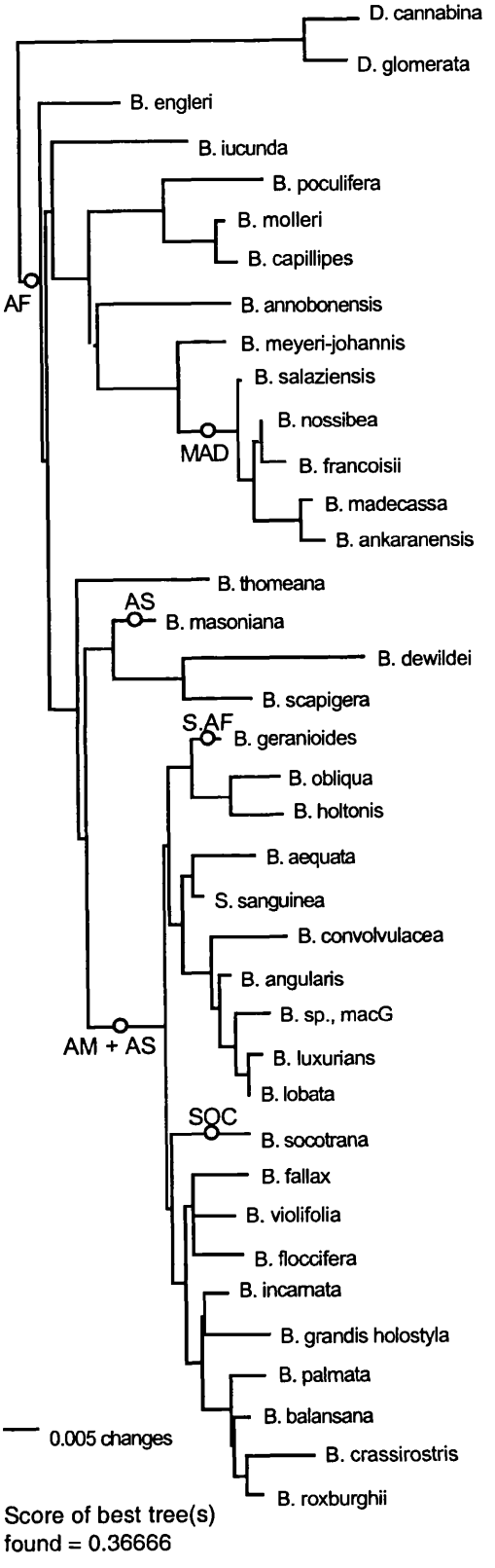
5.3.1.4 ME: One tree was found, with tree-score 0.3667 (see Figure 5.4). There are many clades in common with the trees found by MP and ML, with African taxa basal, but the placement of a few taxa is radically different. *B. iucunda*, *B. annobonensis* and *B. masoniana* have all shifted across clades. *B. iucunda* is not sister to the American/Asian clade, but is basal to an African/Madagascan clade; *B. annobonensis* is not sister to the rest of *Begonia*, but is included within an African/Madagascan clade, and *B. masoniana* is not within the American/Asian clade but is in an African clade.

Figure 5.3: ML, 26S



Score of best tree(s) found = 1856.37520

Figure 5.4: ME, 26S



Score of best tree(s) found = 0.36666

5.3.2 ITS sequence data:

5.3.2.1 Data matrix:

595 characters are excluded; the included data set comprises 223 constant characters, 70 parsimony-uninformative characters and 214 parsimony-informative characters.

Uncorrected pairwise differences range from 0.006 (*B. nossibea* to *B. francoisii*) to 0.279 (*B. floccifera* to *B. iucunda*) within *Begonia*, and from 0.207 (*B. grandis* to *Datisca*) to 0.302 (*B. engleri* to *Datisca*) between the ingroup and outgroup.

The mean base frequencies for the matrix are as follows:

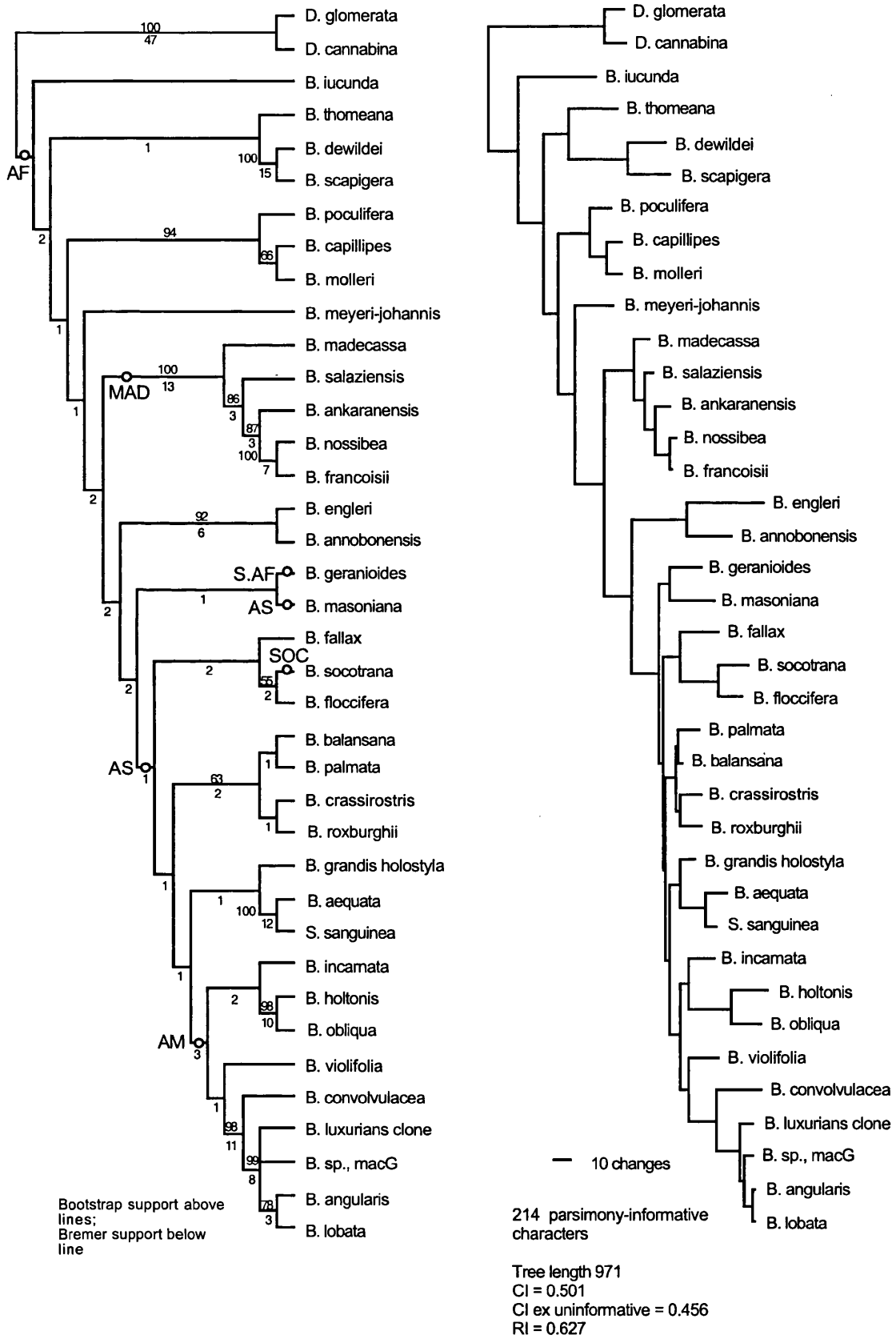
A = 0.217
C = 0.275
G = 0.295
T = 0.213
(GC = 0.570)

The skewedness statistic g1 is -0.694; the probability for the PTP test is 0.010.

5.3.2.2 MP: 3 most parsimonious trees were found, length 971. These were used as starting trees for a second round of searches with no restrictions on number of trees saved. No more or equally parsimonious trees were found. The consistency index is 0.50 (0.46 excluding uninformative characters); retention index is 0.63. 18 clades had over 50% bootstrap support. 35 nodes were resolved in the strict consensus tree. The strict consensus tree and one of the phylograms are presented in Figure 5.5.

African taxa are basal in *Begonia*, although the positions of *B. iucunda* and a *B. annobonensis*/*B. engleri* clade are the reverse of the 26S MP trees. Again, American and Asian taxa are in a derived clade which also includes *B. geranioides* and *B. socotrana*. This clade has generally shorter branch lengths than the more basal part of the tree. Like in the 26S cladograms, there is little bootstrap support for the internal branches; however, a clade of 5 Madagascan species has 100% bootstrap support, and a clade of 5 American species has 98% support.

Figure 5.5: Strict consensus of 3 MPTs and phylogram, ITS data set

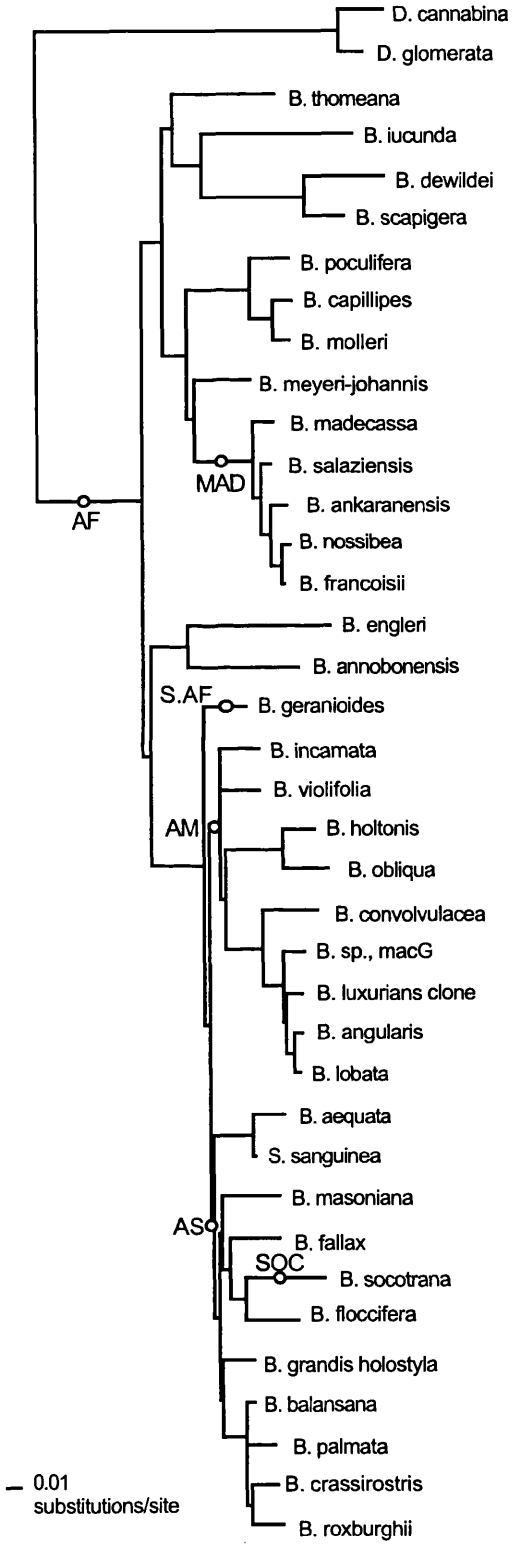


5.3.2.3 ML: Assumed nucleotide frequencies are the mean frequencies for the data matrix; rates assumed to follow a gamma distribution with shape parameter α = 0.2309; transition/transversion ratio = 2.062 (κ = 4.233); number of distinct data patterns under this model = 442; -Ln likelihood of the best tree found is 7756.352 (see Figure 5.6).

Most of the African species are in a clade sister to the rest of *Begonia*; *B. annobonensis* and *B. engleri* are sister to the American/Asian clade. Clades are broadly similar to those on the MP tree, although deeper level relationships of the African taxa are different. As with the 26S cladograms, all clades with bootstrap support in the MP analyses are present. American and Asian taxa are each in monophyletic clades.

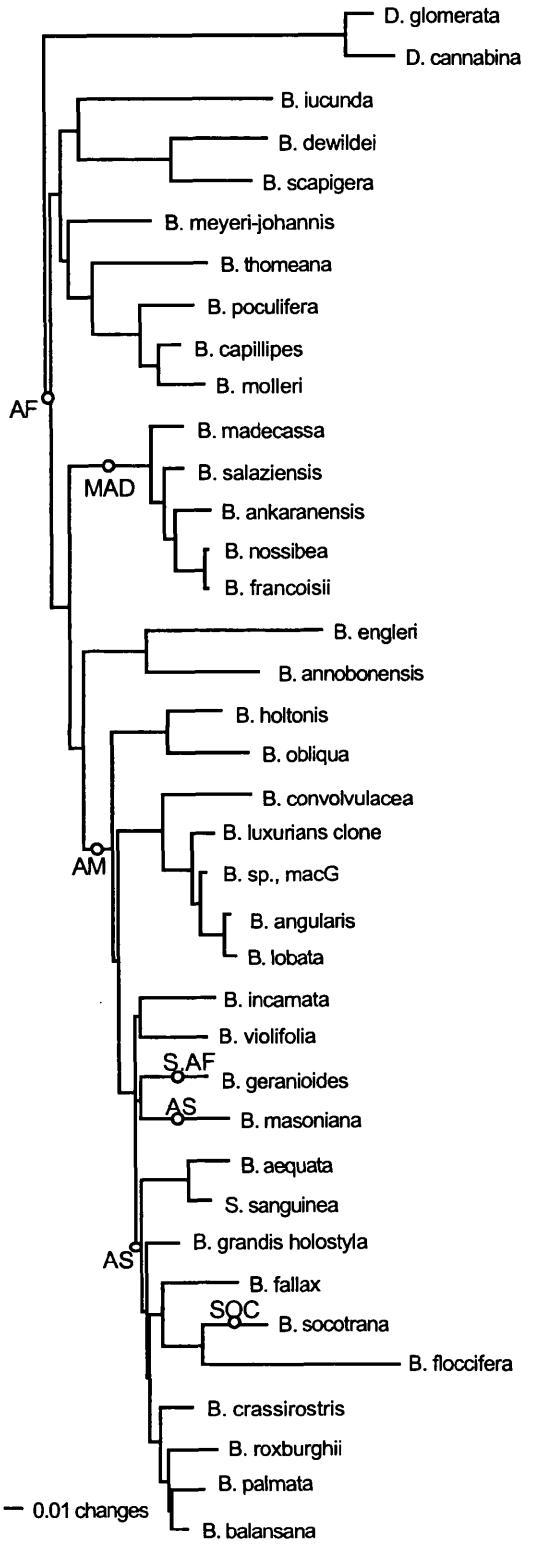
5.3.2.4 ME: The tree found had a minimum evolution score = 1.98289 - see Figure 5.7. Internal branches are very short compared to the terminal branch lengths. Again, many of the clades are similar to those recovered by MP and ML, although their positions relative to each other vary; furthermore, *B. meyeri-johannis* and *B. thomeana* have notably different positions in the ME tree to the MP and ML trees.

Figure 5.6: single ML tree, ITS



-Ln likelihood = 7756.35145

Figure 5.7: single ME tree, ITS



Minimum evolution score = 1.98289

5.3.3 Combined molecular analysis.

5.2.3.1 Data matrix: Values for types of character in the matrices are additive (i.e. the sum of those for the 26S data set and for the ITS data set).

The mean base frequencies for the combined matrix are as follows:

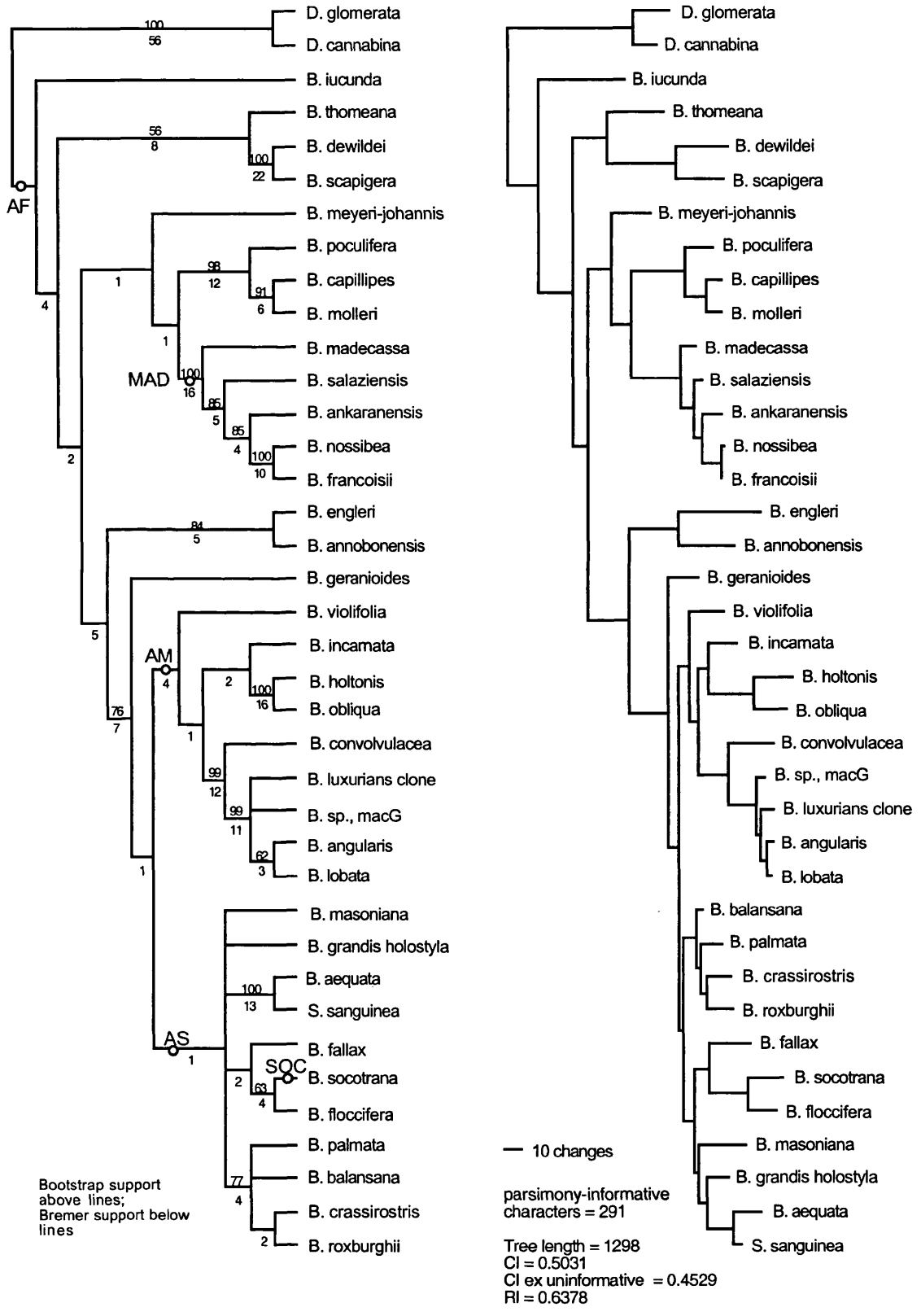
A = 0.216
C = 0.266
G = 0.324
T = 0.194

The skewedness statistic g1 is -0.584; the probability for the PTP test is 0.010.

5.3.3.2 MP: 22 most parsimonious trees were found, of length 1298, consistency index 0.50 (0.45 excluding uninformative characters) and retention index 0.64. 23 clades had over 50% bootstrap support. 31 nodes were resolved in the strict consensus tree. See Figure 5.8 for the strict consensus tree and one of the phylograms.

The topology recovered is quite different from that recovered from the 26S data set alone, but is similar to that recovered from the ITS data set, although the positions of the African taxa *B. geranioides* and *B. meyeri-johannis* have changed, and the Asian and American taxa are both monophyletic in this topology. Again, branch lengths within the Asian and American clade are generally shorter than those in the more basal clades.

Figure 5.8: 26S and ITS combined, MP strict consensus of 22 MPTs and phylogram



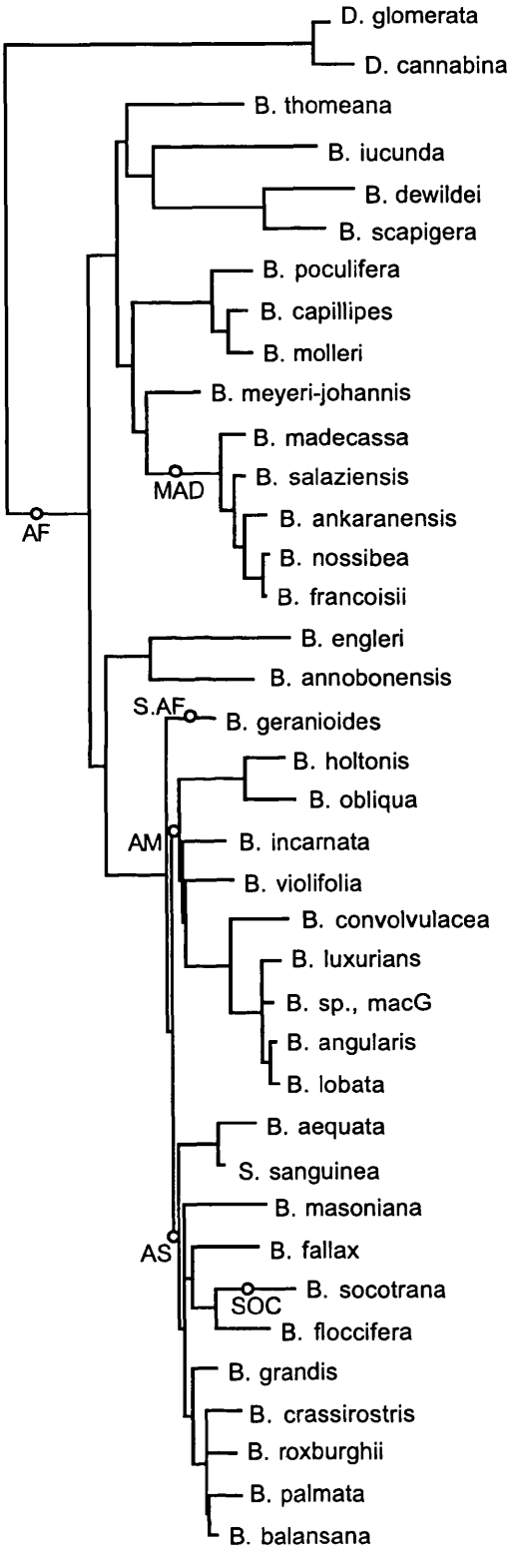
5.3.3.3 ML: Assumed nucleotide frequencies are the mean base frequencies for the matrix; rates assumed to follow a gamma distribution with shape parameter α = 0.2265; transition/transversion ratio = 2.138 (κ = 4.368); number of distinct data patterns under this model = 433; -Ln likelihood of best tree found is 7231.487 (see Figure 5.9 for tree).

The tree produced by ML for the combined data is quite different to that produced by ML for the 26S data set, but is almost identical to that produced from the ITS data set (the positions of *B. incarnata* and *B. violifolia* are slightly different). American and Asian taxa are monophyletic, with African taxa as sister and basal.

5.3.3.4 ME: One tree with minimum evolution score 1.22855 was found (see Figure 5.10).

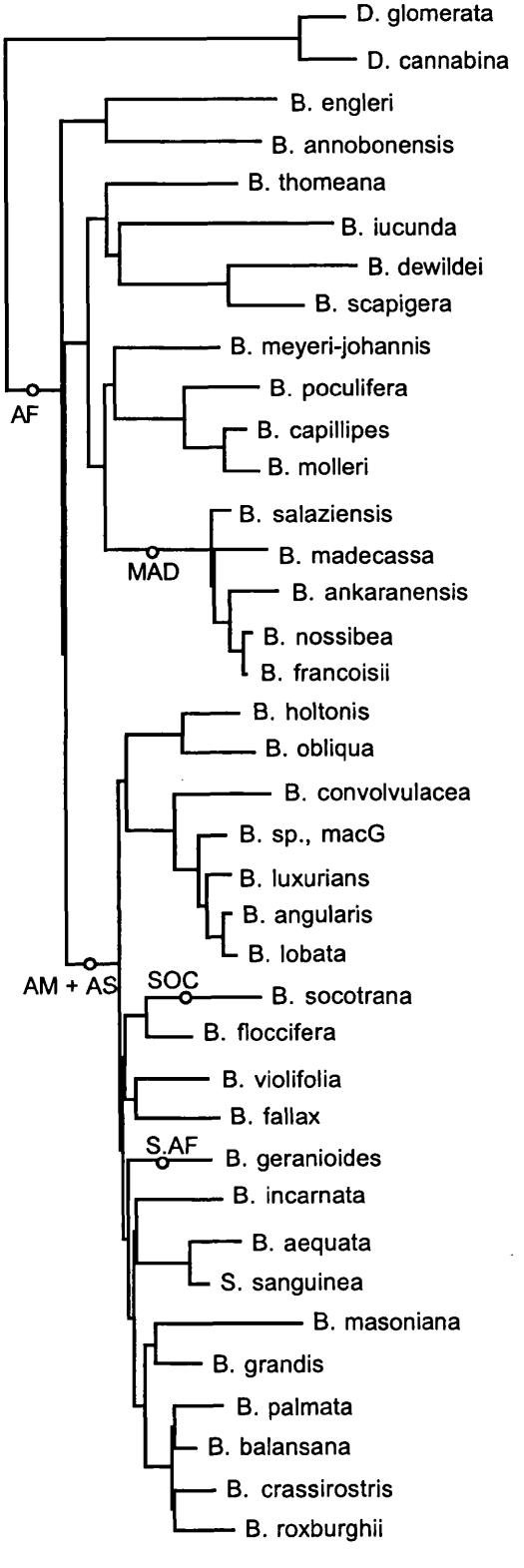
The tree differs from both the 26S and the ITS ME trees, for example in the position of the *B. annobonensis* / *B. engleri* clade.

Figure 5.9: ML tree, combined data



-Ln likelihood = 7231.48689

Figure 5.10: ME tree, combined data

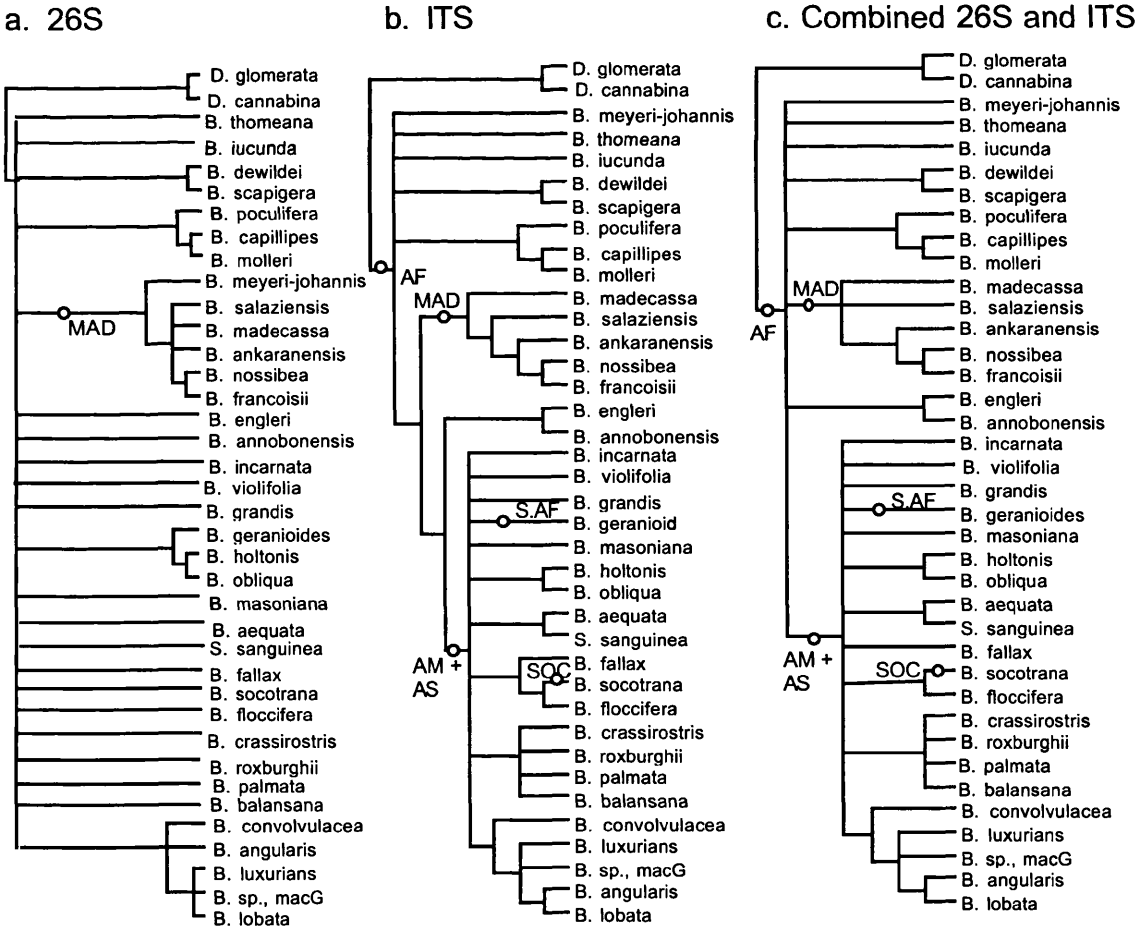


Minimum evolution score = 1.22855

5.3.4 General results

For a summary of how much agreement (or disagreement) there is between the three analytical methods for each of the three data sets, see the strict consensus trees (of the trees produced using each analysis type) for each data matrix, Figure 5.11. There are 11 nodes resolved in the 26S tree, 20 in the ITS tree and 17 in the combined 26S-ITS tree. Therefore, the 26S data set shows least similarity between MP, ML and ME, while the ITS data set shows most (estimated by the amount of resolution in the strict consensus trees). This is partly down to radical clade shifts, in the 26S analyses, by only a limited number of taxa between the ME tree and the other two analyses (ML and MP).

Figure 5.11: Strict consensus of MP, ML and ME trees for 26S, ITS and combined data sets



The tree statistics for the maximum parsimony analyses of the three data sets

are presented below, in Table 5.3. From this it can be seen that, although tree confidence measures are worse for the ITS and combined ITS and 26S data sets, their bootstrap support is better than that of the 26S data set. Also, the number of equally parsimonious trees are lower for the ITS and combined data sets.

Table 5.3: MP tree statistics

Data set	No. inform. char.s	g1	PTP	No. MPTs	Length	CI	CI ex uninf.	RI	nodes over 50%	nodes strict consen
26S	59	-0.3945	0.001	216	185	0.61	0.52	0.79	12	29
ITS	214	-0.694	0.01	3	971	0.5	0.46	0.63	18	35
combined	273	-0.5837	0.01	22	1298	0.5	0.45	0.64	23	31

5.3.4 Molecular evolution in 26S and ITS data sets

The ITS data set has more positions which have more steps over a phylogenetic tree than the 26S data set. 5 positions in ITS have 12 or more steps (see Figure 5.12), while the maximum number of steps on the 26S tree is 8 (see Figure 5.13). Within the ITS data set, the 5.8S region has a greatly reduced number of changes per site than there are in both the ITS 1 and the ITS 2 regions. The distinction between the D2 and D3 regions and the conserved segments in the 26S data are less apparent.

Figure 5.12: ITS 1, 5.8S and ITS 2 for one MPT

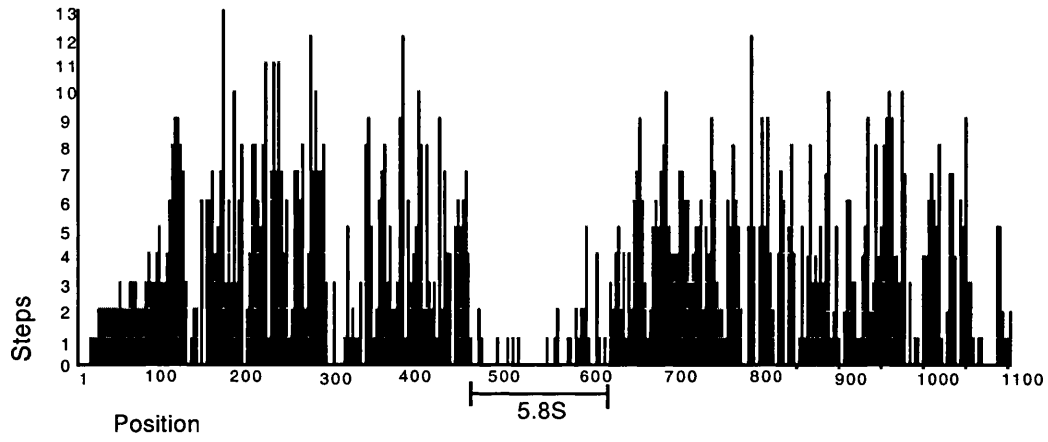
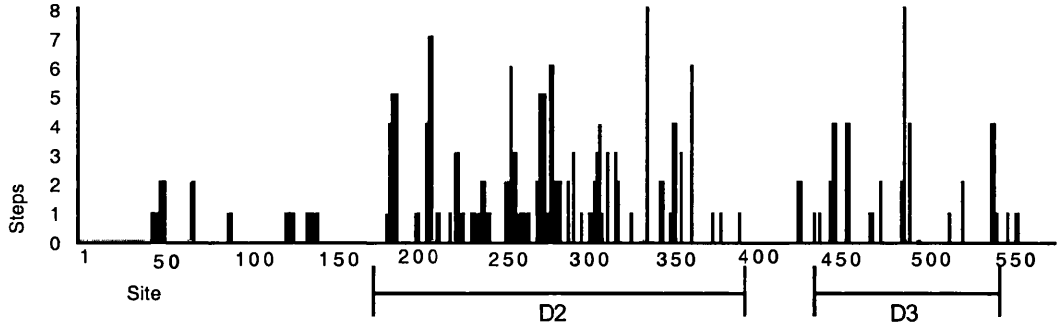
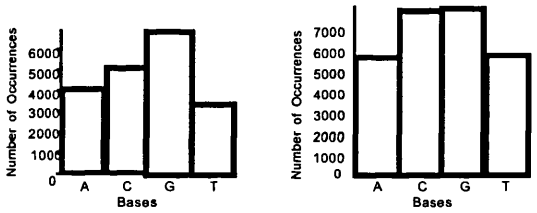


Figure 5.13: 26S change per site for one MPT



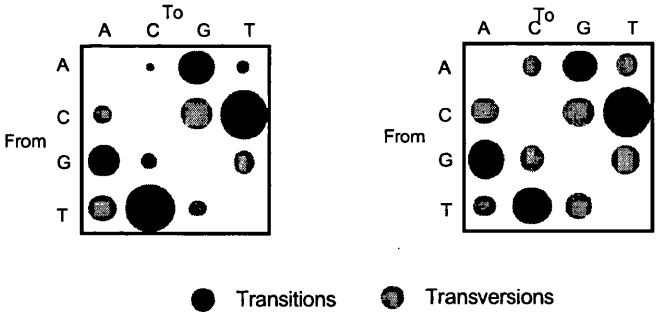
Both ITS and 26S are comparatively GC-rich (see Figure 5.14).

Figure 5.14: 26S ITS



Transitions outnumber transversions for 26S; the balance is less clear for ITS, which most notably has a higher number of changes from A to C than the 26S data set has, and also less changes from T to C (see Figure 5.15).

Figure 5.15: 26S ITS



5.4 26S analyses and taxon sampling

5.4.1 Introduction: In order to investigate the effects of taxon sampling on cladogram topology and support, and the relative support for different parts of the tree within the different data sets, MP analyses were run using different numbers of taxa from the 26S data set (D2 and D3) and from the ITS data set.

5.4.2 Material and methods: For each analysis (see Table 5.4), records were made of the number of parsimony informative characters in the data set, the number of MPTs (exhaustive searches for analyses 1 - 8; branch and bound for analyses 9 - 13, heuristics (1000 random additions, TBR) for analyses 14 and 15), MPT length, CI, RI, RC, g1 (exhaustive searches, analyses 1 - 8; 10,000 random trees, analyses 9 - 15), PTP excluding the outgroup (1000 replicates; branch and bound for analyses 1 - 7; heuristics for analyses 8 - 15 (10 random additions, TBR)), and the number of nodes with over 50% bootstrap support (10,000 replicates fast addition).

Table 5.4: Taxa included in different analyses

Analysis No.	Taxa (inclusive sets)
1. (4 taxa)	*Datisca glomerata* (OG)
	B. annobonensis (AF)
	B. iucunda (AF)
	B. meyeri-johannis (AF)
2. (5 taxa)	*B. thomeana* (AF)
3. (6 taxa)	*B. incarnata* (AM)
4. (7 taxa)	*B. geranioides* (S.AF)
5. (8 taxa)	*B. engleri* (AF)
6. (9 taxa)	*B. fallax* (AS)
7. (10 taxa)	*B. molleri* (AF)
8. (11 taxa)	*B. nossibea* (MAD)
9. (12 taxa)	*B. scapigera* (AF)
10. (13 taxa)	*B. convolvulacea*
11. (15 taxa)	*B. poculifera* (AF)
	B. salaziensis (AF)
12. (18 taxa)	*B. aequata* (AS)
	B. roxburghii (AS)
	B. luxurians (AM)
13. (20 taxa)	*B. francoisii* (MAD)
	B. socotrana (SOC)
14. (21 taxa)	*D. cannabina* (OG)
15. (36 taxa)	all species

(Key: AF = Africa; S.AF. = southern Africa; MAD = Madagascar; AM = America; AS = Asia; SOC = Socotra.)

5.4.3 Results: For the various tree statistics see Table 5.5 for 26S and Table 5.6 for ITS.

Adding taxa makes little difference to the numbers of parsimony-uninformative characters in the matrices, but the number of parsimony-informative characters increases as taxa are added. Consistency indices fall as taxon number increases, while the retention index and rescaled consistency indices both rise. g1 values are not significant (i.e. values are positive) for the analyses with less than 7 taxa. PTP values are insignificant (i.e. values are less than 0.05) for the analyses with less than 9 taxa. There are more clades which have bootstrap support in the ITS trees than there are in the 26S trees; in general, the amount of support rises as the number of taxa rises.

Admittedly, selecting different taxa to include or exclude from any of these analyses could have produced different results for any of these statistics; taxa were selected mainly in order to examine the amount of support for relationships between African taxa, which differ depending on the analytical methods and data sets used.

Table 5.5: Tree Statistics for different sized matrices, 26S

Analysis no.	no. taxa	no. pars. inf. chars	no. pars. uninf. chars	no. MPTs	tree length	CI	CI ex uninf.	RI	RC	g1	PTP	clades >50% bootstrap
1	4	5	37	2	51	0.941	0.625	0.400	0.377	0.716	1.000	0
2	5	9	40	1	63	0.921	0.688	0.444	0.409	0.409	0.866	0
3	6	12	42	1	74	0.892	0.680	0.429	0.382	0.324	0.952	0
4	7	19	36	7	78	0.859	0.711	0.542	0.465	-0.909	0.029	1
5	8	20	36	29	83	0.831	0.674	0.500	0.416	-1.075	0.073	1
6	9	22	34	8	88	0.807	0.667	0.540	0.436	-0.903	0.002	1
7	10	29	28	2	97	0.753	0.647	0.538	0.405	-0.849	0.001	2
8	11	32	28	11	104	0.731	0.627	0.569	0.416	-0.893	0.001	3
9	12	35	29	6	116	0.698	0.588	0.539	0.377	-0.805	0.001	3
10	13	36	30	6	122	0.689	0.578	0.568	0.391	-0.743	0.001	4
11	15	40	30	6	130	0.677	0.571	0.627	0.425	-0.710	0.001	6
12	18	40	36	6	141	0.681	0.559	0.688	0.468	-0.586	0.001	6
13	20	40	36	16	146	0.657	0.533	0.708	0.465	-0.596	0.001	7
14	21	50	28	28	151	0.656	0.573	0.735	0.482	-0.696	0.001	8
15	38	59	34	18	193	0.611	0.519	0.794	0.485	-0.521	0.001	12

Table 5.6: Tree statistics for different sized matrices, ITS

Analysis no.	no. taxa	no. pars. inf. chars	no. pars. uninf. chars	no. MPTs	tree length	CI	CI ex uninf.	RI	RC	g1	PTP	clades >50% bootstrap
1	4	32	204	1	307	0.951	0.681	0.531	0.505	-0.482	0.042	1
2	5	55	201	1	365	0.912	0.692	0.418	0.382	-1.007	0.078	2
3	6	81	187	1	424	0.863	0.676	0.383	0.331	0.002	0.066	2
4	7	104	171	3	456	0.833	0.683	0.441	0.368	-0.818	0.001	2
5	8	123	183	1	521	0.816	0.680	0.442	0.360	-0.779	0.001	3
6	9	133	177	1	557	0.786	0.656	0.431	0.339	-0.560	0.001	3
7	10	147	167	1	611	0.748	0.622	0.421	0.315	-0.611	0.001	3
8	11	159	163	2	669	0.712	0.590	0.401	0.285	-0.518	0.001	3
9	12	176	157	1	739	0.681	0.566	0.384	0.261	-0.582	0.001	4
10	13	186	155	1	792	0.667	0.555	0.397	0.265	-0.546	0.001	5
11	15	201	147	1	821	0.655	0.558	0.485	0.318	-0.779	0.001	7
12	18	217	145	3	917	0.619	0.531	0.504	0.312	-0.601	0.001	6
13	20	224	138	6	967	0.595	0.511	0.528	0.314	-0.676	0.001	7
14	21	256	111	1	989	0.590	0.527	0.581	0.343	-1.248	0.001	9
15	38	291	103	22	1298	0.503	0.453	0.638	0.321	-0.700	0.001	19

5.4.4 Discussion, taxon sampling:

5.4.4.1 Characters: The number of parsimony-uninformative characters does not change greatly with different numbers of taxa (although it changes more for ITS than for 26S), but the number of parsimony-informative characters rises as taxa are added. Some autapomorphies become synapomorphies as taxa are added (obviously it is not possible to turn a constant character into a synapomorphy by adding one taxon to a matrix). Our matrices include a wide range of the total taxonomic, geographic and thus, presumably, sequence, divergence within *Begonia*. Adding taxa appears to be more likely to make autapomorphies informative than to add more autapomorphies. Because adding more taxa turns autapomorphies into synapomorphies, this suggests that the information contents of the matrices are not saturated.

5.4.4.2 Indices: Consistency index falls as taxon number increases; this is a known affect of this statistic and is not necessarily related to the information content of the matrices. Retention index and the rescaled consistency index both rise as taxon number increase (although this is not a linear increase for the ITS data set). RI examines the actual homoplasy in a data set as a fraction of the maximum possible homoplasy, effectively giving a proportion of the similarities on a tree which are interpreted as synapomorphy

(Siebert, 1992). Thus it appears that as taxa are added, homoplasy levels decrease. Bininda-Edmonds, Bryant and Russell (1998) suggest that, in some reduced data sets, the removal of **consistent** parts of the cladogram (from the more inclusive data set) causes RC and RI to decline because, relatively, the regions which remain are more homoplastic. Thus our values for RI may be lower for the smaller data sets, particularly for 26S, because the parts of the matrix which have most cladistic structure are not present therefore are not contributing. The American and Asian sequences, which are less represented in these analyses than African sequences are, are comparatively similar; because there are few substitutions between the American and Asian taxa there may be a lower chance of multiple hits, and this may relate to lower levels of homoplasy within American and Asian clades.

5.4.4.3 Skewedness: The g1 values are not significant for analyses 1, 2 and 3 (which includes *Datisca*, four African taxa and one American taxon), indicating that there are many trees which are not significantly longer than the MPT, and consequently, that the phylogenetic signal in these matrices (i.e. between these taxa) is not strong.

5.4.4.4 Permutation tail probabilities: PTP tests whether character covariation within a matrix is greater than that expected given a random set of characters, and can be defined as the proportion of all data sets which give cladograms which are equal to or shorter than those produced by the unpermuted data. A value of 0.05 is often chosen to imply significant cladistic structure in the data (Kitching et al., 1998). PTP values for the ingroup are insignificant for the groups with less than 7 taxa for 26S and ITS. It is possible that only a few clades contain covarying characters and that these few clades are responsible for the significant statistic values. The addition of *B. geranioides* between analyses 3 and 4, and of *B. fallax* between analyses 5 and 6, both caused big decreases in the PTP value, while adding *B. engleri* between analyses 4 and 5 caused the value to rise. *B. geranioides* and *B. fallax* both form a clade with *B. incarnata*; the implication of the data is that many characters covary in this clade. The addition of *B. engleri*, on the other hand, decreases the covariance of the data.

Both g1 and PTP suggest that there is little cladistic structure in the matrices used in analyses 1, 2 and 3. On this basis, it may be that our data are not suitable for resolving the positions of *B. annobonensis*, *B. iucunda*, *B. meyeri-johannis*, *B. thomeana* and *B. incarnata* relative to each other and to the outgroup, *D. glomerata*.

5.4.4.5 Bootstrap: The number of nodes with over 50% bootstrap support rise as the number of taxa added rise. However, in general the supported nodes are not nodes already present in the less inclusive data sets, but are new nodes created by the addition of closer relatives to the taxa in the matrix. In a data set which includes less taxa, the reduced number of terminals can alter support because the numbers of possible alternative groupings of the taxa are decreased (and also, the number of characters per node can increase). Therefore, bootstrap support for groups on smaller trees can be relatively high even if there are also relatively high levels of homoplasy in the data (Bininda-Edmonds, Bryant and Russell,1998). Effectively this will mean that a bootstrap value, x, in an analysis of five taxa offers less confidence than the same value in an analysis of ten taxa.

5.5 General Discussion and conclusions

Of the nine topologies produced for the 38 taxa (MP, ML and ME analyses for 26S, ITS and combined sequence data sets) some conclusions can be drawn. There is a general agreement that African taxa are basal within *Begonia*, although precisely which African species are basal changes according to the data set and the analysis method used. From looking at reduced data sets, which include only some of these variably-placed taxa, it seems that we do not have sufficient information to resolve these problems.

5.5.1 Taxonomy: Most of the comments about the phylogeny of *Begonia* are reserved for a later chapter. However, a couple of general points can be made here.

Firstly, the placement of the Socotran endemic, *B. socotrana* (section *Peltaugustia*) is worth comment. This is consistently associated with an American and Asian clade (and most commonly, with the Indian species *B. floccifera* and *B. fallax*), rather than with the African species of section *Augustia* (of which, *B. geranioides* is the only one included in this analysis) with which it has traditionally been associated (e.g. Warburg, 1894; Irmscher, 1925, 1961).

Secondly, it is clear that *Begonia* as a genus is paraphyletic without the inclusion of the Asian taxon *Symbegonia*; this taxon always resolves well within the American/Asian clade.

5.5.2 Analysis methods - Which tree is truest?

Data sets: The combined ITS and 26S matrix should be accepted as our best approximation of a rDNA phylogeny for *Begonia*, because it is based on the most information. The assumption that the ITS and 26S data sets are fully congruent is no more problematic than the commonly made assumption that ITS 1 and ITS 2, or ITS and 5.8S, can be treated together (and, depending on the sites of the primers used, parts of the 26S gene are often appended to ITS 2 anyway). The combined data set produces a greater number of nodes with over 50% bootstrap support, therefore confidence levels are greater than for either of the constituent data sets.

Algorithms: Maximum parsimony and maximum likelihood produced broadly similar topologies. Homoplasy ('long branch attraction') can be tested for by comparing ML and MP cladograms; where the placement of long branches differs between analysis methods, the ML tree is frequently accepted as more reliable. For the 26S data set, all topological differences are within the Asia/America clade, where branch lengths are extremely short, and most differences are caused by lack of resolution in the parsimony tree.

For the ITS data set and the combined ITS-26S data set the position of *B. iucunda* changes most radically, from sister to the rest of *Begonia* (MP) to sister to *B. dewildei* (section *Scutobegonia*) and *B. scapigera* (section *Loasibegonia*) in the ML treatment. The placement of *B. meyeri-johannis* also changes, from sister to a wider African clade in the MP analysis, to sister to the mainly Madagascan clade in ML. Otherwise, the main differences lie in the basal relationships (with almost all the African species monophyletic in the ML trees, but paraphyletic in the MP trees). The data sets do not appear to offer enough support to decisively resolve these relationships. (NB: for the sections taxa are currently ascribed to, see the cladogram at the end of this chapter, Figure 5.16.)

The ML analyses, for ITS and for the combined data set, offers the topology most consistent with traditional *Begonia* taxonomy, by keeping the two species from section *Mezieria*, *B. salaziensis* and *B. meyeri-johannis*, closest. Also, although it is currently not placed in a section (*Ignota*, Sosef, 1994; Doorenbos, Sosef & de Wilde, 1998), *B. iucunda* has formerly been placed in section *Scutobegonia* (Irmscher, 1925) and latterly in section *Filicibegonia* (van den Berg, 1985; de Lange & Bouman, 1992), a section traditionally thought to be closely related to sections *Loasibegonia* and *Scutobegonia* (Sosef, 1994); in the ML tree it resolves as sister to *Loasibegonia*/ *Scutobegonia*. It is possible that the inconsistencies in the MP tree are caused by homoplasy, and therefore the ML tree may be a better representation of *Begonia* evolution - ITS has a lot of very variable characters which must heighten the possibility of multiple hits.

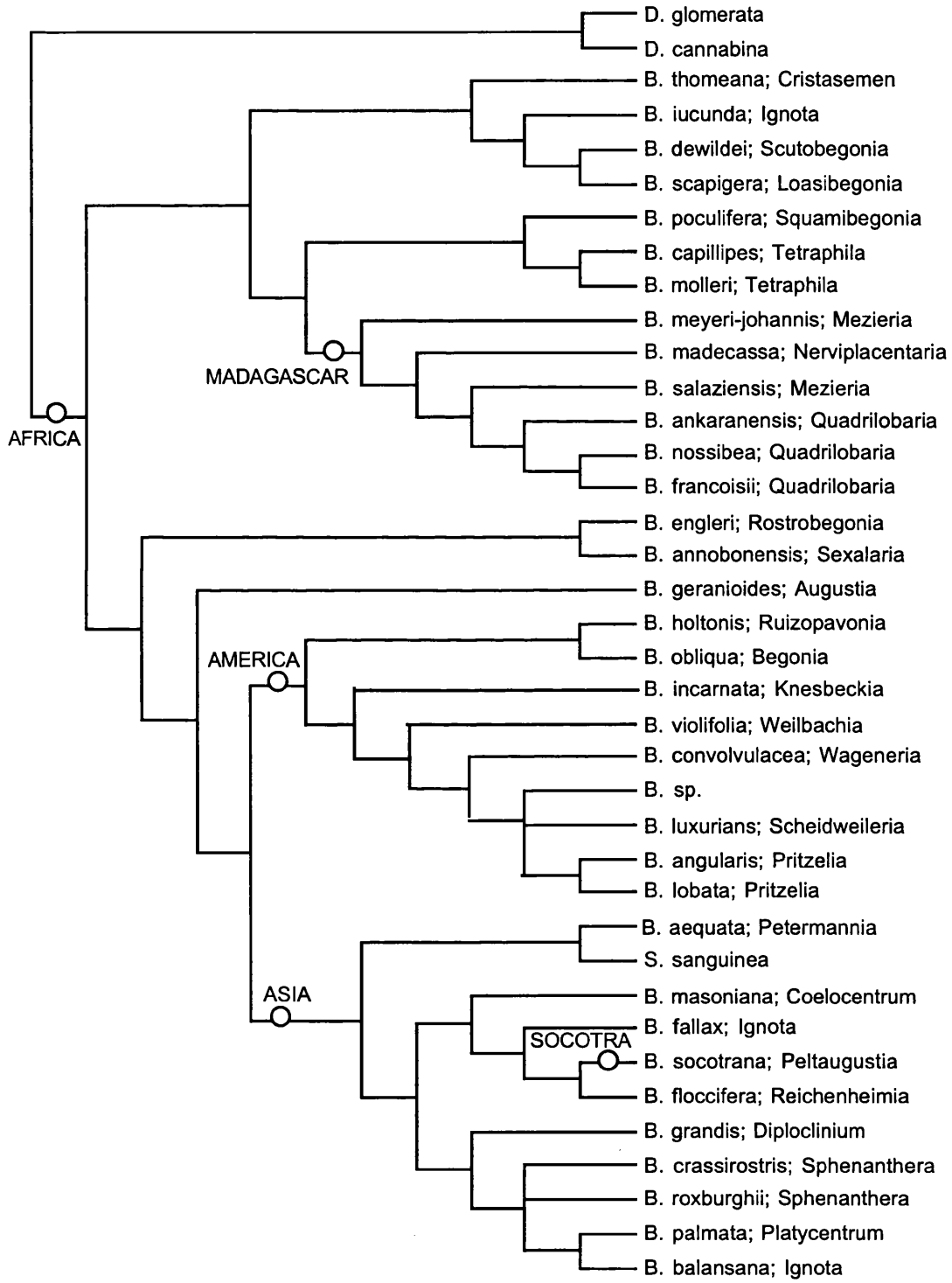
The next chapter concerns the analysis of a far larger ITS data set. ML is not a

practical analysis method for more than c. 60 taxa (Soltis & Soltis, 2000), and part of the point of including trees calculated using ME here was to see how they compared to the ML analyses, and so, whether the technique could be useful for the larger ITS data set as a comparison to MP. While ME produced an identical topology to ML for the 26S data set, it produced a rather different topology for the ITS data set (e.g. novel placements of the African *B. thomeana* (formerly in section *Loasibegonia*, now in a monotypic section, *Cristasemen* - de Wilde, 1985) and the south African *B. geranioides* (section *Augustia*)). Again, for the combined data set, the tree produced by ME was inconsistent with that produced by ML (e.g. the placement of the clade containing *B. engleri* and *B. annobonensis*). In the light of such topological differences between the ML and ME analyses, it was felt that there was little point in carrying out ME analyses in conjunction with MP for the wider ITS analysis.

The 'Truest' tree

The tree presented in Figure 5.16 is the maximum likelihood tree for the combined ITS and 26S data set; it is presented as a cladogram rather than a phylogram in order to emphasis the branching order; current sectional placements of the taxa have been marked on. This topology will be discussed at greater length in a further chapter.

Figure 5.16: 26S and ITS phylogeny for 36 Begoniaceae taxa, produced using ML (sectional placements as given in Doorenbos, Sosef and de Wilde, 1998)



5.6 Summary

Maximum parsimony (MP), maximum likelihood (ML) and minimum evolution (ME) cladograms were produced for 35 species of *Begonia*, one species of *Symbegonia* and two species of *Datisca*, for 26S, ITS and combined 26S-ITS sequence data. While MP and ML both produced highly congruent trees for the 26S-only sequence data, ML produced a tree more in line with traditional (morphological) *Begonia* classification for the ITS and combined 26S-ITS sequence data. This may relate to homoplasy caused by multiple hits in the more rapidly-evolving ITS region 'misleading' the parsimony algorithm. Minimum evolution as implemented here was not felt to produce reliable phylogenetic estimates for these data.

Although there is general agreement about the clades within *Begonia*, the order of branching of these clades can vary dramatically. From parsimony-based analyses of subsets of taxa for ITS and for 26S it appears that neither data set can conclusively resolve the question of what is most basal in *Begonia*. Time constraints did not allow similar analyses to be run on the combined 26S-ITS matrix; as it has more informative characters than either 26S or ITS alone, it may be more reliable. Certainly, Bremer support values for the backbone of the combined analysis MP cladogram (Figure 5.8) are higher than for either of the separate analyses (Figures 5.2, 5.5), suggesting that some characters from each data set (26S and ITS) are supporting the combined topology.

The main conclusion to be drawn from this chapter is that the best estimate we have to date for the phylogeny of *Begonia*, on the basis of information from the rDNA cistron for 36 species from the Begoniaceae, is the ML analysis of the combined data set (Figures 5.8, 5.16).

6. 26S: The wider picture - adding GenBank taxa

6.1 Introduction

There are several published studies which make use of the 26S region, and consequently, there are some 26S sequences available in GenBank. These were used to investigate the utility of the part of 26S sequenced in Begoniaceae within wider analyses of the angiosperms, and also to provide further corroboration of the position of *Datisca* relative to *Begonia* (and so, its utility as an outgroup for *Begonia* for ribosomal DNA sequence data). After all, if the data set is highly homoplastic in the variable positions, it is possible for *Datisca* to resolve within the Begoniaceae.

Due to the location of the primer site for p61 (and its reverse, p71) (Oxelman & Linden, 1995), a short region near the beginning of the 26S region could not be read. Aligning the incomplete sequences to complete sequences from GenBank allowed the size and precise position of this gap to be estimated.

6.2 Material and methods

The *Begonia* 26S region which encompasses D1 to D3 was put into a BLAST search to identify other similar sequences. Sequences from 26 genera in 17 angiosperm families (see Table 6.1) were downloaded. These were added to a data matrix of 6 *Begonia* species and one species of *Datisca*.

Table 6.1: GenBank sequences for 26S analysis

Taxon	Family	GenBank accession no.
Acorus gramineus	Acoraceae	AF036490
Tragopogon dubius	Asteraceae	AF036493
Jeffersonia diphylla	Berberidaceae	U52604
Heliotropium curassavicum	Boraginaceae	AF148274
Arabidopsis thaliana	Brassicaceea	X52320
Brassica napus	Brassicaceae	D10840
Sinapsis alba	Brassicaceae	X57137
Lobelia puberula	Campanulaceae	AF148276
Humulus lupus	Cannabaceae	AF223066
Ipomoea lacunosa	Convolvulaceae	AF146016
Jacquemontia tamnifolia	Convolvulaceae	AF148499
Eucryphia lucida	Cuoniaceae	AF036494
Hamamelis virginiana	Hamamelidaceae	AF036495
Oryza sativa	Poaceae	M11585
Polemonium reptans	Polemoniaceae	AF148282
Phlox divaricata	Polemoniaceae	AF148281
Hydrastis canadensis	Ranunculaceae	U52636

Thalictrum dioicum	Ranunculaceae	U52611
Citrus limon	Rutaceae	X05910
Jepsonia parryi	Saxifragaceae	AF036497
Lithophragma trifoliatum	Saxifragaceae	AF036501
Mitella pentandra	Saxifragaceae	AF036502
Lycopersicon esculentum	Solanaceae	X13557
Nolana humifuse	Solanaceae	AF148272
Physalis angulata	Solanaceae	AF148271
Drimys winteri	Winteraceae	AF036491
Begonia angularis	Begoniaceae	new sequence
Begonia crassirostris	Begoniaceae	new sequence
Begonia grandis	Begoniaceae	new sequence
Begonia molleri	Begoniaceae	new sequence
Begonia obliqua	Begoniaceae	new sequence
Begonia palmata	Begoniaceae	new sequence
Datisca glomerata	Datiscaceae	new sequence

Sequences were aligned in ClustalX (Thompson, Higgins & Gibson, 1997) and then manually augmented in SeqPup (Gilbert, 1995); ambiguous sites were excluded from the analyses. Data was analysed in PAUP* 4.0b2a (Swofford, 2000) using 1000 fast addition bootstrap replicates to isolate well supported clades.

6.3 Results

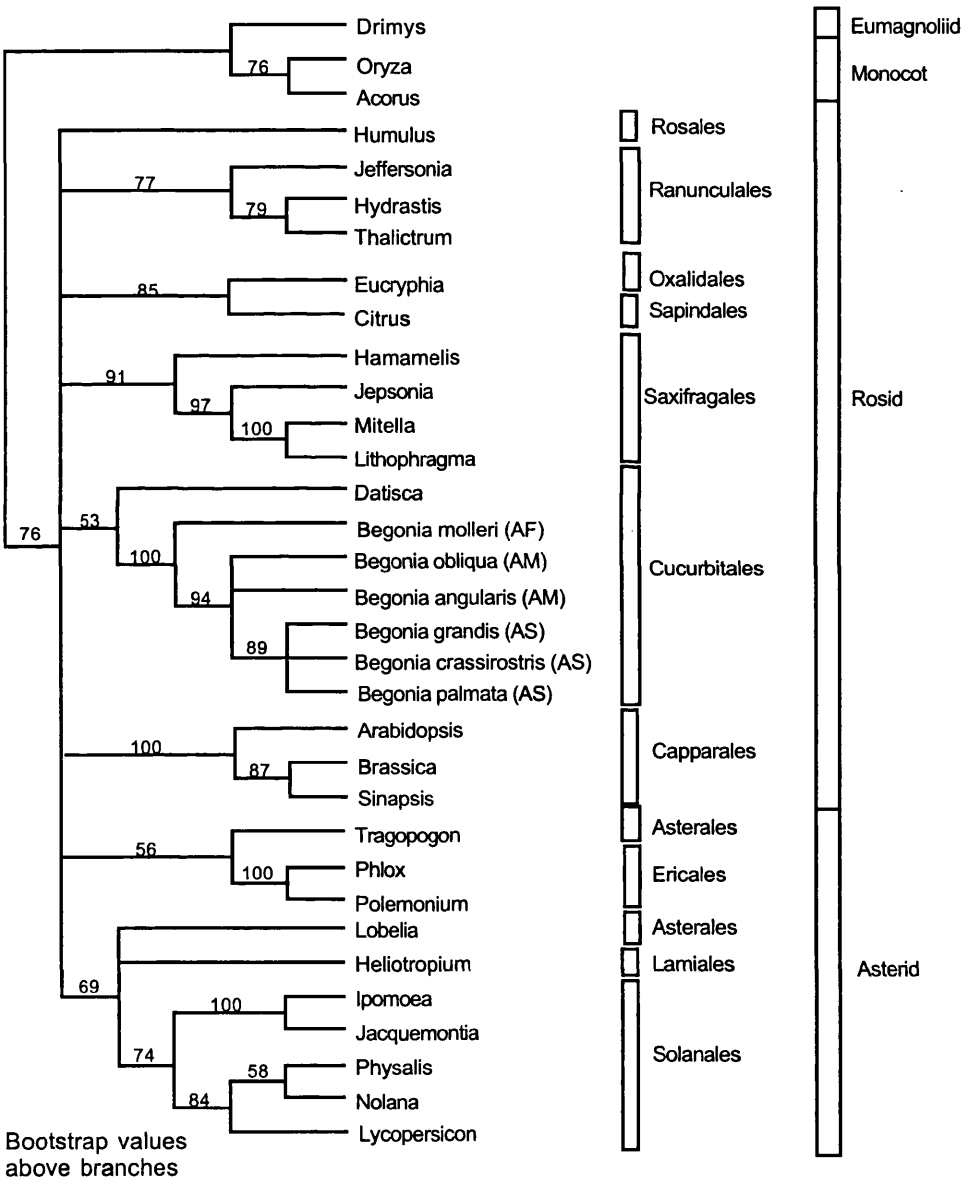
One of the results of this study was to identify the exact location of the end of D1 in the *Begonia* sequence, as it appears to correspond with the primer site for p61/p71. About 70 base pairs are unreadable in the *Begonia* and *Datisca* sequences, about 40 at the 3' end of the D1 region, and about 30 at the 5' end of the second conserved region, C2 (segments identified using the consensus motifs published by Kuzoff et al., 1998, p. 254). P71 (and therefore its reverse, p61) sits exactly at the beginning of the second conserved region, C2. The matrix (and the positions for the primer and for the variable regions D1, D2 and D3) are given in the Appendix, 14.6.

From the bootstrap consensus tree (Figure 6.1), *Begonia* is monophyletic, with 100% bootstrap support, with *Datisca* its sister group (53% bootstrap support). Within *Begonia*, the African species, *B. molleri*, is sister to the American and Asian species (94% support). The Asian species resolve as monophyletic with 89% support.

All angiosperm orders where more than one taxon was sampled (Ranunculales, Saxifragales, Cucurbitales, Capparales, Ericales and Solanales) are recovered by the bootstrap consensus tree, although the

position of these orders relative to each other is not resolved.

Figure 6.1: Bootstrap consensus tree, 10000 fast replicates for 26S D1, D2, D3 and linking regions (classification from Savolainen et al., 2000)



(Key: AF = Africa, Am = America, AS = Asia)

6.4 Discussion

The data support the monophyly of *Begonia* (see bootstrap consensus tree, Figure 6.1) and because the *Datisca* sequences do not nest within *Begonia* we can have additional confidence in the suitability of *Datisca* as an outgroup for the 26S data set. Of the included taxa, *Datisca* appears closest to *Begonia*. Although it is not necessary to root an analysis on the sister group, in cases where there are alignment difficulties even within the ingroup (as is the case for *Begonia* ITS), a close relationship between the ingroup and outgroup is preferable, for practical reasons.

The relationships resolved within *Begonia*, with the African species *B. molleri* basal and the Asian taxa as sister to the American species, are not inconsistent with those recovered by the analyses which included more taxa for 26S and ITS data (although shorter 26S sequences) in the previous chapter.

The lack of resolution between dicot clades is probably due to the region of 26S used; the region of sequence includes D1, D2 and D3, which are recommended for analysis within families (Kuzoff et al., 1998), while only short stretches of core regions, more suitable for between-family reconstruction, are included.

7. Building the cladogram - ITS

7.1 Introduction

There is no published molecular phylogeny for the Begoniaceae; previous authors (e.g. van den Berg, 1983, 1985 ; Legros & Doorenbos, 1969, 1971, 1973; Reitsma, 1984; Shui, Li & Huang, 1999; de Wilde & Arends, 1989) have attempted to interpret morphological and cytological patterns in the absence of any reliable genus-level phylogeny; some, like van den Berg (1985), de Wilde and Arends (1989) and Shui, Li and Huang (1999) have been tempted into hypothesising evolutionary directions based on supposition about what construes a primitive character within the genus or species group they are interested in. Schemske, Agren and le Corff (1996) resist this temptation, but consider that "information on the phylogeny of *Begonia* would be very useful for determining if the evolution of male and female floral traits is consistent with the intersexual mimicry hypothesis" (p. 313).

In this chapter, a molecular phylogeny is produced in order that evolutionary processes and patterns within the family may be discussed in its light in subsequent chapters.

7.2 Material and methods

7.2.1 Plant Material: The sources of plant material and vouchers used in this analysis are listed on the CD-ROM. In total I obtained ITS sequences from 163 individuals; 177 sequences were obtained, two from the outgroup and 175 from Begoniaceae. Sequences of *Hillebrandia* and *Datisca cannabina* accession 2 were kindly donated by Susan Swensen (Ithaca, New York). All three genera which are included in the Begoniaceae (*Begonia*, *Hillebrandia* and *Symbegonia*) were included in the analysis. The choice of *Datisca* as outgroup has been discussed in the previous chapter.

Identifications: Some of the taxa included in the analyses are unidentified. Several of these taxa are from China; the new Flora of China account has not been completed. Until it is, putting names onto species is highly problematic. Other unidentified taxa either do not exactly match known species from the region they were collected (e.g. the species collected in Bolivia, CAP 566), or are thought to be new species (e.g. the species from the Philippines, RBGE

1997 2566, for which a description is in preparation). In a few cases, morphologically interesting taxa were sampled using the names they were being cultivated under in E and/or GL, and only later was it realised that they were completely misidentified or mislabelled. *Begonia* species are very problematic to identify; there are relatively few keys, and even fewer which work. Given good collection details it can be possible to put names on plants; in the case of apparent labelling errors, where there is no information about the geographical origins of a taxon, it is virtually impossible. Of course, after the taxa have been sequenced and a phylogeny produced, there is more chance of naming the problem plant. Thus plants sequenced under the names *'B. macrocarpa'* (E and GL), an African species, are now known to each belong to different clades of American species. Likewise, the supposedly Asian *'B. sychnanthera'* (GL) and *'B. guttata'* (GL) are also mislabelled American plants.

One could argue that, given that there is no reliable way of knowing what these plants are or where they originated, the sequences should be deleted from the analysis. However, morphological data can be (and has been) gathered for these taxa; deleting them would amount to not using all the available information on the genus.

Unpublished names: *B. gabonensis* is the name given to a new species by de Wilde. It has not been published yet, but is given as a "provisional name" in de Lange and Bouman (1992, p. 29).

7.2.2 Molecular methods

The methods were as described in Chapter 5.2.

7.2.3 Sequence alignment

7.2.3.1. Automated alignment: Sixteen ClustalX (Thompson, Higgins & Gibson, 1997) alignments were produced using a range of gap opening and extension penalties (Table 7.1) (alignment one uses the default settings). The alignments were imported into SeqPup (Gilbert, 1995) and converted into Nexus format, then imported into MacClade 3.07 (Maddison & Maddison, 1992), where the edges of each alignment were checked and trimmed consistently.

Table 7.1: Automated Alignment Parameters:

Alignment Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Gap Opening Penalty	15	15	30	30	6.66	6.66	45	45	3.33	3.33	20	10	25	20	25	10
Gap Extension Penalty	6.67	15	6.67	15	6.66	15	6.66	15	6.66	3.33	6.66	6.66	6.66	15	15	15
Length of Matrix	1012	910	952	878	1087	928	905	859	1159	1413	1022	1035	1000	894	893	901

Files were saved without interleaving and opened in PAUP* 4.0b2a (Swofford, 2000). The data matrices were then copied into Word and arranged so that the taxa were in alphabetical order. Matrices were then copied back into the Nexus files in PAUP*, and saved. Each matrix was copied and pasted into one large Nexus file which was then interleaved to produce an elision data set (15848 characters long).

7.2.3.2. Manual alignment: An initial alignment was done in ClustalX (Thompson, Higgins & Gibson, 1997). This was transported into SeqPup 0.6f (Gilbert, 1995) for manual augmentation. Conserved ITS 1 and ITS 2 regions were identified from Hershkovitz, Zimmer and Hahn (1999) and Hershkovitz and Zimmer (1996). Secondary structure was determined at 20°c using MulFold (Zuker, 1989; Jaeger, Turner & Zuker, 1989a; Jaeger, Zuker & Turner, 1989b) and viewed in LoopDloop (Gilbert, 1995) (as described in chapter 8). Secondary structure information was used to help clarify ambiguous regions of alignment by identifying homologous regions (loops or stems) in ITS 2 (ITS 1 secondary structure proved too variable to be useful).

7.2.4 Phylogenetic analyses

Data matrices were analysed using the parsimony algorithm of the software package PAUP* 4.0b2a (Swofford, 2000). Searches were conducted on the 16 automatic alignments produced by ClustalX (Thompson, Higgins & Gibson, 1997), on the elision matrix, and on the manually aligned matrix. When searching on the automatic alignments and on the elision matrix, all data were included. For the manual alignment, regions were identified where the hypotheses of primary homology for bases were very tentative, and these regions were excluded (culled) from the analysis. For purposes of comparison, an analysis was also run with these regions included.

7.2.4.1 Automated alignments: Individual analyses were run for each of the 16 alignments, with 10 random addition replicates and no more than 10 trees of any length saved, swapping with TBR; the shortest trees from these searches were then used as starting trees in a second search, TBR swapping (MaxTrees set to 5100). Lengths and tree statistics from these analyses were compared with those from the elision analysis topology. To save time, bootstrap support measures were not obtained for these data sets, as the relative support for clades was not thought relevant to the study and the trees were not intended for discussing evolutionary scenarios (following Morrison & Ellis,1997).

7.2.4.2 Elision alignments: An heuristic analysis of the elision data set was carried out (1000 random addition replicates, TBR). Bootstrap support was estimated using 10,000 fast addition replicates.

7.2.4.3 Manual alignment:

A. Unculled: For the manual alignment, variable characters were first included and an analysis was run with 1000 random addition replicates, TBR, saving no more than 10 trees per step. These were then used as starting trees in a round of TBR swapping, MaxTrees set to 10,000. To check that there were no equally parsimonious trees with very different topologies, a constrain (the topology of the strict consensus of the MPTs) was imposed on a further round of analyses. 1000 random additions were performed, TBR, saving only trees which were not compatible with this topology. The shortest trees found were one step longer than the MPTs from the unconstrained analyses. Bootstrap support was estimated using 10,000 fast addition replicates.

B. Culled: Variable positions were excluded from the matrix, and the analysis was rerun (as above) with only two differences, that 10,000 random addition replicates were performed in the first step of the analysis, and Bremer support was estimated using AutoDecay (Erikkson, 1998) (10 random addition replicates per constraint tree, TBR).

Excluded characters are the same as those given in the previous chapter (1-183, 188, 200, 204, 211-217, 223-225, 230-249, 255-256, 266, 274-329, 340-366, 378, 383-384, 406-407, 415, 419-421, 426-428, 435-437, 444, 449-451, 460, 466-469, 475-483, 493-497, 503-507, 513-514, 539, 571, 577, 603, 606, 615, 649, 686, 688, 693-856, 886-901, 930-931, 944, 957-966, 983-984, 992-

993, 1013-1014, 1018, 1023, 1029-1035, 1041-1053, 1064-1093, 1110-1114, 1121-1122 and 1137-1154, see CD-ROM).

7.2.4.4 Tree comparisons: Because the total number of MPTs for all 16 alignments is very high (43,433) and because the numbers of MPTs for different alignments vary considerably (from six to 5100), which would effectively weight some alignments over others, consensus methods were used to compare the results of the analyses of the different alignments; thus only one tree was compared per alignment.

A strict consensus was computed of trees produced from each of the 16 alignments. These 16 strict consensus trees were used to compute a further strict consensus tree; this tree was unresolved except for a clade containing the three *Datisca* sequences. A majority rule tree (50%) had considerably more structure, and was used to look at areas of agreement between the cladograms produced from the different alignments.

Majority rule trees (50%) were also calculated for each of the 16 alignments, with other compatible groupings included. This was done in order to produce trees with the maximum possible resolution, as using unresolved trees makes some tree comparison statistics relatively meaningless. The symmetric-difference metric (or partition metric, PM)[3] and the agreement-subtree metric D_1[4], as implemented by PAUP* 4.0 (Swofford, 2000), were calculated between all pairs of trees.

A majority rule consensus of the fundamental elision trees was compared to the majority rule tree for each of the 16 alignments, by using the symmetric-difference metric PM and D_1.

[3] The partition metric (PM) can be defined as the minimum number of branch contractions and decontractions needed to transform one of two trees into the other. It is analogous, however, to a strict consensus tree in that the differential placement of just one taxon on two pectinate trees can give the highest possible value of PM (the maximum value possible is 2N - 6, where N = number of taxa) (Johnson & Soltis, 1998). PM was implemented in PAUP* 4.0 under the command TREEDIST METRIC = SYMDIFF (Swofford, 2000).

[4] PAUP* 4.0 (Swofford, 2000) can be used to compute a 'largest common pruned tree' (LCPT) which summarises a number of input trees, using the AGREE command. Two distance measures can be computed, D_1 and D (TREEDIST METRIC = AGD1 or AGR). Agree D is the number of taxa pruned to form the LCPT weighted by the distance between the taxa, while Agree D_1 is the minimum number of taxa which must be pruned to make two trees identical (Johnson & Soltis, 1998).

The strict consensus trees for each of the 16 alignments were used to look at the presence/absence of certain clades, the monophyly of certain geographical regions and the position both of the outgroup and of *Hillebrandia*.

7.3 Results

7.3.1 Statistics: For statistics concerning the various data sets and concerning the trees produced from their analysis, see Table 7.2.

Table 7.2: Statistics for the various automated alignments and for the elision, the manual, culled and the manual, unculled alignments.

	Al. 1	Al. 2	Al. 3	Al. 4	Al. 5	Al. 6	Al. 7	Al. 8	Al. 9	Al. 10	Al. 11	Al. 12	Al. 13	Al. 14	Al. 15	Al. 16	Elision	Man., Total	Man., Cull
Opening penalty	15	15	30	30	6.7	6.7	45	45	3.3	3.3	20	10	25	20	25	10			
Extension penalty	6.7	15	6.7	15	6.7	15	6.7	15	6.7	3.3	6.7	6.7	6.7	15	15	15			
No. char.s	1012	910	952	878	1087	928	905	859	1159	1413	1022	1035	1000	894	893	901	15848	993	522
No. invariant char.s	120	78	104	54	191	104	77	58	213	405	171	159	154	69	43	72	2067	191	101
No. pars. uninf. char.s	171	112	142	116	182	124	107	90	190	273	149	160	158	121	112	114	2321	137	92
No. pars. inform. char.s	721	720	706	708	714	700	721	711	756	735	702	716	688	704	738	715	11455	665	329
No. trees	5100	5100	360	5100	120	5100	128	5100	80	5100	140	1656	6	5100	143	5100	6	100	10000
Length	9983	10325	11188	11205	9468	9594	12456	11769	9621	8615	10264	9634	10234	10206	11074	9741	175227	5979	2844
CI ex uninf.	0.188	0.183	0.163	0.17	0.194	0.188	0.152	0.161	0.198	0.215	0.175	0.192	0.174	0.183	0.176	0.192	0.17	0.271	0.267
RI	0.48	0.531	0.472	0.54	0.484	0.564	0.466	0.556	0.455	0.475	0.476	0.495	0.488	0.528	0.525	0.551	0.4717	0.691	0.691
RC	0.099	0.103	0.084	0.098	0.104	0.114	0.075	0.094	0.1	0.118	0.091	0.105	0.093	0.103	0.099	0.113	0.087	0.202	0.205
invar./ total char.s	0.119	0.086	0.109	0.062	0.176	0.112	0.085	0.068	0.184	0.287	0.167	0.154	0.154	0.077	0.048	0.080	0.130	0.192	0.193
uninform/ total char.s	0.169	0.123	0.149	0.132	0.167	0.134	0.118	0.105	0.164	0.193	0.146	0.155	0.158	0.135	0.125	0.127	0.146	0.138	0.176
inform./ total char.s	0.712	0.791	0.742	0.806	0.657	0.754	0.797	0.828	0.652	0.520	0.687	0.692	0.688	0.787	0.826	0.794	0.723	0.670	0.630

The number of characters in the different alignments range from 859 (alignment 8) to 1413 (alignment 10) (i.e. by 554 characters); the number of invariant characters range from 43 (alignment 15) to 405 (alignment 10) (i.e. by 362 characters); the number of parsimony uninformative characters range from 90 (alignment 8) to 273 (alignment 10) (i.e. by 183 characters). However, the number of parsimony-informative characters only range from 688 (alignment 13) to 756 (alignment 9) (a difference of 68 characters).

The percentage of invariant characters in each alignment ranges from 6.2% (alignment 4) to 28.7% (alignment 10); uninformative characters range from 10.5% (alignment 8) to 16.9% (alignment 1) and informative characters range from 52.0% (alignment 10) to 82.8% (alignment 7).

Characters - automated alignments: Changes in gap opening and gap extension penalties have an effect on the proportion of different types of characters in the matrices. Table 7.3 is a summary of the values from Table 7.2.

Table 7.3: The Effect of Alignment on Characters

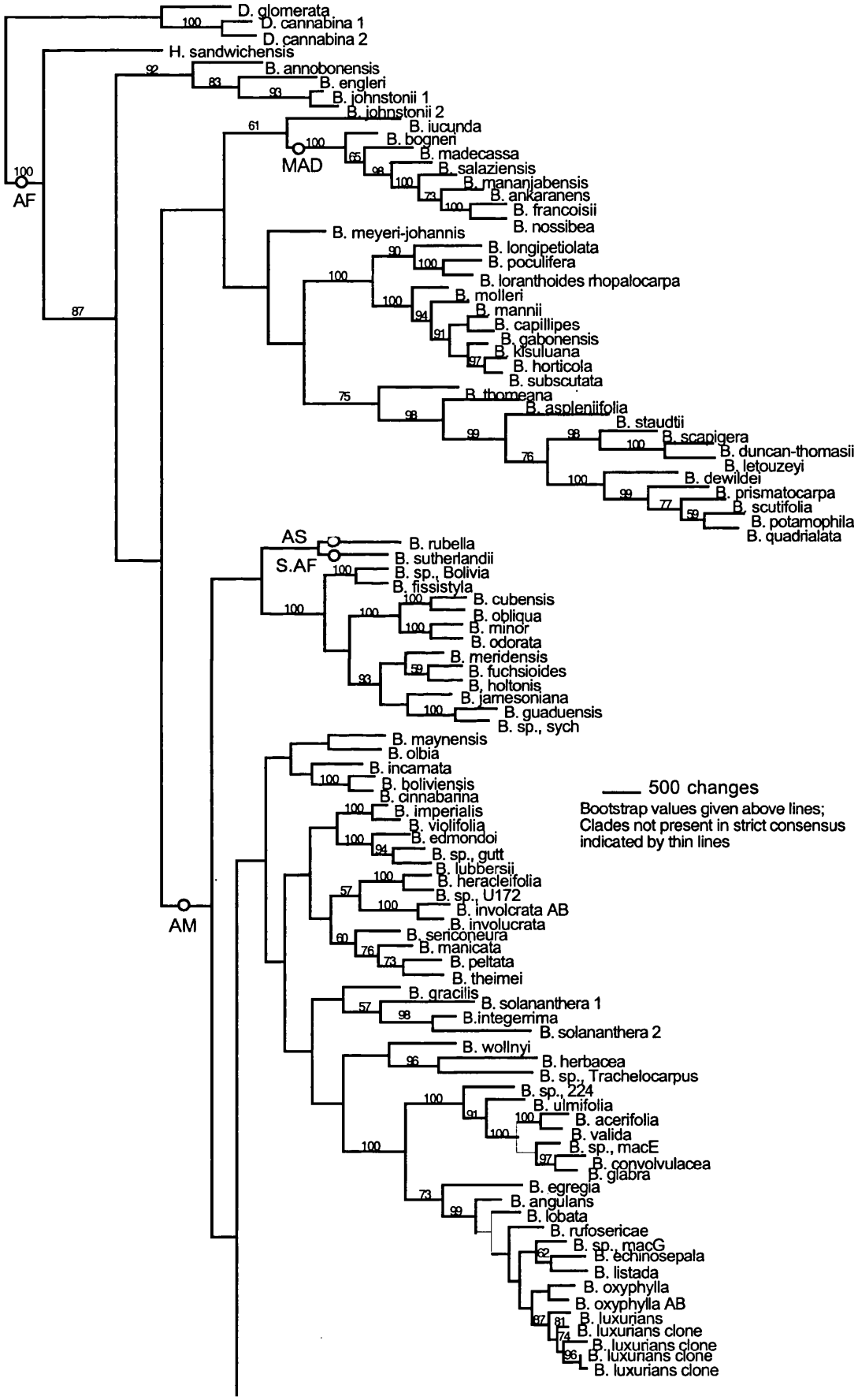
	TOTAL CHARACTER NUMBER	CONSTANT CHARACTER NUMBER	PARSIMONY-UNINFORMATIVE CHARACTER NUMBER	PARSIMONY-INFORMATIVE CHARACTER NUMBER
opening penalty increasing, extension penalty constant	usually increases	often increases	usually increases	no clear trend
extension penalty increasing, opening penalty constant	always increases	always increases	always increases	no clear trend

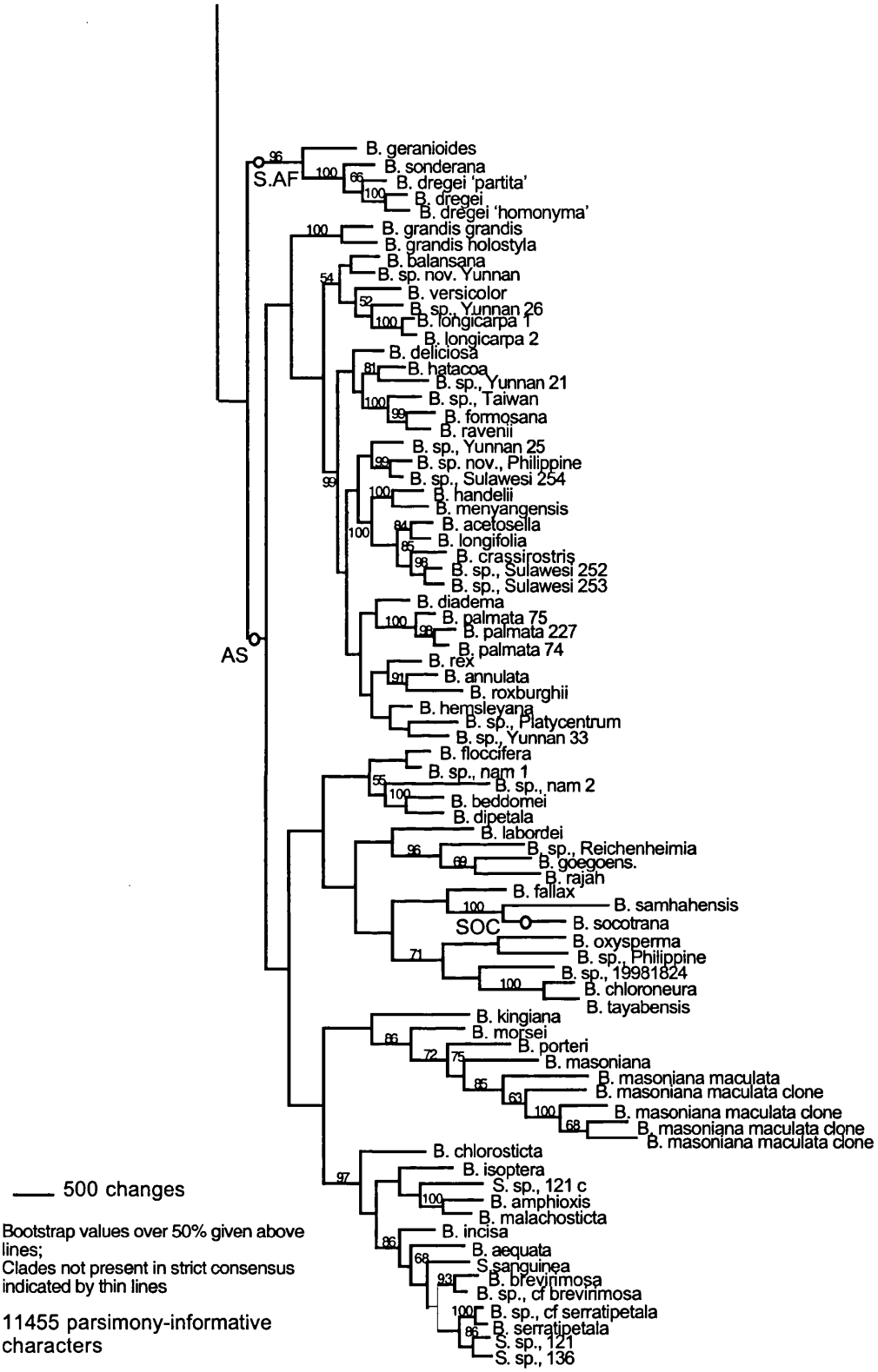
Because eight different gap opening penalties were used, while only three gap extension penalties were used, the clear effect of increasing the gap extension penalty when the gap opening penalty is constant may be lost if more extension penalties are tested.

7.3.2. Trees: The tree statistics obtained from the elision topology and from swapping each of the constituent data sets (automated alignments) are presented in Table 7.2. One of the phylograms produced from analysis of the elision data set is presented (Figure 7.1); clades which collapse in the strict consensus tree are marked on. Also a majority rule tree of the strict consensus trees for the 16 alignments is shown, as a summary of the areas of most agreement between alignments (Figure 7.2); strict consensus trees and phylograms are also shown for the culled (Figures 7.3, 7.4) and the unculled (Figures 7.5, 7.6) manual alignments.

For all the cladograms presented, where geographical clades are marked on, AF is Africa, S. AF is southern Africa, MAD is Madagascar, SOC is Socotra, AM is America, and AS is Asia.

Figure 7.1: Phylogram from analysis of the ITS elision matrix



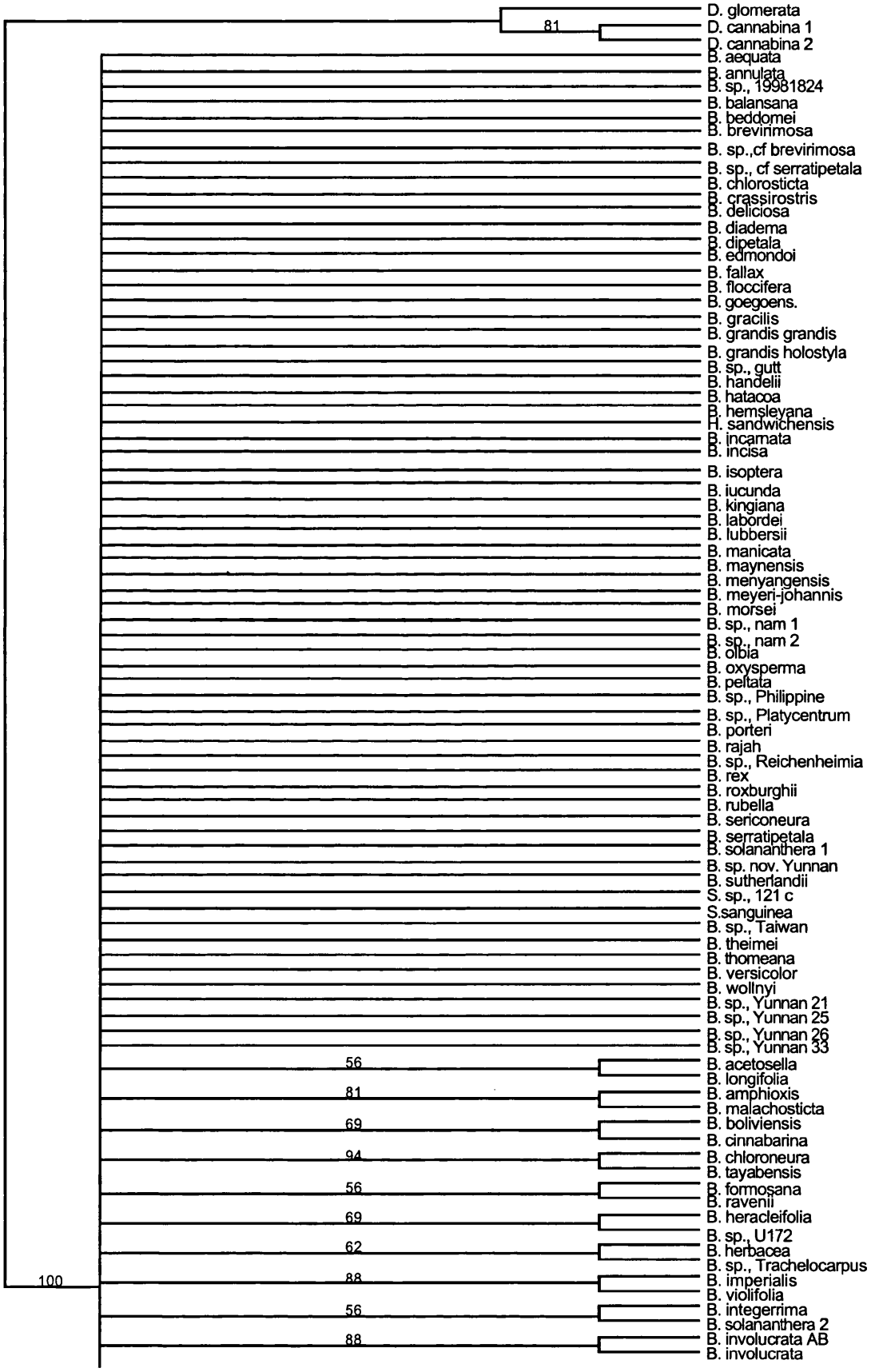


Bootstrap values over 50% given above lines;
Clades not present in strict consensus indicated by thin lines

11455 parsimony-informative characters

Tree length = 175227
CI = 0.1846
CI ex uninformative = 0.1697
RI = 0.4717

Figure 7.2: Majority rule tree of the strict consensus trees from analyses of the 16 automated alignments



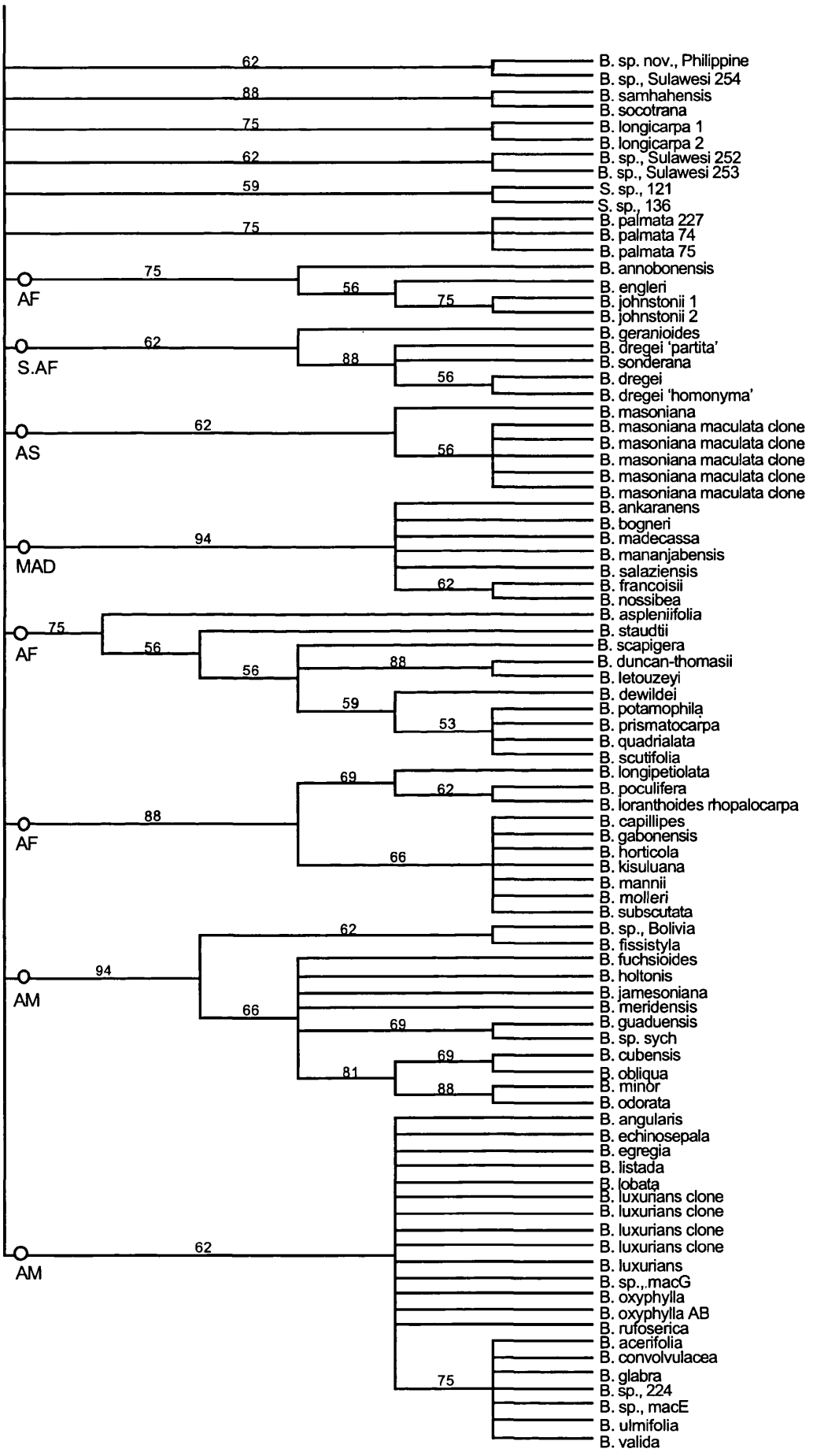
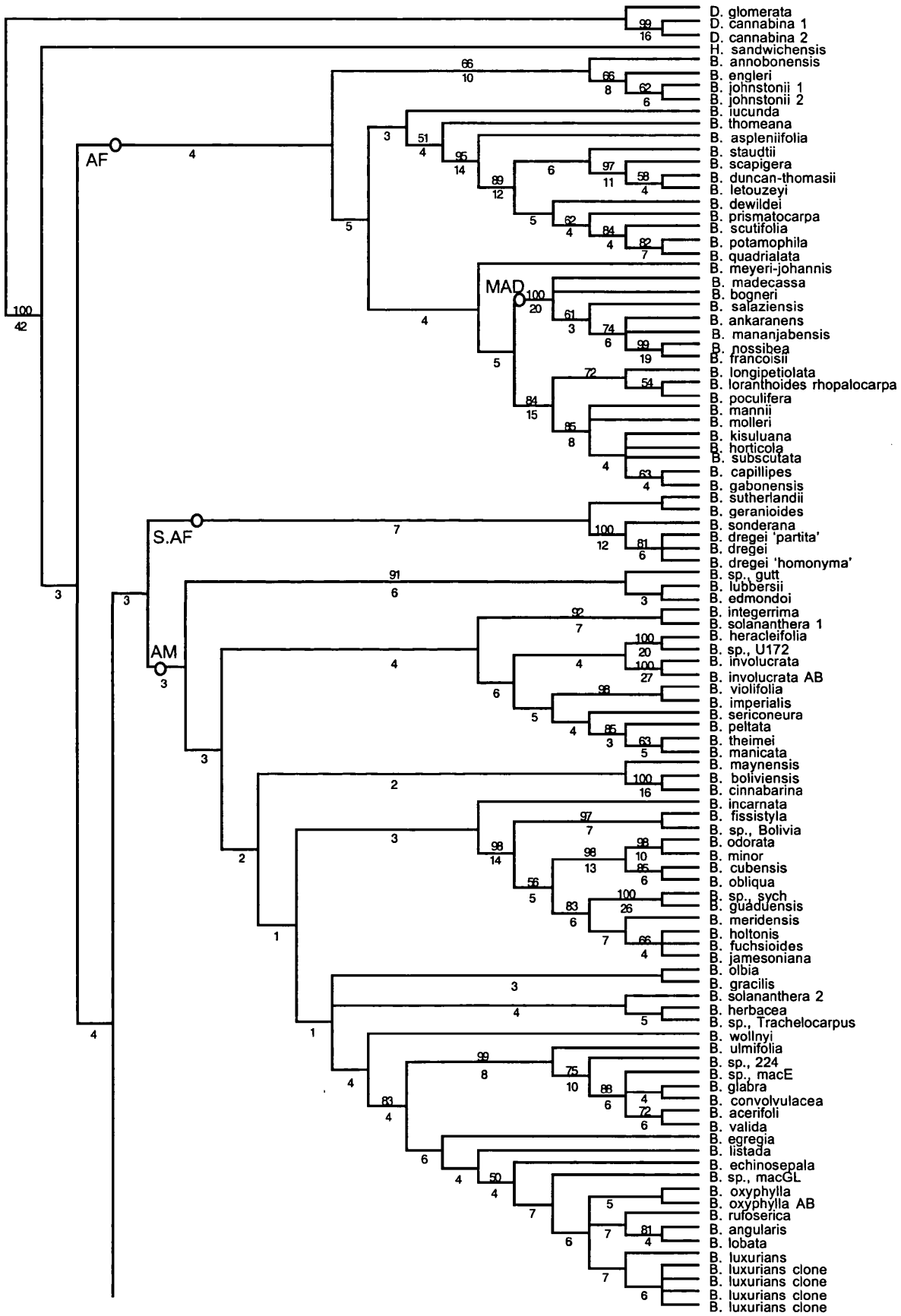

B. sp. nov., Philippine
B. sp., Sulawesi 254
B. samhahensis
B. socotrana
B. longicarpa 1
B. longicarpa 2
B. sp., Sulawesi 252
B. sp., Sulawesi 253
S. sp., 121
S. sp., 136
B. palmata 227
B. palmata 74
B. palmata 75
B. annobonensis
B. engleri
B. johnstonii 1
B. johnstonii 2
B. geranioides
B. dregei 'partita'
B. sonderana
B. dregei
B. dregei 'homonyma'
B. masoniana
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. ankaranens
B. bogneri
B. madecassa
B. mananjabensis
B. salaziensis
B. francoisii
B. nossibea
B. aspleniifolia
B. staudtii
B. scapigera
B. duncan-thomasii
B. letouzeyi
B. dewildei
B. potamophila
B. prismatocarpa
B. quadrialata
B. scutifolia
B. longipetiolata
B. poculifera
B. loranthoides rhopalocarpa
B. capillipes
B. gabonensis
B. horticola
B. kisuluana
B. mannii
B. molleri
B. subscutata
B. sp., Bolivia
B. fissistyla
B. fuchsioides
B. holtonis
B. jamesoniana
B. meridensis
B. guaduensis
B. sp. sych
B. cubensis
B. obliqua
B. minor
B. odorata
B. angularis
B. echinosepala
B. egregia
B. listada
B. lobata
B. luxurians clone
B. luxurians clone
B. luxurians clone
B. luxurians clone
B. luxurians
B. sp., macG
B. oxyphylla
B. oxyphylla AB
B. rufoserica
B. acerifolia
B. convolvulacea
B. glabra
B. sp., 224
B. sp., macE
B. ulmifolia
B. valida
AF
S.AF
AS
MAD
AF
AF
AM
AM
62
88
75
62
59
75
75
56
75
62
88
56
62
56
94
62
75
56
56
88
59
53
69
62
88
66
62
94
66
69
69
81
88
62
75

Figure 7.3: Strict consensus of 10,000 MPTs, culled manual ITS alignment



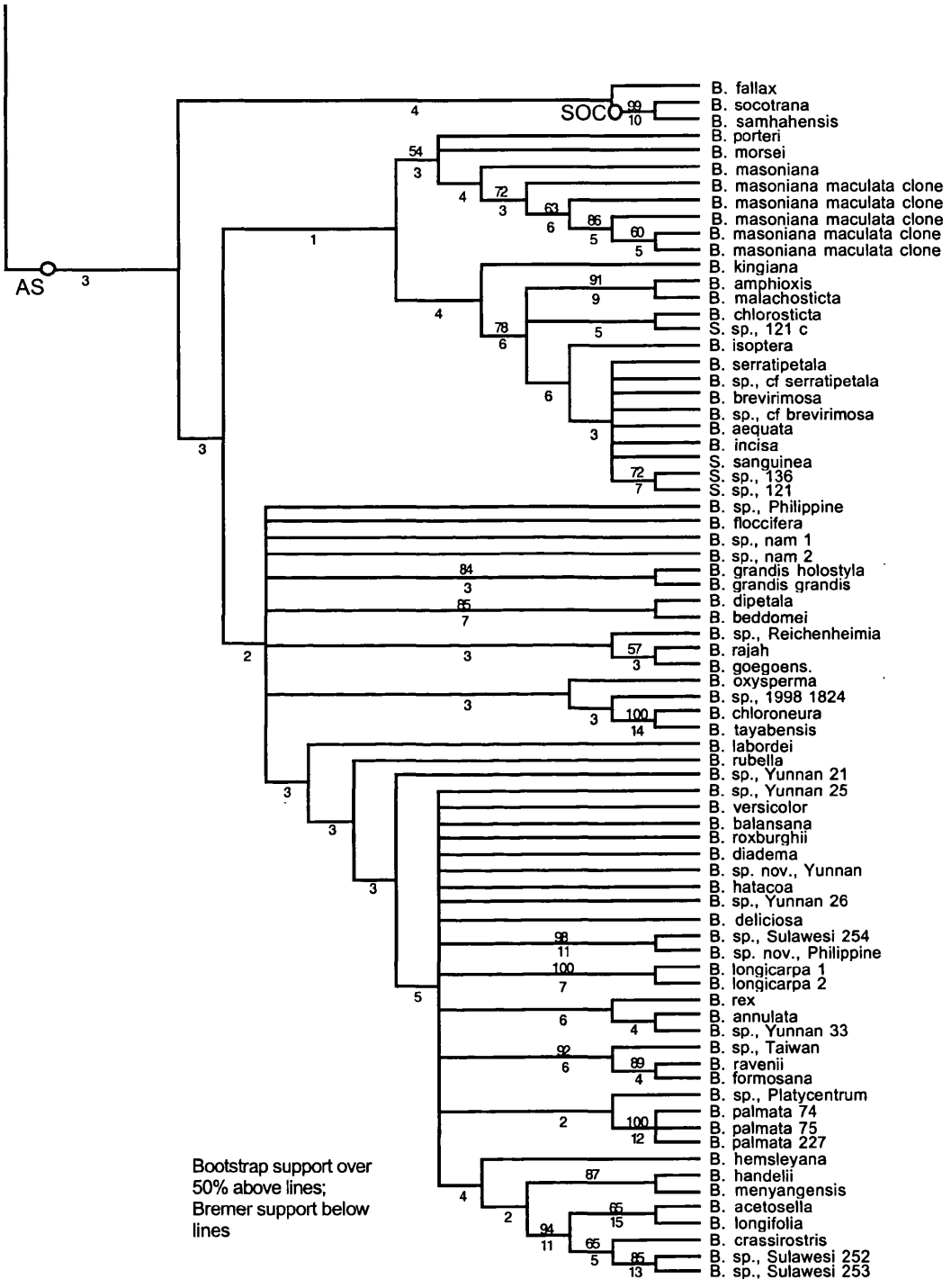
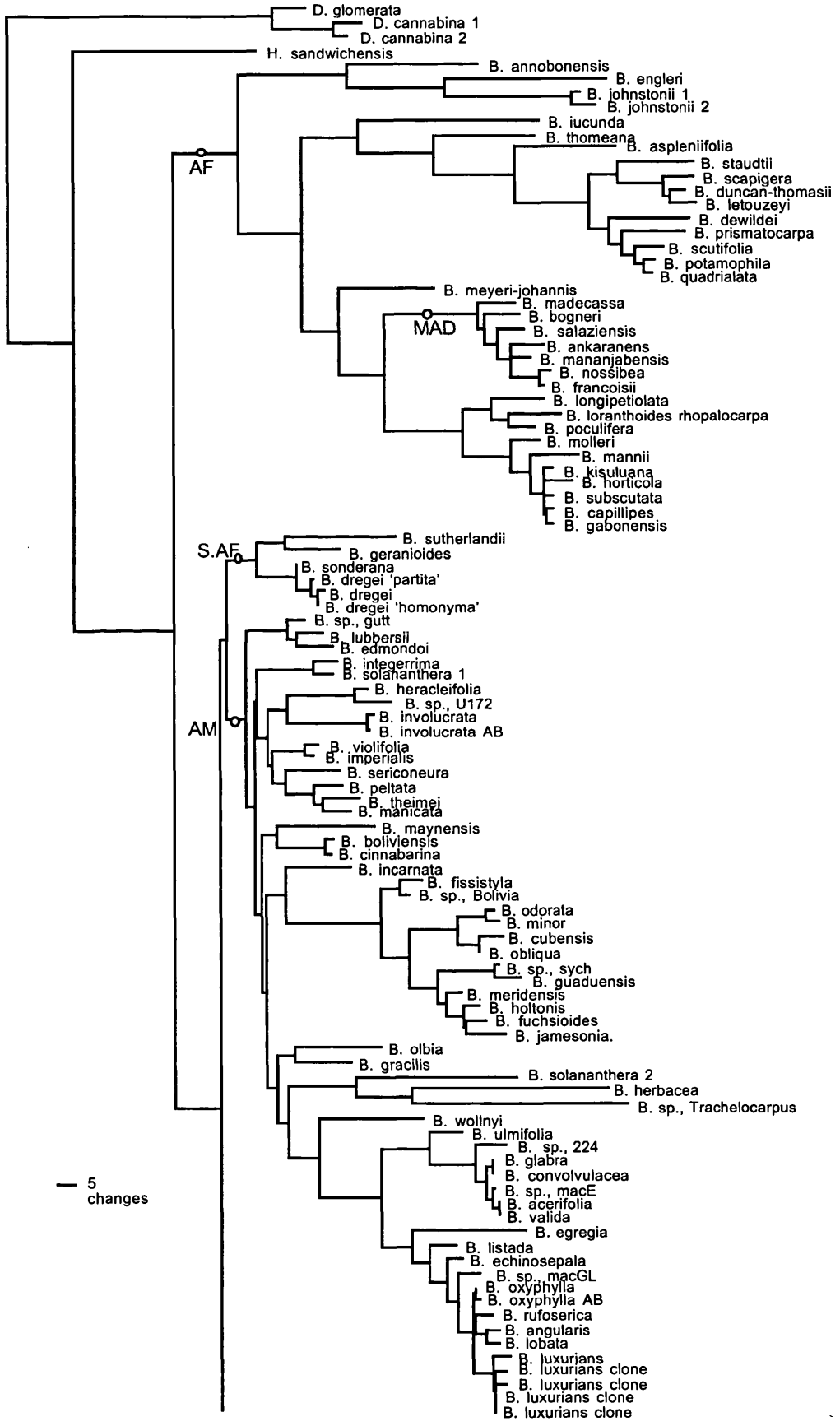

AS
SOCO
B. fallax
B. socotrana
B. samhahensis
B. porteri
B. morsei
B. masoniana
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. kingiana
B. amphioxis
B. malachosticta
B. chlorosticta
S. sp., 121 c
B. isoptera
B. serratipetala
B. sp., cf serratipetala
B. brevirimosa
B. sp., cf brevirimosa
B. aequata
B. incisa
S. sanguinea
S. sp., 136
S. sp., 121
B. sp., Philippine
B. floccifera
B. sp., nam 1
B. sp., nam 2
B. grandis holostyla
B. grandis grandis
B. dipetala
B. beddomei
B. sp., Reichenheimia
B. rajah
B. goegoens.
B. oxysperma
B. sp., 1998 1824
B. chloroneura
B. tayabensis
B. labordei
B. rubella
B. sp., Yunnan 21
B. sp., Yunnan 25
B. versicolor
B. balansana
B. roxburghii
B. diadema
B. sp. nov., Yunnan
B. hatacoa
B. sp., Yunnan 26
B. deliciosa
B. sp., Sulawesi 254
B. sp. nov., Philippine
B. longicarpa 1
B. longicarpa 2
B. rex
B. annulata
B. sp., Yunnan 33
B. sp., Taiwan
B. ravenii
B. formosana
B. sp., Platycentrum
B. palmata 74
B. palmata 75
B. palmata 227
B. hemsleyana
B. handelii
B. menyangensis
B. acetosella
B. longifolia
B. crassirostris
B. sp., Sulawesi 252
B. sp., Sulawesi 253
Bootstrap support over 50% above lines; Bremer support below lines

Figure 7.4: Phylogram for the culled manual ITS alignment; one of 10,000 MPTs



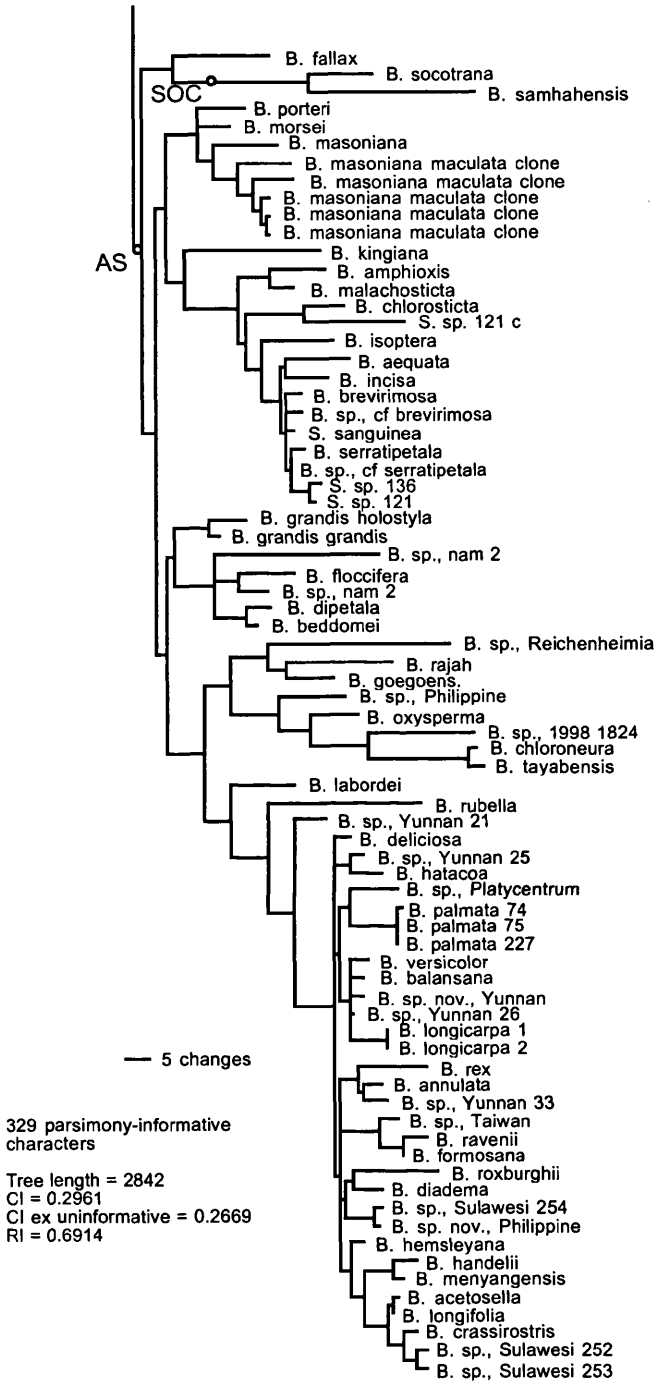
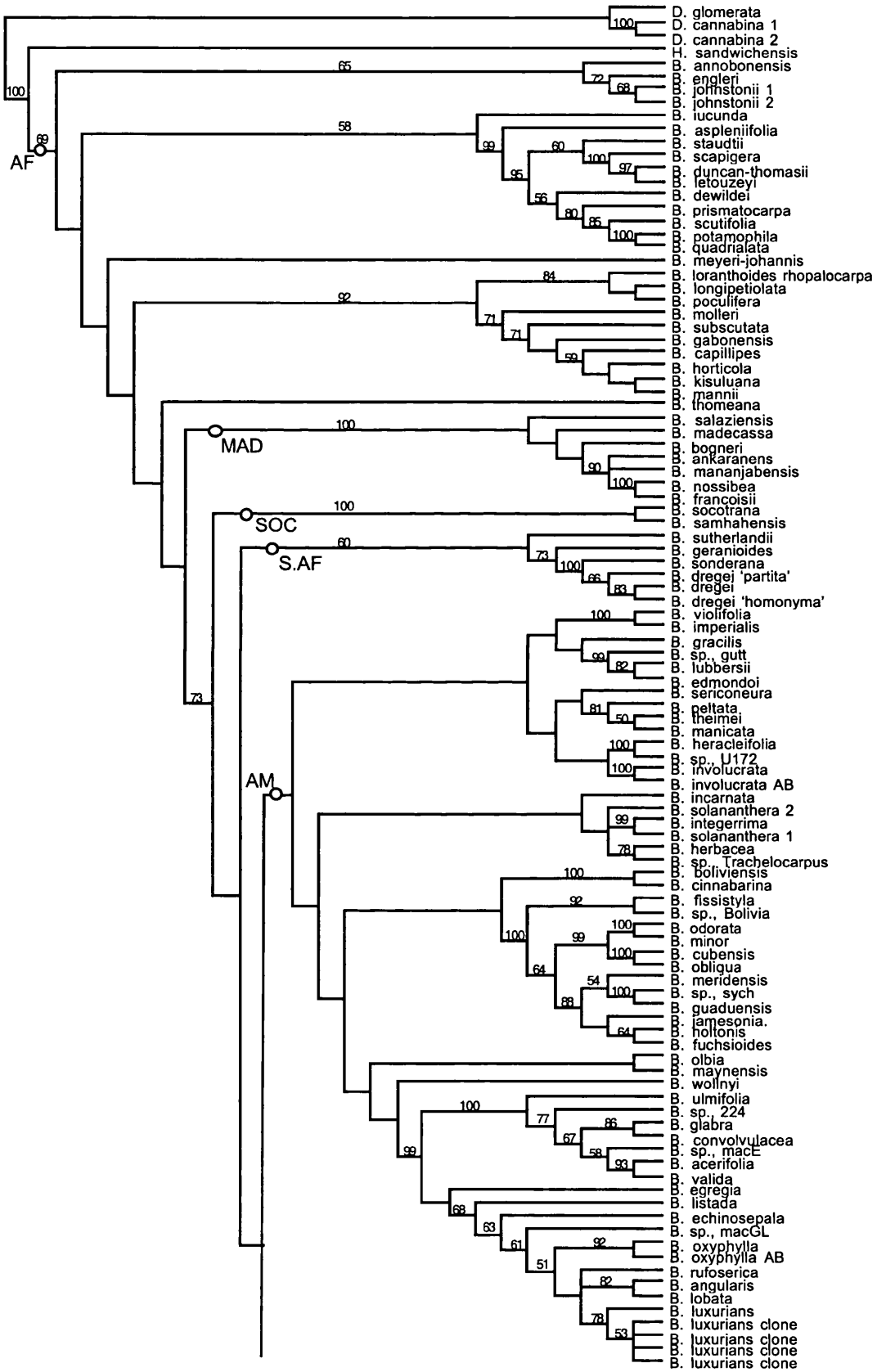

B. fallax
B. socotrana
SOC
B. samhahensis
B. porteri
B. morsei
B. masoniana
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
AS
B. kingiana
B. amphioxis
B. malachosticta
B. chlorosticta
S. sp. 121 c
B. isoptera
B. aequata
B. incisa
B. brevirimosa
B. sp., cf brevirimosa
S. sanguinea
B. serratipetala
B. sp., cf serratipetala
S. sp. 136
S. sp. 121
B. grandis holostyla
B. grandis grandis
B. sp., nam 2
B. floccifera
B. sp., nam 2
B. dipetala
B. beddomei
B. sp., Reichenheimia
B. rajah
B. goegoens.
B. sp., Philippine
B. oxysperma
B. sp., 1998 1824
B. chloroneura
B. tayabensis
B. labordei
B. rubella
B. sp., Yunnan 21
B. deliciosa
B. sp., Yunnan 25
B. hatacoa
B. sp., Platycentrum
B. palmata 74
B. palmata 75
B. palmata 227
B. versicolor
B. balansana
B. sp. nov., Yunnan
B. sp., Yunnan 26
B. longicarpa 1
B. longicarpa 2
— 5 changes
B. rex
B. annulata
B. sp., Yunnan 33
B. sp., Taiwan
B. ravenii
B. formosana
B. roxburghii
B. diadema
B. sp., Sulawesi 254
B. sp. nov., Philippine
B. hemsleyana
B. handelii
B. menyangensis
B. acetosella
B. longifolia
B. crassirostris
B. sp., Sulawesi 252
B. sp., Sulawesi 253
329 parsimony-informative characters
Tree length = 2842
CI = 0.2961
CI ex uninformative = 0.2669
RI = 0.6914

Figure 7.5: Strict consensus of 100 MPTs, unculled manual ITS alignment



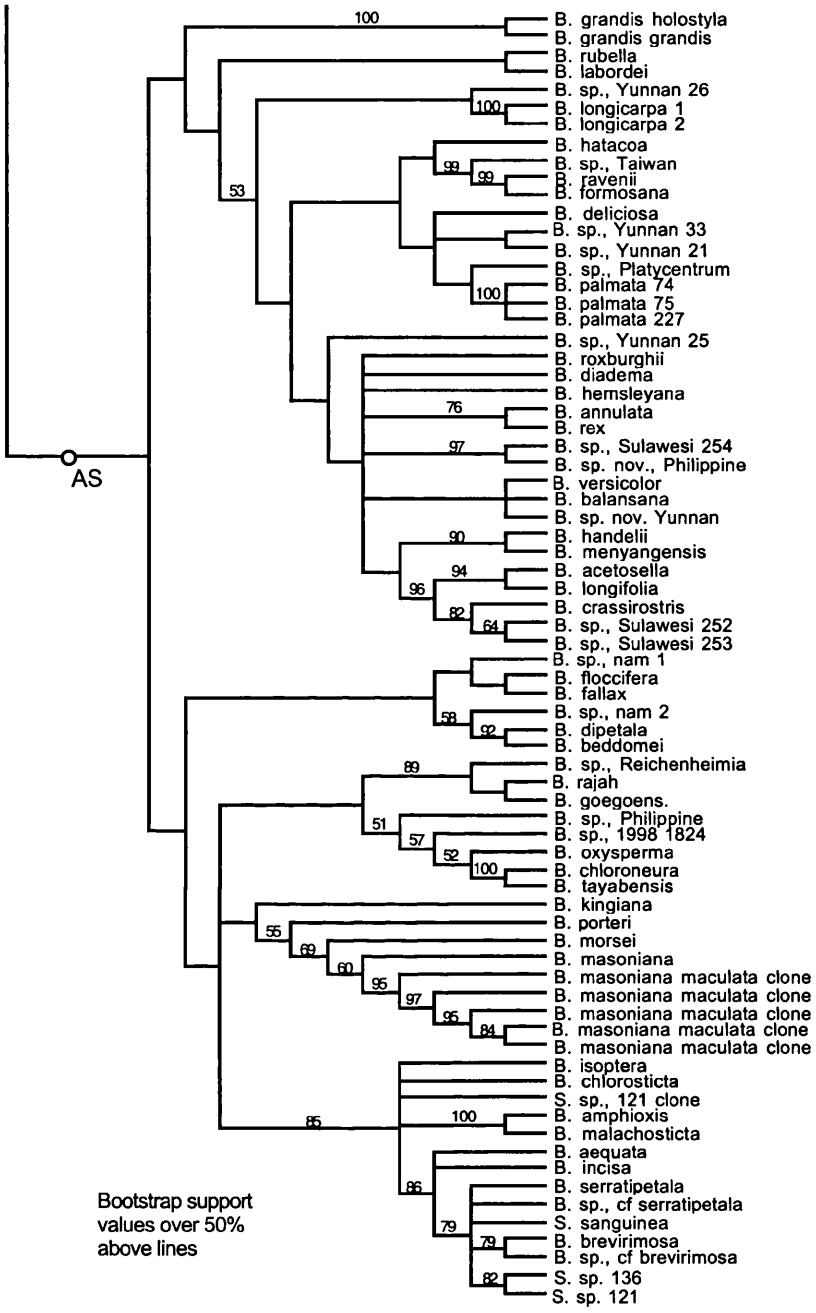
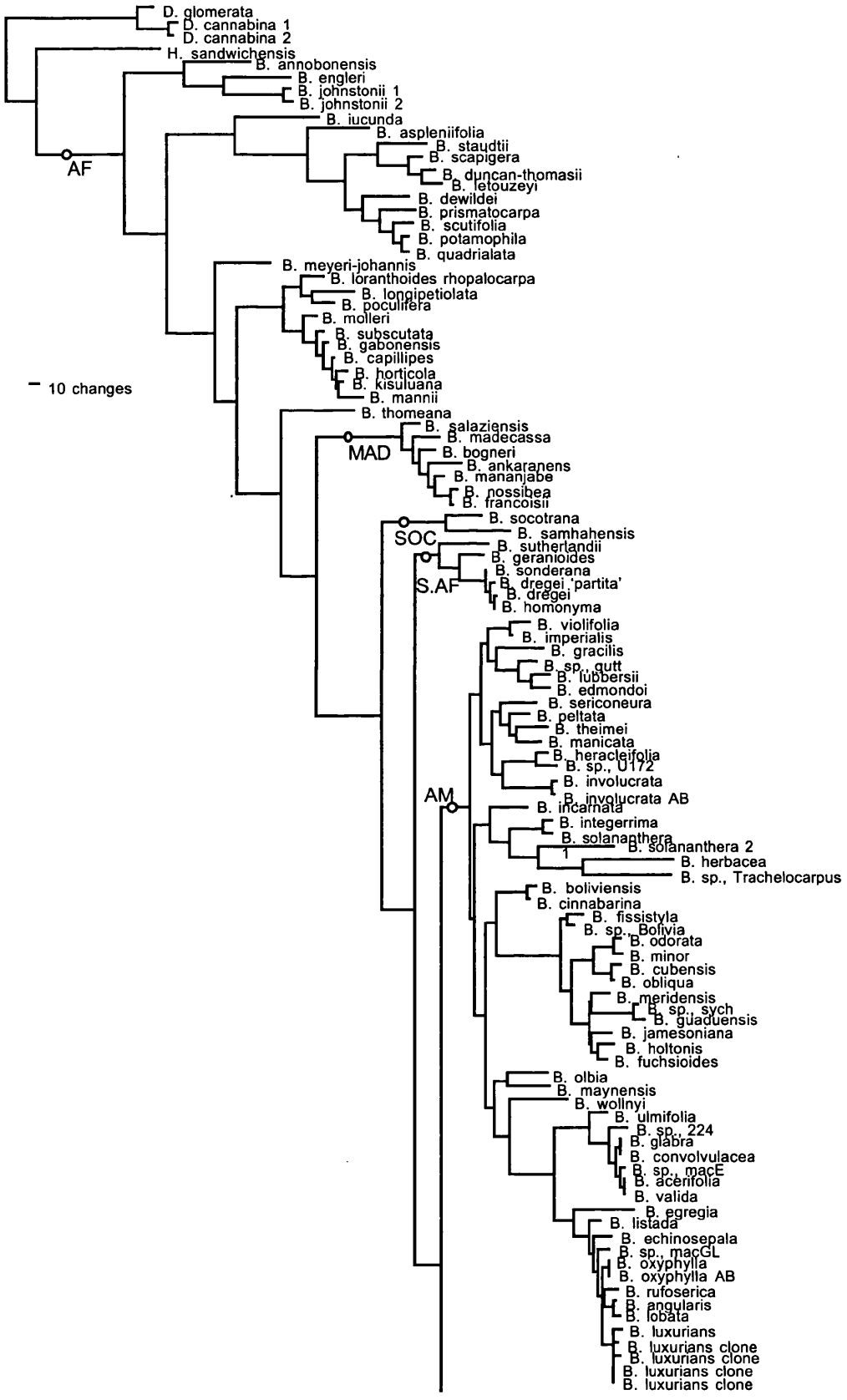
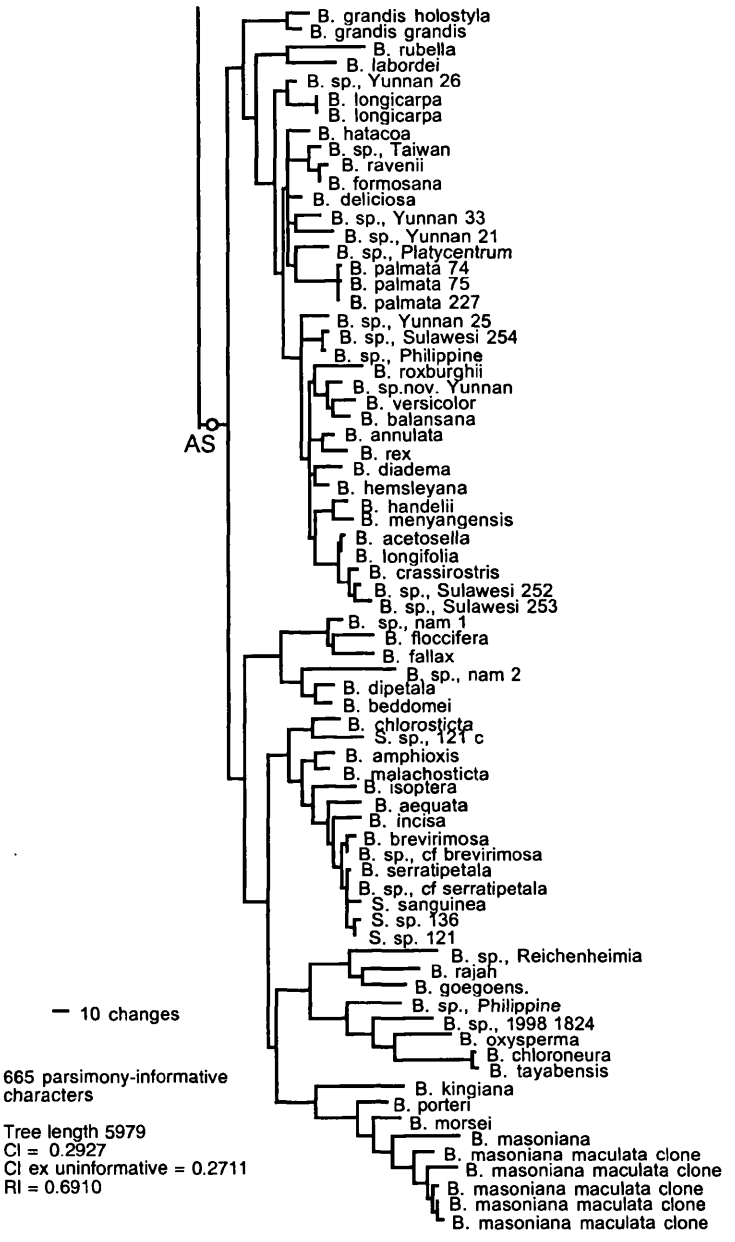

B. grandis holostyla
B. grandis grandis
B. rubella
B. labordei
B. sp., Yunnan 26
B. longicarpa 1
B. longicarpa 2
B. hatacoa
B. sp., Taiwan
B. ravenii
B. formosana
B. deliciosa
B. sp., Yunnan 33
B. sp., Yunnan 21
B. sp., Platycentrum
B. palmata 74
B. palmata 75
B. palmata 227
B. sp., Yunnan 25
B. roxburghii
B. diadema
B. hemsleyana
B. annulata
B. rex
B. sp., Sulawesi 254
B. sp. nov., Philippine
B. versicolor
B. balansana
B. sp. nov. Yunnan
B. handelii
B. menyangensis
B. acetosella
B. longifolia
B. crassirostris
B. sp., Sulawesi 252
B. sp., Sulawesi 253
B. sp., nam 1
B. floccifera
B. fallax
B. sp., nam 2
B. dipetala
B. beddomei
B. sp., Reichenheimia
B. rajah
B. goegoens.
B. sp., Philippine
B. sp., 1998 1824
B. oxysperma
B. chloroneura
B. tayabensis
B. kingiana
B. porteri
B. morsei
B. masoniana
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. isoptera
B. chlorosticta
S. sp., 121 clone
B. amphioxis
B. malachosticta
B. aequata
B. incisa
B. serratipetala
B. sp., cf serratipetala
S. sanguinea
B. brevirimosa
B. sp., cf brevirimosa
S. sp. 136
S. sp. 121
AS
Bootstrap support values over 50% above lines

Figure 7.6: Phylogram for unculled manual ITS alignment; one of 100 MPTs




B. grandis holostyla
B. grandis grandis
B. rubella
B. labordei
B. sp., Yunnan 26
B. longicarpa
B. longicarpa
B. hatacoa
B. sp., Taiwan
B. ravenii
B. formosana
B. deliciosa
B. sp., Yunnan 33
B. sp., Yunnan 21
B. sp., Platycentrum
B. palmata 74
B. palmata 75
B. palmata 227
B. sp., Yunnan 25
B. sp., Sulawesi 254
B. sp., Philippine
B. roxburghii
B. sp.nov. Yunnan
B. versicolor
B. balansana
AS
B. annulata
B. rex
B. diadema
B. hemsleyana
B. handelii
B. menyangensis
B. acetosella
B. longifolia
B. crassirostris
B. sp., Sulawesi 252
B. sp., Sulawesi 253
B. sp., nam 1
B. floccifera
B. fallax
B. sp., nam 2
B. dipetala
B. beddomei
B. chlorosticta
S. sp., 121 c
B. amphioxis
B. malachosticta
B. isoptera
B. aequata
B. incisa
B. brevirimosa
B. sp., cf brevirimosa
B. serratipetala
B. sp., cf serratipetala
S. sanguinea
S. sp. 136
S. sp. 121
B. sp., Reichenheimia
B. rajah
B. goegoens.
B. sp., Philippine
B. sp., 1998 1824
B. oxysperma
B. chloroneura
B. tayabensis
B. kingiana
B. porteri
B. morsei
B. masoniana
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
B. masoniana maculata clone
— 10 changes
665 parsimony-informative characters
Tree length 5979
CI = 0.2927
CI ex uninformative = 0.2711
RI = 0.6910

7.3.2.1. Topology: clades which occur in trees produced by the majority of alignments (see Figure 7.2, majority rule consensus tree) are as follows (with bootstrap values given, firstly from the elision tree, secondly from the culled manual ITS alignment):

1. **'*Rostrobegonia*'**: *B. annobonensis*, *B. engleri*, *B. johnstonii* (92%; 66%).
2. ***Augustia*:** *B. geranioides*, *B. dregei 'partita'*, *B. sonderana*, *B. dregei*, *B. dregei homonyma* (96%; <50%).
3. **the *B. masoniana* clones** (75%; 72%).
4. **Madagascar:** *B. ankaranensis*, *B. bogneri*, *B. madecassa*, *B. mananjensis*, *B. salaziensis*, *B. francoisii*, *B. nossibea* (100%; 100%).
5. **'*Loasibegonia*'**: *B. aspleniifolia*, *B. staudii*, *B. scapigera*, *B. duncan-thomasii*, *B. letouzeyi*, *B. dewildei*, *B. potamophila*, *B. prismatocarpa*, *B. quadrialata*, *B. scutifolia* (98%; 95%).
6. **'*Tetraphila*'**: *B. longipetiolata*, *B. poculifera*, *B. loranthoides* ssp *rhopalocarpa*, *B. capillipes*, *B. gabonensis*, *B. horticola*, *B. kisuluana*, *B. mannii*, *B. molleri*, *B. subscutata* (100%; 84%).
7. **American clade A (*Begonia* and *Ruizopavonia*):** Bolivian *B.* sp., *B. fissistyla*, *B. fuchsioides*, *B. holtonis*, *B. jamesoniana*, *B. meridensis*, *B. guaduensis*, *B.* sp. 'sychnanthera', *B. cubensis*, *B. obliqua*, *B. minor*, *B. odorata* (100%; 98%).
8. ***Begonia*:** *B. cubensis*, *B. obliqua*, *B. minor*, *B. odorata* (100%; 98%).
9. **'*Pritzelia*'**: *B. angularis*, *B. echinosepala*, *B. egregia*, *B. listada*, *B. lobata*, *B. luxurians* (and its clones), *B.* sp. 'macrocarpa' GL, *B. oxyphylla*, *B. rufosericae*, *B. acerifolia*, *B. convolvulacea*, *B. glabra*, *B.* sp. 224, *B.* sp 'macrocarpa' E, *B. ulmifolia*, *B. valida* (100%; 83%).
10. **American clade B pro parte:** *B. acerifolia*, *B. convolvulacea*, *B. glabra*, *B.* sp. 224, *B.* sp 'macrocarpa' E, *B. ulmifolia*, *B. valida* (100%; 99%).

Where section names are given in inverted commas, the vast majority of species within the clade belong to that section (Doorenbos, Sosef & de Wilde, 1998), but the underlined species belong to related sections.

The strict consensus trees for the individual alignments were checked to see which trees contained which of these clades. A few other features have also been considered:

11. *Hillebrandia* has appeared as sister to *Begonia* in most of the analyses in which it has been included, including the analyses of ITS which only use conserved regions of sequence, Swensen, Luthi & Rieseberg (1998), Wagstaff and Dawson (2000) and unpublished studies with *trnL* (Plana, 2000).
12. The placement of the outgroup on the trees has also been considered: in many cases the 'conventional' tree of [*Datisca* [*Hillebrandia* [*Begonia*]]] has not been

obtained and so trees have been checked to see whether or not *Datisca* has been resolved as sister to *Hillebrandia* or *Begonia* taxa. The elision tree gives 87% bootstrap support to a monophyletic *Begonia*, with *Hillebrandia* as sister. **13, 14.** Lastly, in some preliminary analyses of manually aligned ITS data, the African taxa have been paraphyletic including the American and Asian taxa as follows: [Africa[Africa[Asia][America]]]. Therefore trees were checked to see which indicated monophyly of American and/or Asian taxa.

This information is all summarised in Table 7.4.

Table 7.4: Topological features of the cladograms produced by different alignments

Clade	Al. 1	Al. 2	Al. 3	Al. 4	Al. 5	Al. 6	Al. 7	Al. 8	Al. 9	Al. 10	Al. 11	Al. 12	Al. 13	Al. 14	Al. 15	Al. 16	Elision	Manual Culled	Manual entire
1. 'Rostrobegonia'	+	+	+			+	+	+		+	+	+	+	+		+	+	+	+
2. Augustia	+		+		+	+	+	+	+	+				+	+	+	+	+	+
3. B. masoniana	+				+	+	+		+	+	+	+		+	+	+	+	+	+
4. Madagascar	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
5. 'Loasibegonia'	+	+	+		+	+	+	+		+	+	+	+	+		+	+	+	+
6. 'Tetraphila'	+		+		+	+	+	+	+	+	+	+	+	+		+	+	+	+
7. American clade A	+	+	+		+	+	+	+	+		+	+	+		+	+	+	+	+
8. Begonia	+		+	+	+	+	+		+	+	+	+	+		+	+	+	+	+
9. American clade B	+		+	+	+	+	+	+	+		+	+	+		+	+	+	+	+
10. American clade B pro parte	+		+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+
11. Hillebrandia sister to Begonia		+			+	+	+		+			+		+	+		+	+	+
12. Datisca sister to Hillebrandia OR African taxa	+	+		+	+	+	+	+	+			+		+	+		+	+	+
13. American taxa monophyletic					+				+	+	+	+	+					+	+
14. Asian taxa monophyletic										+		+						+	+

Alignments 1 and 7 contain the greatest number of the described clades, although in alignment 1 *Hillebrandia* does not appear as sister to *Begonia*.

The most consistently resolved clade is of the Madagascan species, which is recovered from 15 out of 16 alignments. In only two out of 16 alignments are the Asian taxa monophyletic. In several of the trees where the Asian taxa are not monophyletic, this is because they have been rendered paraphyletic by American taxa, not because taxa are scattered widely across the trees.

7.3.2.2. Tree Distance Measures: The highest possible value for the partition metric (PM) for these data is 2N - 6 = 2 x 177 - 6 = 348. Values for the 16 different Clustal alignments ranged from 212 (between the trees produced from alignments 1 and 10, and alignments 5 and 12) to 314 (between alignments 4 and 7). Values of D_1 were lowest (97) between alignments 10 and 11, and highest (140) between alignments 8 and 13. Between the six elision MPTs, PM ranged from two to six, and D_1, from one to three. Between the majority rule elision tree and each of the 16 majority rule alignment trees, PM ranged from 176 (alignment 10) to 272 (alignment 4) and D_1, from 85 (alignment 12) to 134 (alignment 4).

7.3.3 Compartmentalization

A. Methods: Well-supported clades were selected on the basis of strict consensus trees from:

1. the 26S data set of 38 taxa
2. the 26S and ITS combined data set of 38 taxa
3. the elision ITS data set of 177 sequences
4. the individual automated alignments
5. the culled ITS data set of 177 sequences
6. the complete ITS data set of 177 sequences

Clades were selected which were monophyletic in all trees except those produced from the automated alignments (which occasionally have widely misplaced taxa in clades which otherwise agree with the other data sets and data analyses) and the 26S alignment for taxa within the Asian - American clade (because some clades here are based on only one or two characters, with no bootstrap support, and differ from clades recovered by all other treatments). Not all clades have bootstrap support in all data sets. The clades which were isolated thus are:

1. **'*Loasibegonia*'**: *B. aspleniifolia*; *B. staudtii*; *B. scapigera*; *B. duncan-thomasii*; *B. letouzeyi*; *B. dewildei*; *B. prismatocarpa*; *B. scutifolia*; *B. potamophila*; *B. quadrialata* (total of 10 taxa).
2. **'*Tetraphila*'**: *B. loranthoides* ssp *rhopalocarpa*; *B. longipetiolata*; *B. poculifera*; *B. molleri*; *B. subscutata*; *B. gabonensis*; *B. capillipes*; *B. horticola*; *B. kisuluana*; *B. mannii* (total of 10 taxa).
3. **Madagascar**: *B. bogneri*; *B. salaziensis*; *B. madecassa*; *B. mananjensis*; *B. ankaranensis*; *B. nossibea*; *B. francoisii* (total of seven taxa).
4. ***Coelocentrum***: *B. porteri*; *B morsei*; *B. masoniana* and the *B. masoniana* var. *maculata* clones (total of eight sequences).
5. **'*Petermannia*'**: *B. cholorosticta*; *B. amphioxis*; *B. malachosticta*; *B. isoptera*; *B. aequata*; *B. incisa*; *B. serratipetala*; *B.* cf. *serratipetala*; *B. brevirimosa*; *B.* cf. *brevirimosa*; *Symbegonia sanguinea*; *S.* sp. 136; *S.* sp. 121 (two sequences) (total of 14 sequences).
6. **'*Platycentrum*'**: Yunnan sp. 21; Yunnan sp. 25; Yunnan sp. 26; Yunnan sp. 33; *B. versicolor*; *B. balansana*; *B.* sp. nov. 20; *B. longicarpa* (two individuals); *B. hatacoa*; *B.* sp., Taiwan; *B. ravenii*; *B. formosana*; *B. deliciosa*; *Platycentrum* sp. 215; *B. palmata* (three accessions); *B. hemsleyana*; *B. roxburghii*; *B. diadema*; *B. annulata*; *B. rex*; *B.* sp., Sulawesi 254; *B.* sp. nov.*:* Philippines; *B. handelii*; *B. menyangensis*; *B. acetosella*; *B. longifolia*; *B. crassirostris*; *B.* sp, Sulawesi 252; *B.* sp, Sulawesi 253 (total of 32 taxa).
7. **'*Begonia*'**: *B. fissistyla*; *B.* sp., Bolivia; *B. odorata*; *B. minor*; *B. cubensis*; *B. obliqua*; *B.* sp. 'sych'; *B. guaduensis*; *B. meridensis*; *B. holtonis*; *B. fuchsioides*; *B. jamesoniana* (total of 12 taxa).
8. **'*Pritzelia*'**: *B. ulmifolia*; *B.* sp. 224; *B. glabra*; *B. convolvulacea*; *B.* sp. 'macE'; *B.* sp. 'macG'; *B. acerifolia*; *B. valida*; *B. egregia*; *B. listada*; *B. echinosepala*; *B. rufosericae*; *B. angularis*; *B. lobata*; *B. oxyphylla* (two sequences); *B. luxurians* and its clones (total of 21 sequences).

For each clade, the region of the manual alignment which includes its taxa and outgroups from within other well-defined clades (not from the closest taxa in the larger phylogenies, because often the placement of these is uncertain over several trees, and rooting each compartment clade on them may bias subsequent reanalysis) were selected in MacClade 3.07 (Maddison & Maddison, 1992, 1997) and saved as separate files.

The compartments were removed from the total alignment and realigned manually in SeqPup 0.6f (Gilbert, 1995), before being exported to PAUP* 4.0 (Swofford, 2000). The form of parsimony analysis used depended on the size of the data sets. For smaller compartment data sets, exhaustive (less that 12 taxa) or branch and bound (12 to 14 taxa) searches were run; heuristics were used for compartment 6 (*Platycentrum*) and compartment 8 (*Pritzelia*) (1000 random addition replicates, TBR). Values for g1 (10,000 random trees) and PTP (outgroup excluded) were also calculated. Searches were run with and

without ambiguous sites excluded. Bootstrap support was calculated using 10,000 replicates of fast addition; Bremer support values were calculated using AutoDecay (Erikkson, 1998) (10 random addition replicates and TBR per constraint tree).

For seven of the eight data sets, analyses were run both for the complete alignments and for a culled subsection of the alignments (ambiguous positions excluded). For compartment 6, the *Platycentrum* data set, no culled analyses were run, as no positions appeared ambiguous.

B. Results: In all analyses except that for compartment 1, *Loasibegonia*, the topologies for both analyses (culled and unculled) were the same. In all analyses except that for the section *Begonia* data set (compartment 7) the topologies are also consistent with the cladogram produced by analysis of the culled 177 taxon ITS matrix. In the first analysis of compartment 5, the *Petermannia* data set, the cloned sequence of *Symbegonia* sp. 121 does not cluster with the consensus sequence for *Symbegonia* sp. 121. Removal of this cloned sequence (*Petermannia* analysis 2) increases tree support and decreases the number of MPTs. Although removing a taxon simply because one does not like the effect it has on an analysis is hard to justify, the placement of this one cloned sequence, far from the other *Symbegonia* species, may mean that it is a disfunctional paralogue and perhaps best excluded.

Furthermore, looking at the 5.8S sequence of the *Symbegonia* clone, there are eight point mutations (G to A, character 567; C to T, character 601; C to T, character 640; A to T, character 642; C to T, character 646; C to T, character 654; C to T, character 661 and C to T, character 670). These add weight to the hypothesis that the copy may be paralogous and can be safely excluded.

For a summary of tree statistics for these analyses, see Table 7.5. Individual trees are presented subsequently under the headings for each compartment. Where sectional placements are marked onto the trees, these are taken from Doorenbos, Sosef and de Wilde (1998).

Table 7.5: Tree statistics for compartment analyses

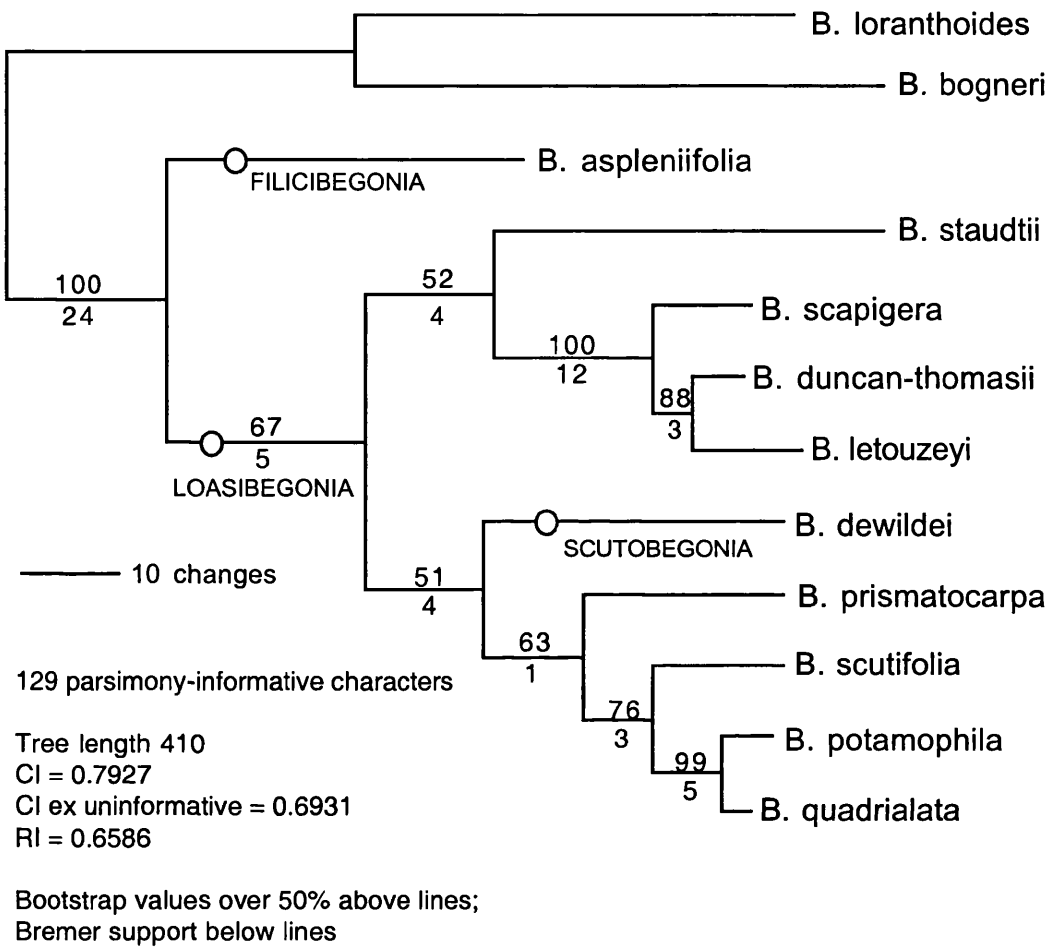
Comp.	Clade	No. taxa (+ OG)	Constant chars	Uninform. chars	Inform. chars	No. trees	MPT length	CI	CI ex uninf.	RI	g1	PTP
1	Loasibegonia	10 (12)	441	148	149	1	494	0.80	0.69	0.66	-1.036	0.001
1	Loasibegonia culled	10 (12)	399	112	129	1	410	0.79	0.69	0.66	-1.082	0.001
2	Tetraphila	10 (12)	513	161	83	1	371	0.84	0.67	0.64	-0.779	0.01
2	Tetraphila culled	10 (12)	444	110	73	1	282	0.82	0.66	0.65	-0.797	0.01
3	Madagascar	7 (9)	534	150	77	2	289	0.90	0.75	0.71	-1.678	0.001
3	Madagascar culled	7 (9)	485	138	74	1	270	0.90	0.75	0.72	-1.780	0.001
4	Coelocentrum	8 (10)	513	144	60	1	263	0.92	0.78	0.77	-1.254	0.01
4	Coelocentrum culled	8 (10)	478	110	46	1	195	0.92	0.79	0.79	-1.144	0.01
5	Petermannia	14 (16)	554	164	81	36	357	0.82	0.62	0.64	-0.748	0.001
5	Petermannia culled	14 (16)	524	148	75	36	328	0.81	0.61	0.64	-0.714	0.001
5	Petermannia 2	13 (15)	574	149	76	4	318	0.83	0.64	0.66	-0.814	0.001
5	Petermannia 2 culled	13 (15)	543	132	72	4	292	0.82	0.64	0.66	-0.799	0.001
6	Platycentrum	32 (34)	473	138	172	10	559	0.70	0.58	0.67	-1.291	0.001
7	Begonia	12 (15)	463	129	151	1	452	0.83	0.75	0.79	-1.179	0.002
7	Begonia culled	12 (15)	440	113	142	1	415	0.83	0.75	0.78	-1.076	0.002
8	Pritzelia	21 (24)	426	140	159	2	521	0.78	0.67	0.81	-0.852	0.001
8	Pritzelia culled	21 (24)	397	121	145	2	467	0.77	0.66	0.82	-0.776	0.001

7.3.3.1. Compartment 1: *Loasibegonia* (Africa)

The PTP probability is 0.001; g1 is -1.082. There were 259 characters excluded, leaving 399 constant, 112 uninformative and 129 informative characters. The furthest pairwise distances within *Loasibegonia / Scutobegonia* are 0.130 (*B. staudtii* to *B. dewildei*) and the closest are 0.013 (*B. potamophila* to *B. quadrialata*); the greatest distance between *Filicibegonia* and *Loasibegonia* is 0.151 (*B. aspleniifolia* to *B. staudtii*).

Analysis of the culled and unculled data sets both produced the same single MPT (Figure 7.7).

Figure 7.7: Phylogram of single MPT for culled '*Loasibegonia*' data set:



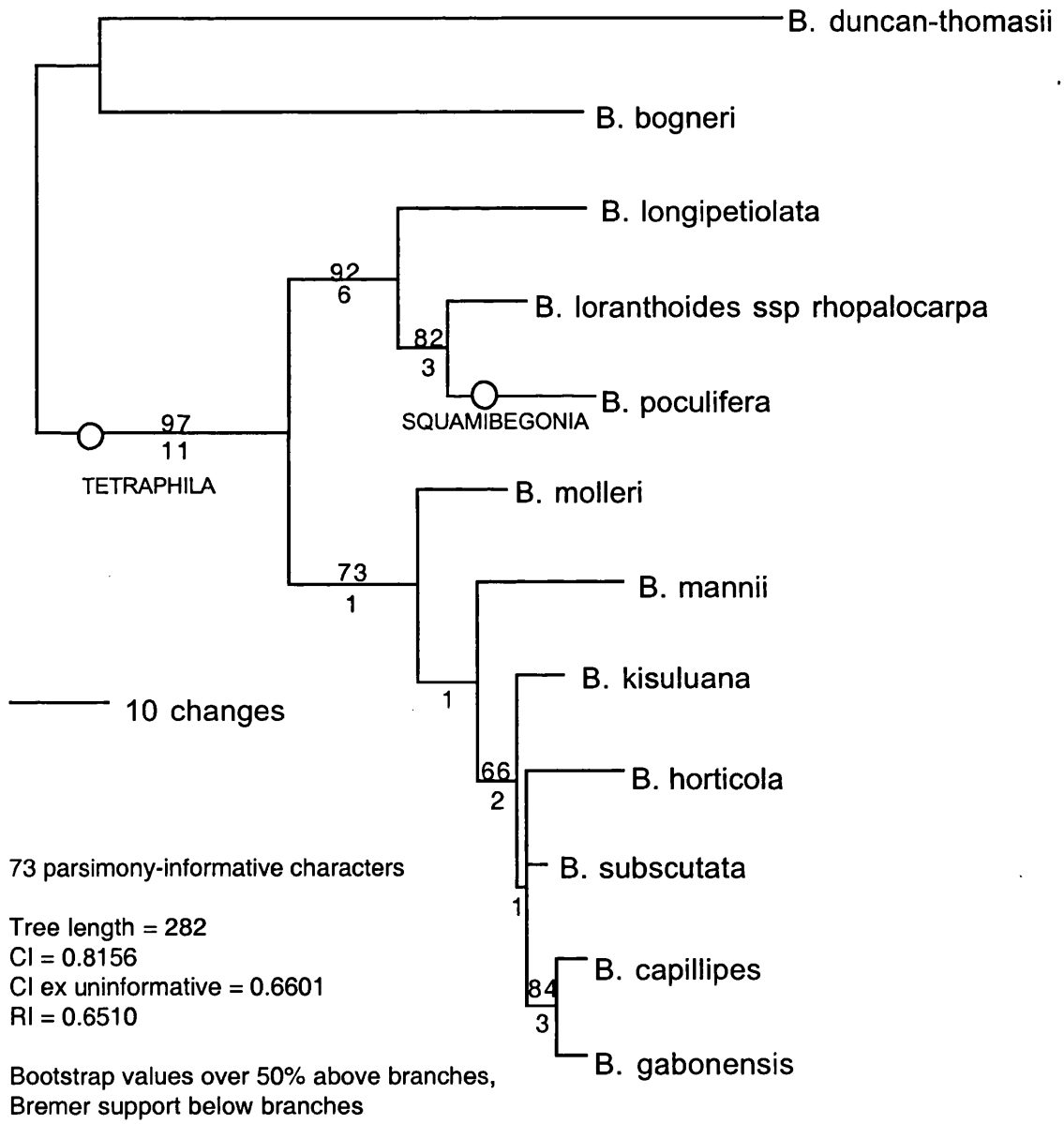
On the basis of this ITS phylogram (Figure 7.7), section *Filicibegonia* is sister to section *Loasibegonia*, and section *Loasibegonia* is paraphyletic without the inclusion of section *Scutobegonia*. The topology of the ingroup here is the same as the elision, unculled manual and culled manual 177-sequence analysis topologies for the same taxa (Figures 7.1, 7.3 and 7.5). (See Appendix 14.4 for a comparison of this tree topology with a morphological cladogram topology produced by Sosef, 1994.)

7.3.3.2. Compartment 2: *Tetraphila* (Africa)

The PTP probability is 0.010; g1 is -0.776. There were 249 characters excluded, leaving 444 constant, 110 uninformative and 73 informative characters. The maximum pairwise distance from ingroup to outgroup is 0.231 (*B. longipetiolata* to *B. duncan-thomasii*). Within the ingroup, the most divergence is 0.103 (*B. manii* to *B. longipetiolata*) and the least is 0.011 (*B. gabonensis* to *B. capillipes*).

Because the ingroup and outgroup were quite divergent, analyses were run using each outgroup species separately. Both outgroups produced the same topology (one MPT). Analyses were also run including all the matrix and excluding variable positions. Both produced the same topology (one MPT) (Figure 7.8). The tree from the culled matrix is presented here, as a more conservative estimate.

Figure 7.8: Phylogram of single MPT for *Tetraphila* matrix



On the basis of this ITS tree (Figure 7.8), section *Tetraphila* resolves into two clades, one of which includes section *Squamibegonia*. The topology of the ingroup has several differences to the elision and unculled manual 177-

sequence analysis topologies (Figures 7.1 and 7.5), but is consistent with the topology from the culled manual 177-sequence analysis (Figure 7.3), although this topology is more resolved.

7.3.3.3. Compartment 3: Madagascar

The PTP probability is 0.001; g1 (evaluated during the exhaustive search) is -1.753. The maximum pairwise distance from outgroup to ingroup is 0.210 (*B. duncan-thomasii* to *B. nossibea*); within the ingroup, the maximum distance is 0.071 (*B. ankaranensis* to *B. madecassa*) and the minimum is 0.011 (*B. nossibea* to *B. francoisii*).

Using each outgroup separately and using the culled and unculled matrices both found the same two MPTs (Figures 7.9 and 7.10).

Figure 7.9: First phylogram (of two MPTs) for Madagascan matrix

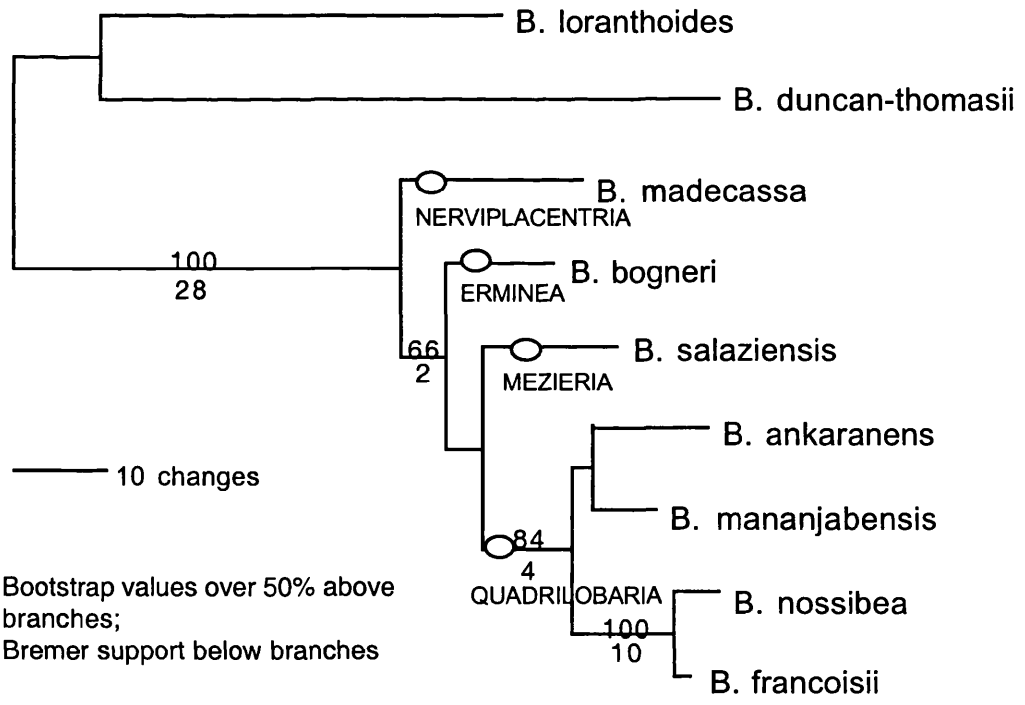
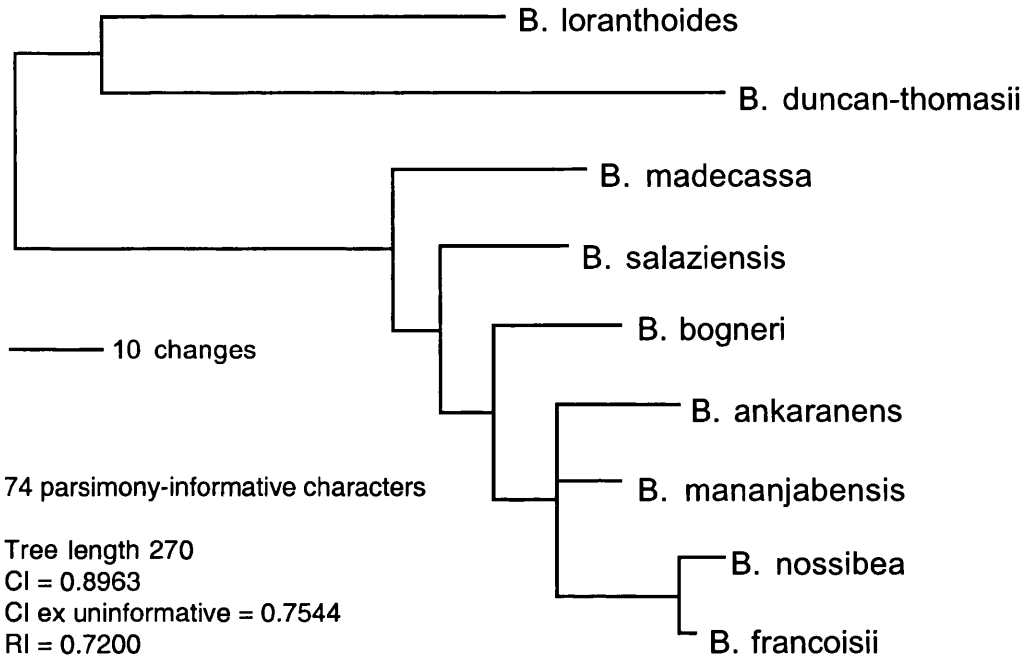


Figure 7.10: Second Phylogram (of two MPTs) for Madagascan matrix



The trees differ by the position of *B. bogneri* and *B. salaziensis*, and the *Quadrilobaria* clade is not fully resolved in the second tree. These ITS trees suggest that section *Quadrilobaria* is monophyletic. Sampling does not allow consideration of the monophyly of sections *Nerviplacentaria* or *Erminea*. The monophyly of section *Mezieria* is not considered here, as it has been assumed to be paraphyletic based on prior analyses (Figures 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 7.1, 7.3 and 7.5).

The first phylogram is congruent with the culled manual 177-sequence analysis (Figure 7.3), but is more resolved; it differs from the unculled manual 177-sequence analysis (Figure 7.5) in that the unculled alignment gives *B. salaziensis* as basal. It also differs from the elision 177-sequence analysis (Figure 7.1).

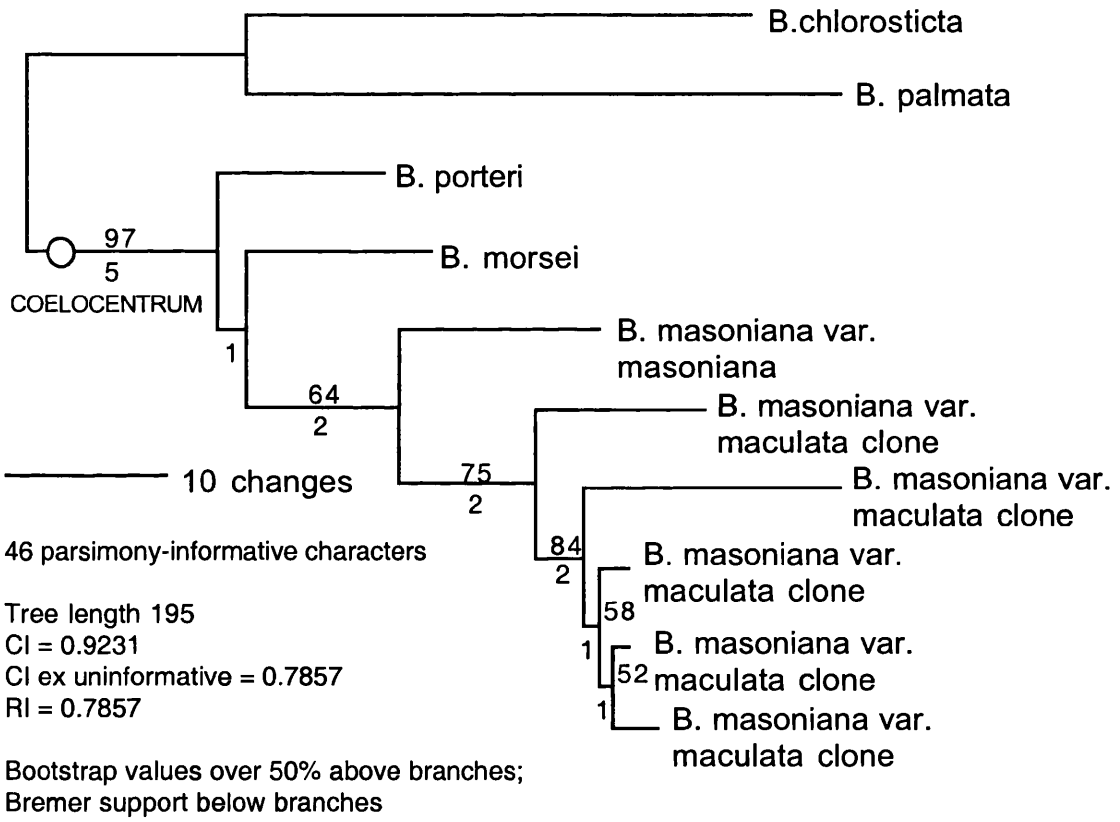
7.3.3.4. Compartment 4: *Coelocentrum* (Asia)

The PTP probability is 0.010; g1 (estimated from an exhaustive search) is -1.073. There were 272 characters excluded, leaving 478 constant, 110 uninformative and 46 informative characters. Pairwise distances range from 0.007 (between two clones of *B. masoniana* var. *maculata*) to 0.113 (*B. masoniana* var. *masoniana* to *B. masoniana* var. *maculata*). The largest

distance between the other species in section *Coelocentrum* and *B. masoniana* is 0.106 (*B. porteri* to *B. masoniana*) and between *B. porteri* and *B. morsei*, the distance is 0.072. The maximum distance from the outgroup to the ingroup is 0.171 (*B. masoniana* to *B. palmata*).

Analysis of the unculled and of the culled data sets found a single MPT (Figure 7.11).

Figure 7.11: Single MPT for *Coelocentrum* matrix



Section *Coelocentrum* appears monophyletic. Within the section, all sequences from *B. masoniana* resolve as monophyletic. The topology found is the same as that from the unculled 177-sequence analysis (Figure 7.5); it is congruent with, but less resolved that, the topology from the culled 177-sequence analysis (Figure 7.3), and differs from the topology from the elision 177-sequence analysis (Figure 7.1).

When a locus which has been homogenised by concerted evolution is compared across species, "all the paralogues within a species appear as each others' closest relatives in a gene tree" (Doyle & Gaut, 2000, p. 3). If the

different ITS sequences from clones of *B. masoniana* var. *maculata* represent paralogous copies, their monophyly in this ITS gene tree is evidence for concerted evolution in this species. However, it is also possible that the different copies represent recent allelic variation rather than paralogues.

7.3.3.5. Compartment 5: *Petermannia* (Asia)

The PTP probability is 0.001; g1 is -0.799. There were 181 characters excluded; of the remaining characters, 543 were constant, 132 were parsimony uninformative and 72 were parsimony informative. The minimum pairwise distance within *Petermannia* is 0.006, between *B. brevirimosa* and *B.* cf. *brevirimosa*. The maximum distance is 0.081, between *B. chlorosticta* and *B. aequata*. Between the ingroup and the outgroup, the maximum distance is 0.130 (*B. masoniana* to *B. aequata*).

Four equally parsimonious trees were found. The strict consensus is shown in Figure 7.12, and one of the MPTs in Figure 7.13:

Figure 7.12: Strict consensus of four MPTs, *Petermannia* matrix

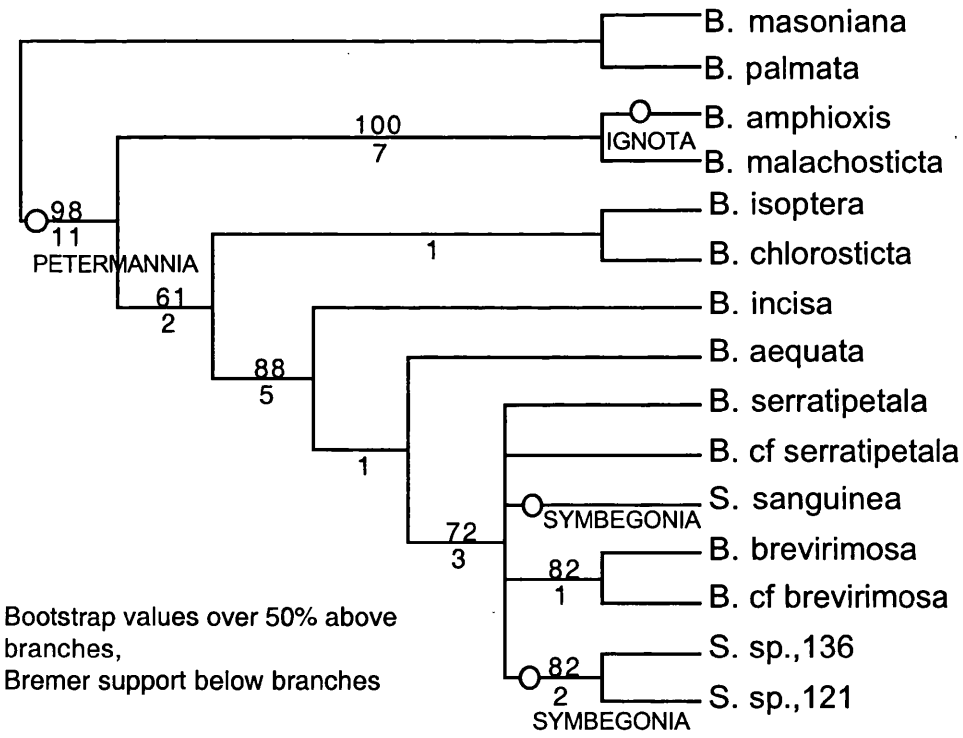
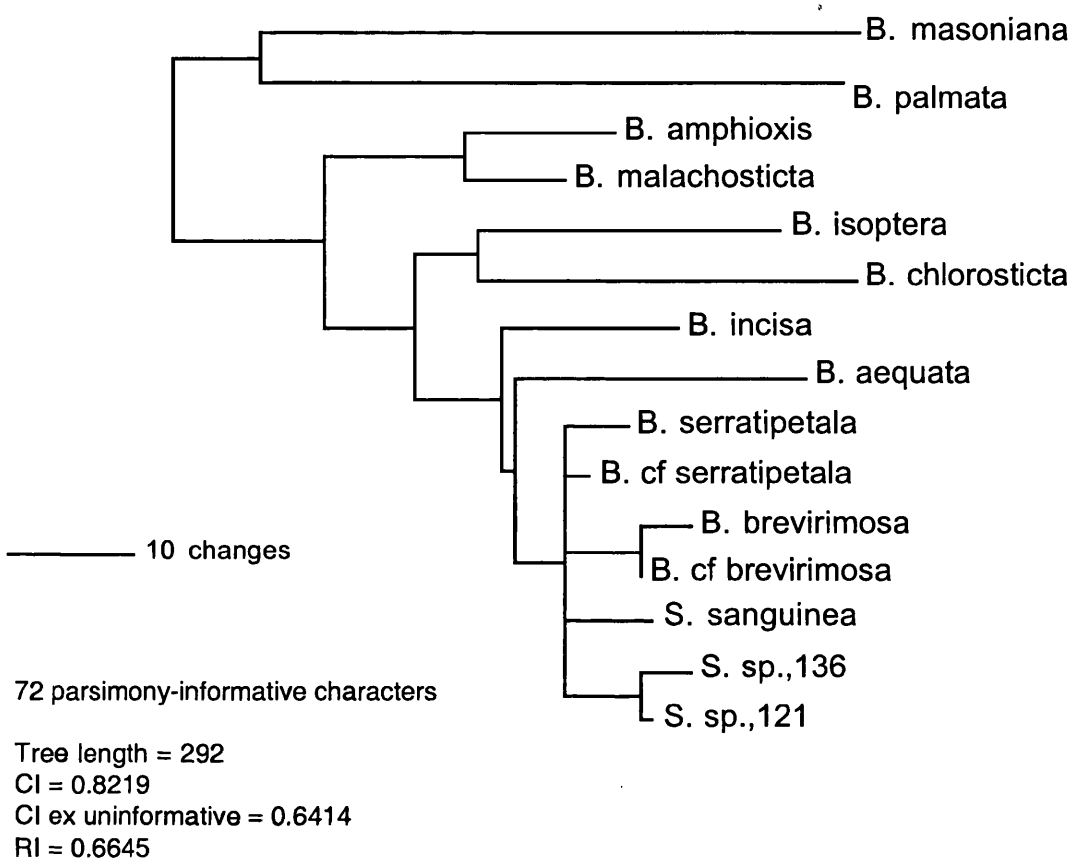


Figure 7.13: Phylogram for *Petermannia* matrix, one of four MPTs



On the basis of this ITS data, *Petermannia* is clearly paraphyletic without the inclusion of the genus *Symbegonia*.

The topology of the ingroup is consistent with, but more resolved than, those from the culled and unculled 177-sequence analyses (Figures 7.3, 7.5); it differs from the elision 177-sequence alignment topology (Figure 7.1).

7.3.3.6. Compartment 6: *Platycentrum* (Asia)

The PTP probability is 0.001; g1 is -1.291. The data set comprises 473 constant, 138 uninformative and 172 informative characters. There are few indels in this data set and so the alignment is not ambiguous; consequently no data are excluded. Pairwise distances between species ranged from 0.001 (*B. longifolia* to *B. acetosella*) to 0.095 (*B. roxburghii* to *B. palmata* 74).

An heuristic search was run with 1000 random additions, TBR. 10 MPTs were found (see Figures 7.14, 7.15).

Figure 7.14: Strict consensus of 10 MPTs for *Platycentrum* matrix

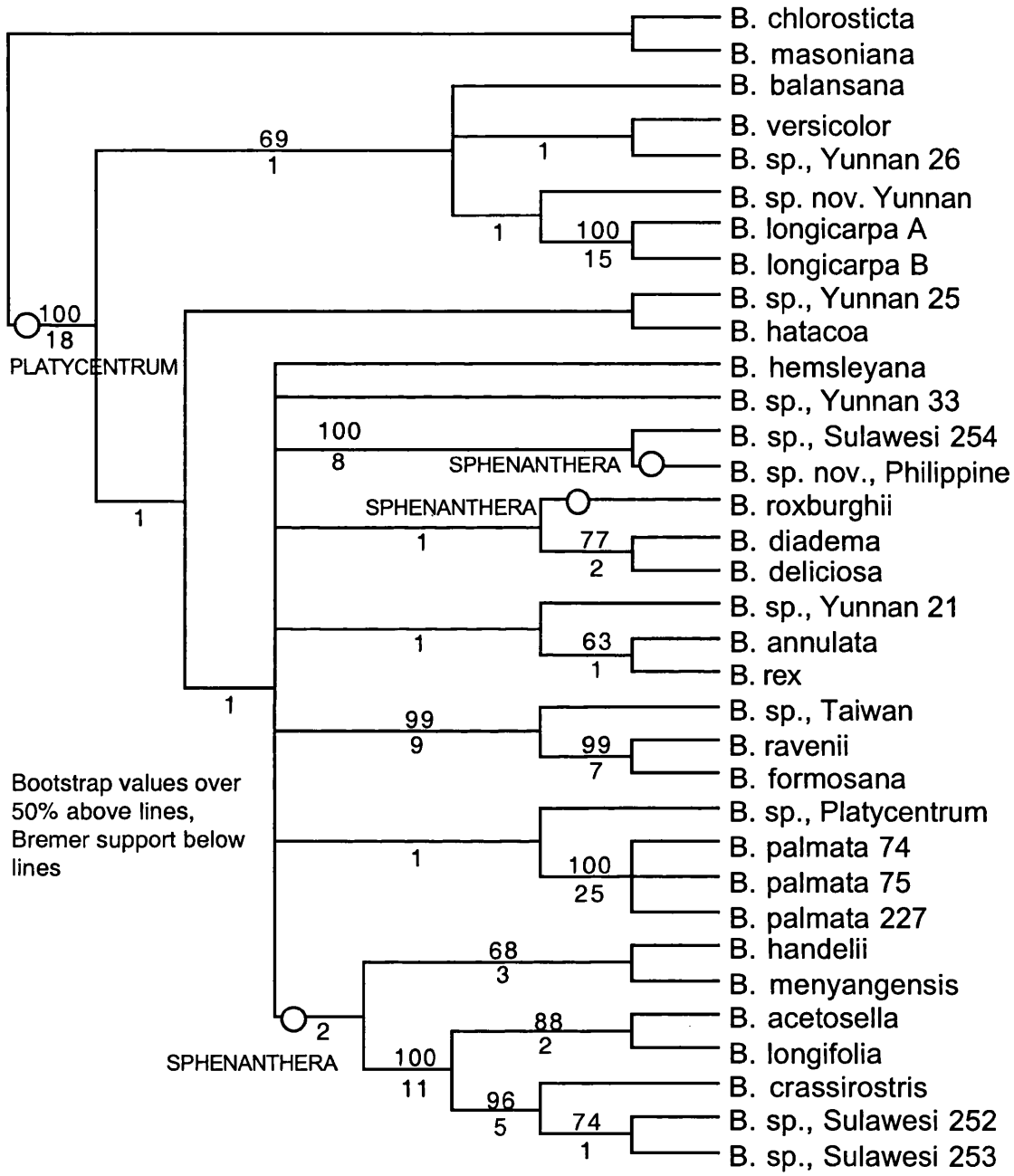
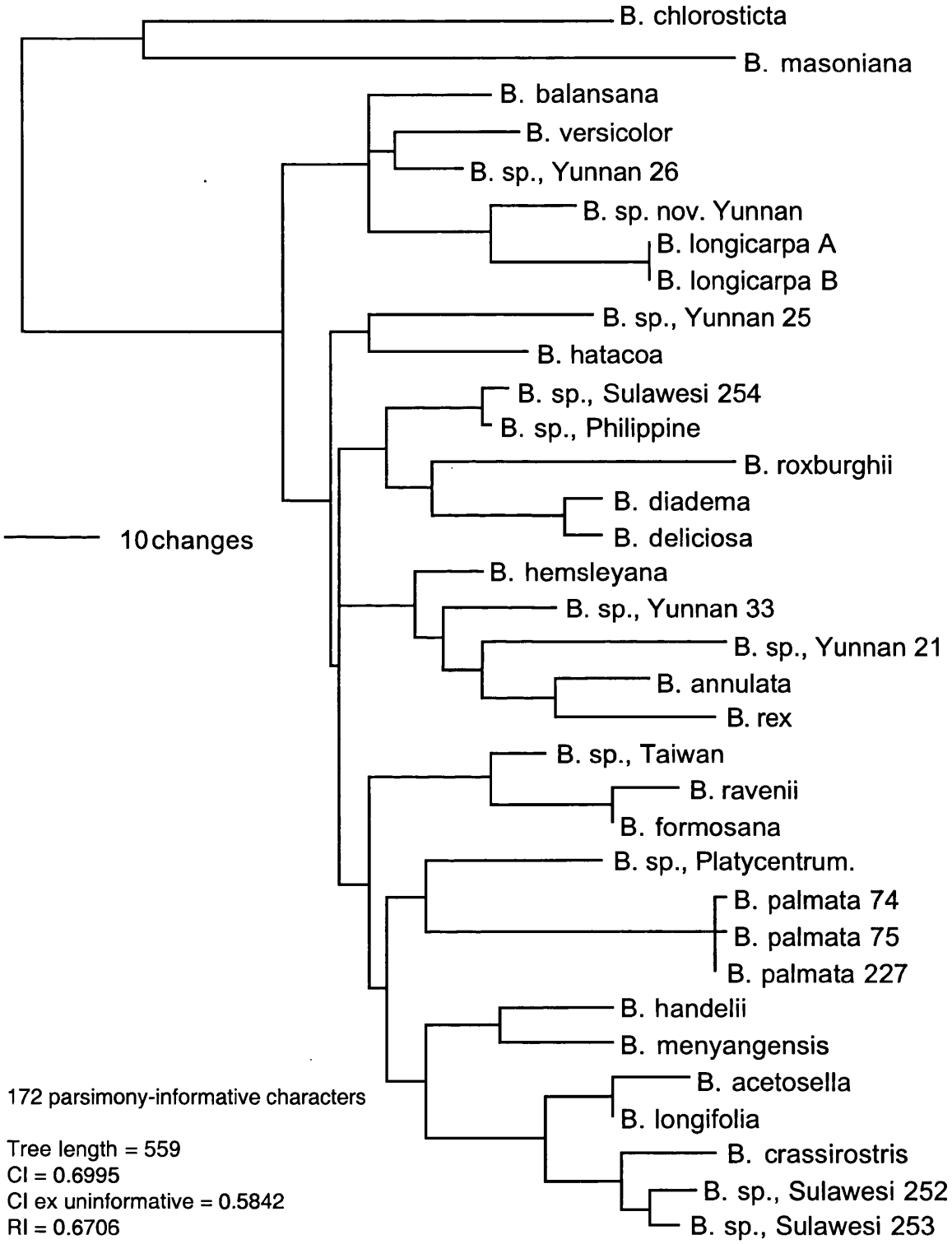


Figure 7.15: Phylogram for *Platycentrum* matrix, one of 10 MPTs



The phylogram (Figure 7.15) shows internal branches short relative to the terminal branches. Section *Platycentrum*, on the basis of ITS data, is paraphyletic without the inclusion of section *Sphenanthera*. The position of section *Sphenanthera* within section *Platycentrum* is not resolved here, with

Sphenanthera appearing polyphyletic in several of the MPTs.

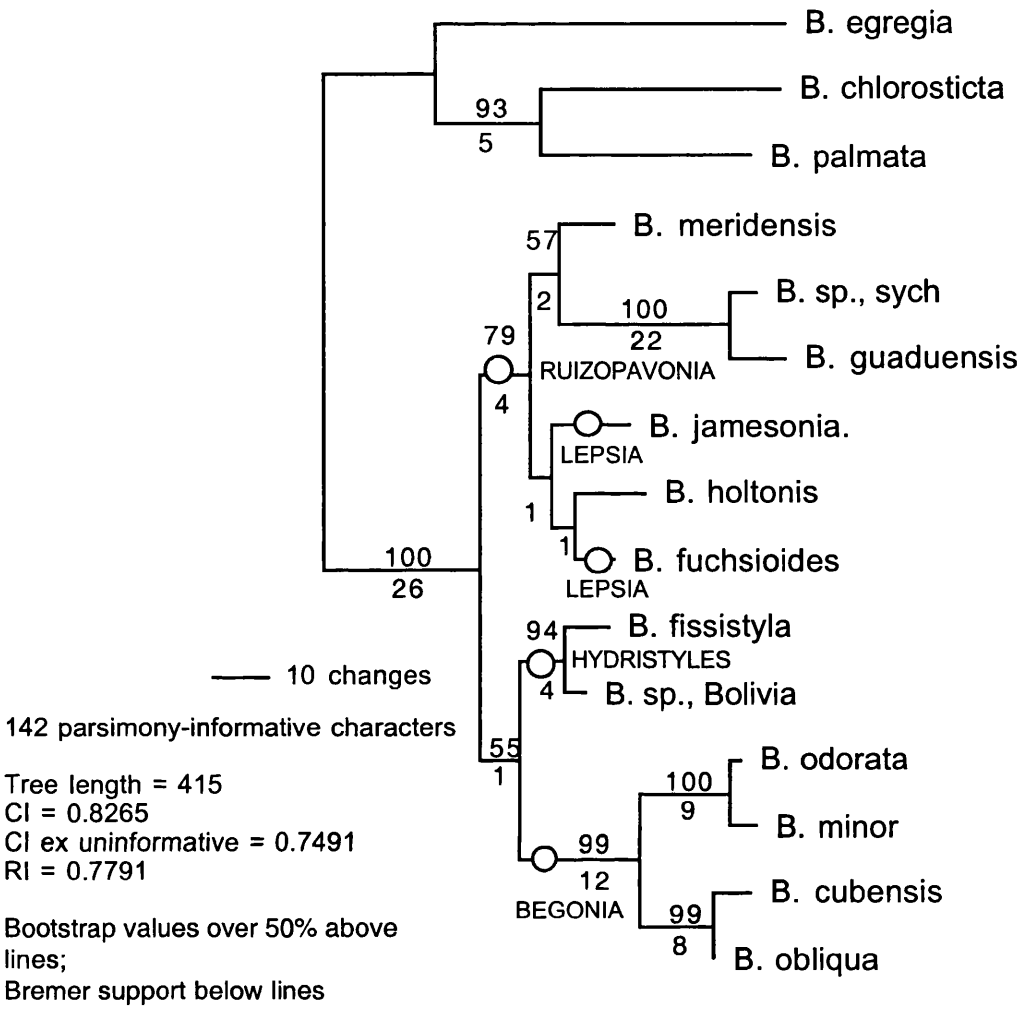
The topology of the ingroup is inconsistent with those from the elision and unculled manual 177-sequence analysis topologies (Figures 7.1, 7.5). It is largely congruent with, but more resolved than, the topology from the culled manual 177-sequence analysis (Figure 7.3); it differs by the placements of *B.* sp., Yunnan 21 and *B.* sp., Yunnan 33.

7.3.3.7. Compartment 7: '*Begonia*' (America)

The PTP probability for this matrix is 0.002; g1 is -1.087. There were 179 characters excluded, leaving 440 constant, 113 uninformative and 142 informative characters. Pairwise distances between taxa range from 0.011 (*B. obliqua* to *B. cubensis*) to 0.128 (*B. minor* to *B. guaduensis*). The widest distance to the outgroup is 0.258 (*B. guaduensis* to *B. egregia*).

Both the culled data set and the unculled data set found the same topology (Figure 7.16. The branch-lengths shown here are from the culled data set.

Figure 7.16: Single MPT for '*Begonia*' matrix



This phylogram resolves sections *Hydristyles* and *Begonia* as monophyletic, while sections *Ruizopavonia* and *Lepsia* are not well differentiated.

The topology of the ingroup here differs from the topologies produced by the elision, culled and unculled manual 177-sequence analyses (Figures 7.1, 7.3, 7.5). However, the inconsistency is not due to the placement of a few taxa, but to the rooting of the trees. Given that the wider analyses include closer relatives of the ingroup taxa here, it is likely that they provide a more accurate representation of evolution within this clade.

7.3.3.8. Compartment 8: '*Pritzelia*' (America)

The PTP probability is 0.001; g1 is -0.776. There were 201 characters excluded, leaving 397 constant, 121 uninformative and 145 informative characters. Pairwise distances range from 0.000 between two of the *B.*

luxurians clones and between *B. valida* and *B. acerifolia*, to 0.202, between *B. egregia* and *B.* sp. 224.

The same two MPTs were found from analysis of the culled and unculled data matrices (see Figures 7.17, 7.18). Data given here are from the analysis of the culled data set.

Figure 7.17: Strict consensus of two MPTs for *Pritzelia* matrix

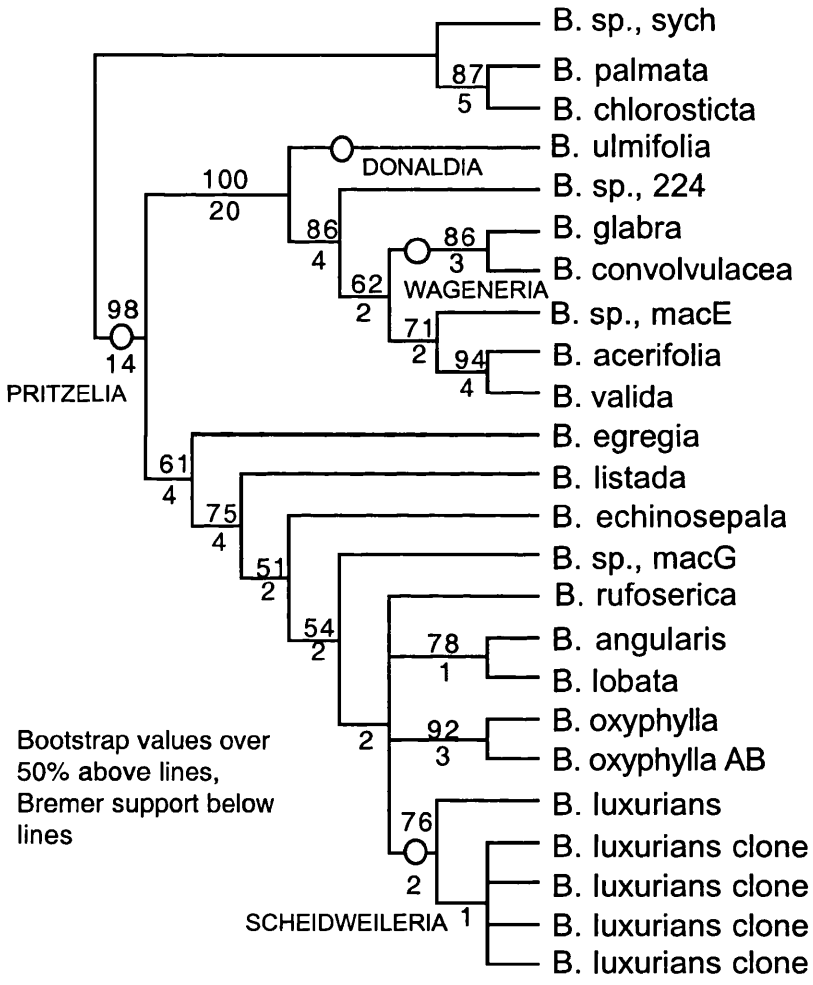
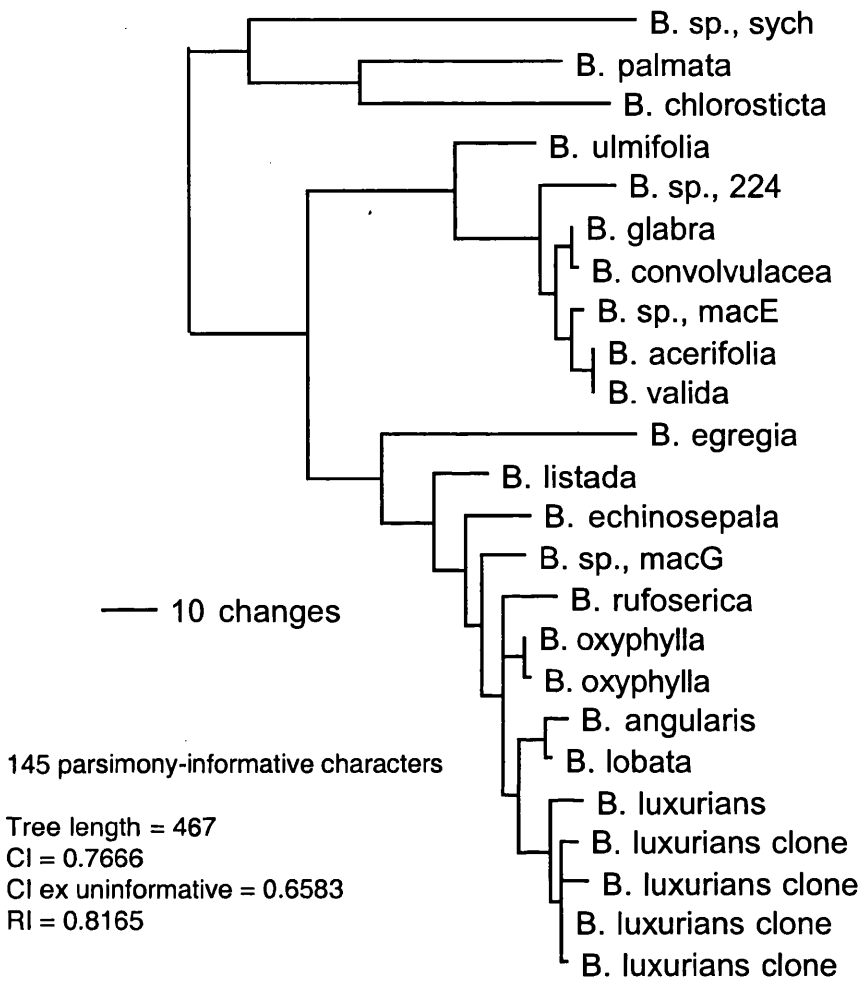


Figure 7.18: Phylogram for *Pritzelia* matrix, one of two MPTs



The two MPTs differ in the placement of *B. oxyphylla*, and the *B. angularis*/*B. lobata* clades. Section *Pritzelia* is paraphyletic in the basis of this ITS phylogeny, including the sections *Donaldia*, *Wageneria* and *Scheidweilaria*.

The topology of the ingroup is congruent with, but more resolved than, the topologies of the same taxa in the culled and unculled manual 177-sequence analyses (Figures 7.3, 7.5); it differs from the topology produced from the elision 177-sequence analysis (Figure 7.1).

As with the clones of *B. masoniana* in the analysis of compartment 4 (*Coelocentrum*), the clones of *B. luxurians* are monophyletic, indicating concerted evolution (Doyle & Gaut, 2000) (if they are paralogs).

Comparisons with previous topologies - summary: The compartment analyses have the benefit of more alignable data sets than the 177-taxon analyses, as they represent clades from within those wider analyses. Therefore estimates of character homology, and therefore the resulting topologies, are more reliable. Furthermore, less data needs to be excluded due to uncertainly. Thus we can use these topologies as a guide to how well the larger analyses performed. Table 7.6 summarises these comparisons. From this it can be seen that the method which was least reliable was elision, while that which performed best was the culled manual alignment.

Table 7.6: Summary of comparisons between compartment analysis topology and 177-sequence analysis topologies for different alignments

COMPARTMENT	ELISION	UNCULLED	CULLED
1. Loasibegonia	same	same	same
2. Tetraphila	differs	differs	congruent, more resolved
3. Madagascar	differs	differs	congruent, more resolved
4. Coelocentrum	differs	same	congruent, more resolved
5. Petermannia	differs	congruent, more resolved	congruent, more resolved
6. Platycentrum	differs	differs	differs (2 taxa interchange)
7. Begonia	differs	differs	differs - root
8. Pritzelia	differs	congruent, more resolved	congruent, more resolved

7.3.3.9 The remaining taxa

a. Introduction

Yeates (1995) points out that the character states which best represent a supra-specific taxon (by keeping it at the same position in a cladogram that the clade it was derived from took) are those of its common ancestor. If there is no homoplasy in a data matrix, replacing a monophyletic group by a terminal taxon using either the exemplar method or an hypothetical ancestor will not alter the inferred relationships (Bininda-Edmonds, Bryant & Russell, 1998). However, if some members of a clade have homoplasies with taxa outside the clade, using them as exemplar taxa can lead to incorrect phylogenetic reconstructions (through 'branch attraction' problems). Using an hypothetical ancestor reduces this problem with homoplasy because apomorphies of some clade members (which can be homoplasies with other clades) are ignored (Bininda-Edmonds, Bryant & Russell, 1998).

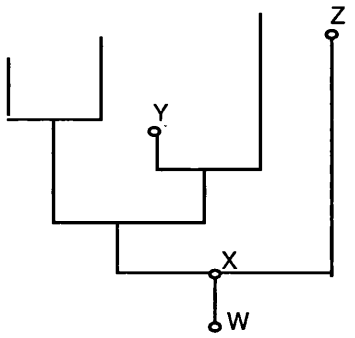
Three different procedures for obtaining compartmentalised trees were initially considered:

1. A constraint tree was constructed, including the topology of each of the compartment analyses, with all other taxa in a basal polytomy; this was used to direct phylogenetic analyses of the polytomous taxa. However, this did not reduce analysis times because the search algorithm in PAUP* 4.0 (Swofford, 2000) still considered trees based on any topology; PAUP then rejected trees which did not fit the constraint. This offers no perceivable advantage in terms of data analysis time.

2. An alternate method was tried, wherein the states on the branch leading to each compartment were reconstructed (using the 'describe trees' command in PAUP* 4.0 (Swofford, 2000), ACCTRAN) and added to the data matrix as hypothetical ancestors. Hovever, subsequent reanalysis of the compartment this was tested on ('*Platycentrum*'), including the hypothetisised ancestor as a taxon, did not place the ancestor in a basal position, and also grossly inflated the number of equally parsimonious trees for the data set. Furthermore, the ancestral states could have been affected not only by whether ACCTRAN or DELTRAN were used but also by the topology of the tree chosen (where there were more than one MPT for a data set) and by the choice of ancestor. Given these concerns, and coupled with the fact that an hypothesised ancestor is neither real nor testable, this method was then rejected.

3. Exemplar taxa can be used instead of hypothetical ancestors. Using exemplar taxa as place-holders also has disadvantages; the branch lengths may be far longer than they would be if an hypothetical ancestor was used. For example, in Figure 7.19, the distance to the hypothetical ancestor is only W to X, while that to the nearest exemplar taxon is W to Y. Thus long-branch attraction is more likely to cause problems in analyses which use exemplar taxa. A further consideration is whether it may be better to use taxon Z, which occupies a basal position relative to taxon Y, or taxon Y, which is on a shorter branch than taxon Z.

Figure 7.19: Choosing exemplar taxa



b. Material and methods: It was decided to select the exemplar taxa from each of the compartments, on the basis of a combination of factors - firstly, some taxa have missing data and were immediately rejected. Secondly, taxa on shorter branches were selected. Where there were more than one taxon with comparable branch lengths, the more basal taxon was chosen. Finally, some taxa cause more alignment difficulties than others, and were less liable to be chosen.

Further to excluding non-exemplar taxa from the compartment analyses for the reanalysis of the ITS matrix, multiple accessions or sequences were removed. The remaining data set contained 66 taxa. The matrix was imported into SeqPup 0.6f (Gilbert, 1995) and manually realigned. Variable positions and highly 'gappy' sites were then culled.

Maximum parsimony analysis was run using 1000 random addition replicates, TBR, saving no more than 10 equally parsimonious trees for each replicate (MaxTrees set at 5000). The saved trees were then used as the starting trees in a second round of searching, with TBR and swapping to completion, saving all most parsimonious trees. Bremer support was calculated using AutoDecay (Erikkson, 1998) (10 random addition replicates per constraint tree, TBR); Bootstrapping was performed using the 'fast addition' option in PAUP* 4.0 (Swofford, 2000) with 10,000 replicates; PTP (1000 replicates, simple addition, TBR, outgroup excluded) and g1 (10,000 random trees) were also calculated in PAUP* 4.0 (Swofford, 2000).

The strict consensus tree of the MPTs was saved and used as a topological constraint for a further round of analyses, 10,000 random addition sequences, TBR, keeping trees not compatible with the constraint tree, in order to test whether there were any equally parsimonious topologies which differed from

the strict consensus tree.

Taxa which were not resolved in the strict consensus of MPTs were identified, and the PTP values between some of these unresolved taxa were calculated using a branch and bound algorithm.

c. Results and Discussion: There are 555 characters excluded; of the 564 included characters, 133 are constant, 115 are uninformative and 316 are parsimony-informative across the entire data set of 66 taxa. The g1 value for the matrix is -0.816; PTP (excluding the outgroup and *Hillebrandia*) is 0.001.

There were 554 MPTs found, length 2186, which were used to construct the topological constraint tree. Rerunning the analysis with the constraint tree in place found 8525 trees of length 2187 (i.e. one step longer than the MPTs). The strict consensus of these less parsimonious trees had very little resolution. Twenty-one clades have over 50% bootstrap support. The consistency index is 0.37 (0.33 excluding uninformative characters); the retention index is 0.45.

PTP values for partitions of the matrix (unrooted, so no outgroup excluded; branch and bound search; 1000 replicates) are as follows:

1. *B. iucunda*, *B. molleri*, *B. madecassa*, *B. meyeri-johannis* (unresolved on parsimony tree): 1.000 (insignificant).
2. *B. balansana*, *B.* sp., *Reichenheimia*, *B. rubella*, *B. labordei*, *B. sutherlandii*, *B. geranioides* (unresolved): 0.704 (insignificant).
3. *B. beddomei*, *B. sonderana*, *B. oxysperma*, *B. theimei* (unresolved): 0.059 (insignificant).
4. *B. floccifera*, *B.* sp. 'nam', *B. dipetala*, *B. beddomei* (resolved): 0.031 (significant).
5. *B. theimei*, *B. meridensis*, *B. incarnata*, *B. olbia*, *B. gracilis*, *B. maynensis*, *B. sericoneura*, *B. peltata*, *B. manicata* (unresolved): 0.001 (significant).
6. *B. oxysperma*, *B.* sp., Philippine, *B.* sp. 1998 1824, *B. chloroneura*, *B. tayabensis* (resolved): 0.001 (significant).
7. *B. violifolia*, *B. imperialis*, *B. edmondoi*, *B. lubbersii*, *B.* sp. 'guttata' (resolved): 0.001 (significant).

This indicates that there is significant character covariance (taken as indicative of cladistic structure) in at least one part of this data set which is not resolved in the strict consensus of MPTs (partition 5); however, other unresolved taxa show no character covariance and so, as there is no cladistic structure to the data, any attempts to obtain further resolution between them would be superfluous (partitions 1, 2 and 3). Partition 4 was resolved in the strict consensus tree;

there is however only a relatively low level of character covariance between the taxa and so the reconstructed topology must have relatively low confidence. More reassuringly, partitions 6 and 7, which are resolved in the strict consensus trees, have significant character covariance.

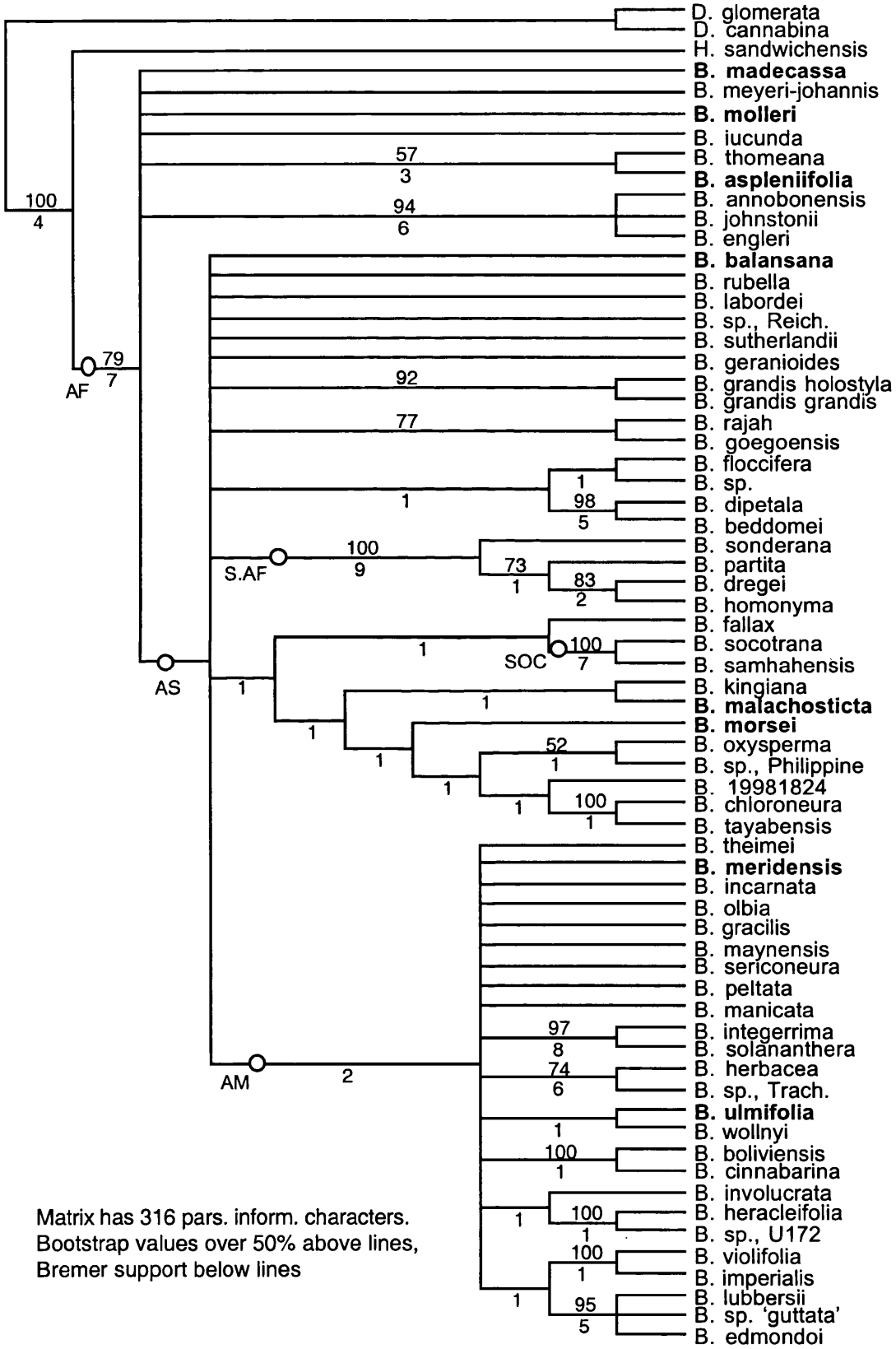
In order to compare the amount of information provided by this analysis (which only included exemplar taxa from well-structured clades) with the total analysis, the taxa which were not included in this analysis were pruned from the strict consensus of 10,000 MPTs for the culled manual ITS alignment (Figure 7.3), and the topologies were compared. Figure 7.20 is the strict consensus tree for the exemplar-included (compartment-removed) ITS data set. Figure 7.21 is the pruned strict consensus for the culled ITS analysis. Both trees have bootstrap and Bremer support values on the branches (obviously, for the pruned tree, these have been taken from the complete 177-sequence analysis). Taxa which are highlighted in bold are those which have been selected as exemplars. There is more resolution in the pruned tree. Both support monophyly for America, but only the pruned tree supports monophyly for Asia (if Socotra is included).

Tree comparison statistics for the two trees are $D_t = 47$, PM = 40 (31.7% of maximum possible PM value 126). The strict consensus tree for the compartment-removed data set has 33 resolved nodes, while the pruned consensus tree for the 177-sequence analysis has 56 resolved nodes.

Some clades are shared by both trees, but there is also some conflict, e.g. in the position of *B. morsei*. Many of the exemplar taxa are in more resolved positions on the pruned tree, suggesting that presence of other related taxa in the analysis affected their placement.

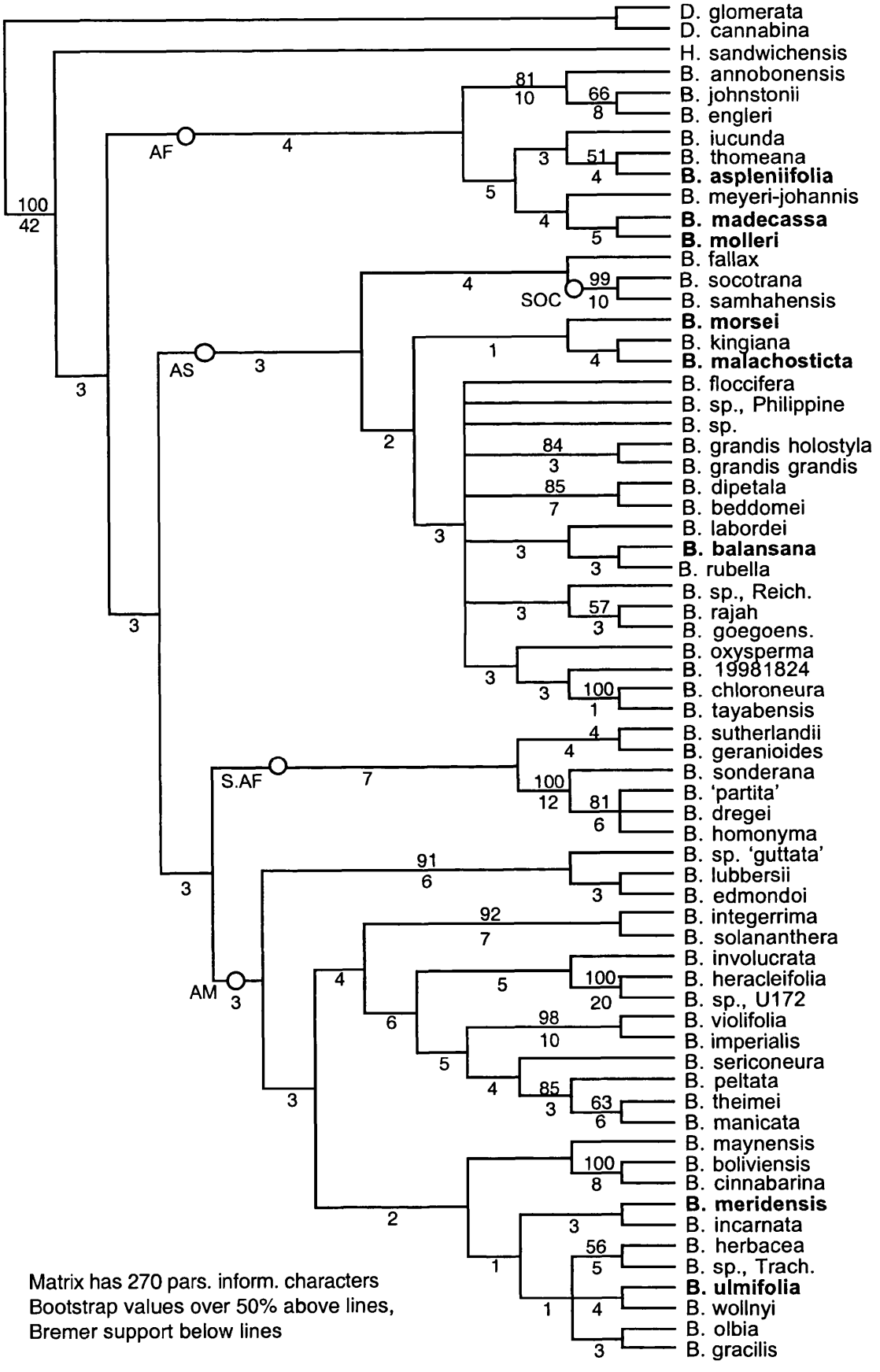
It seems therefore that the best resolution is produced by the 177-sequence analysis, and compartmentalization has little to offer in analyses of this data set. Tree support measures are not greatly increased in the reduced-taxon analysis. In fact, given that most of the compartment analyses are congruent with the culled 177-sequence analysis topology, and that the culled 177-sequence analysis offers higher resolution across the spine of the tree than analysis of a smaller data set, it seems that the best estimate of *Begonia* phylogeny is the culled ITS tree.

Figure 7.20: Strict consensus of 554 MPTs, compartment-removed ITS data set



Matrix has 316 pars. inform. characters.
Bootstrap values over 50% above lines,
Bremer support below lines

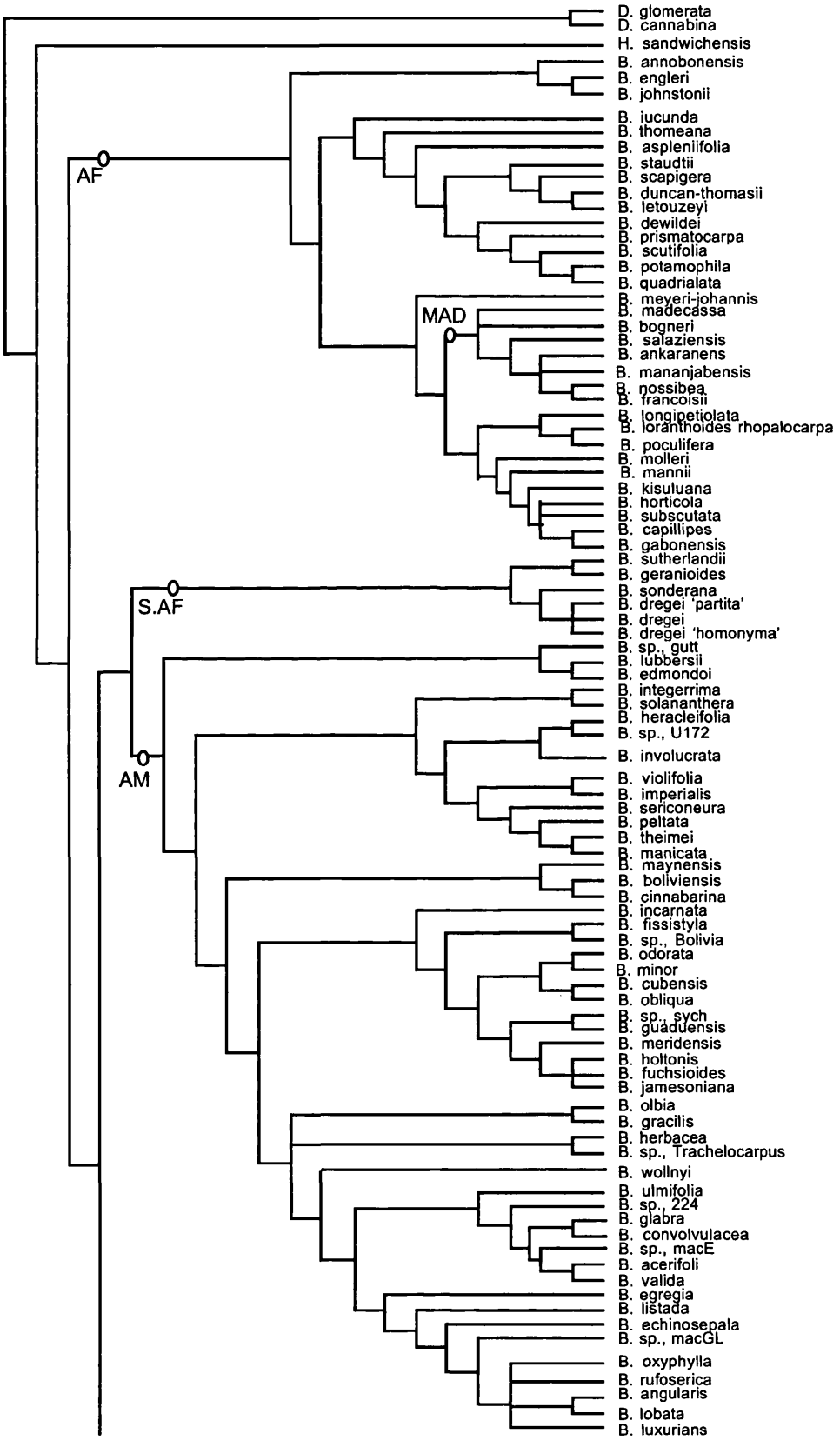
Figure 7.21: Pruned strict consensus of 10,000 MPTs, culled ITS data set

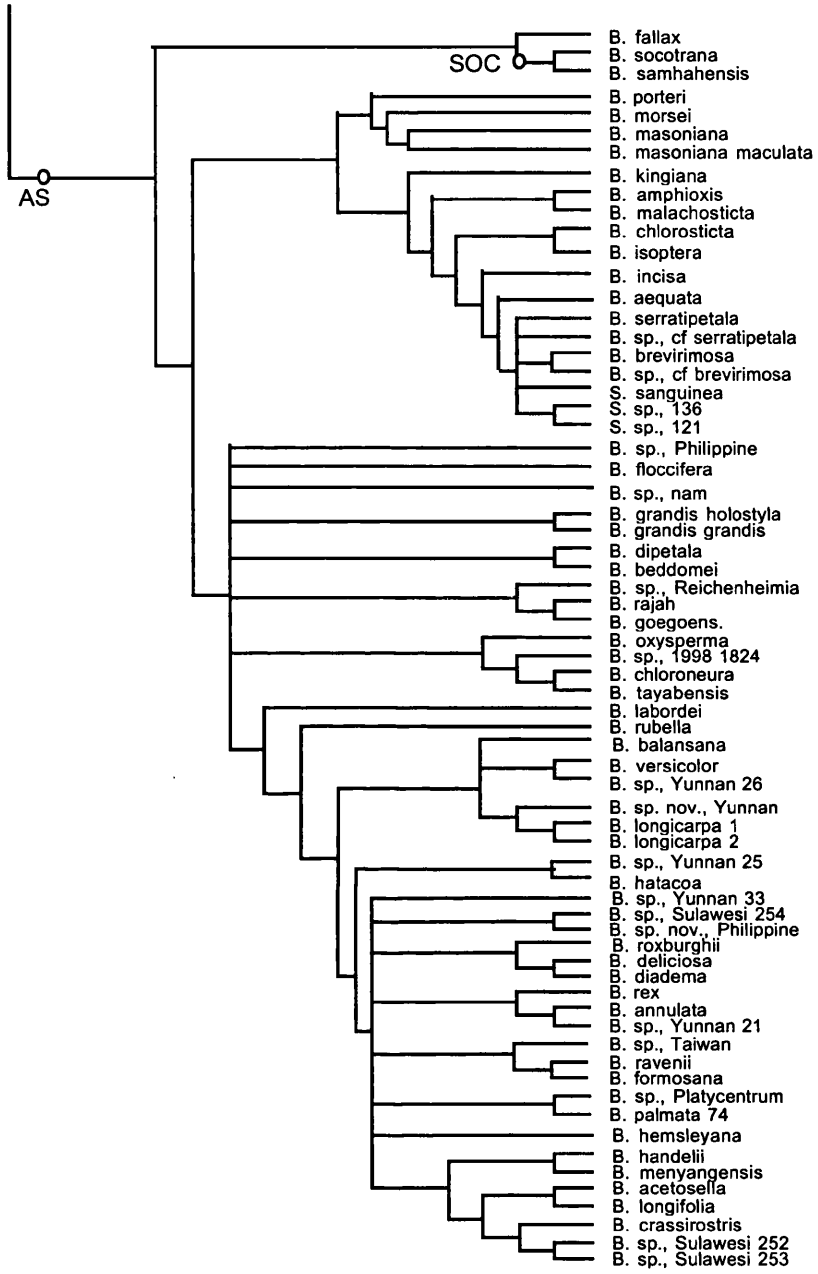


Matrix has 270 pars. inform. characters
Bootstrap values over 50% above lines,
Bremer support below lines

d. The Jigsaw Tree

Figure 7.22: ITS phylogeny of Begoniaceae (culled ITS sequence analysis, altered to reflect compartment topology)





The backbone from the 177 sequence analysis of variable-position culled ITS sequence data was taken; multiple sequences from the same individual were pruned from it for purposes of clarity. The topology was then altered to reflect the increased resolution within clades from the individual compartment analyses. Although this may mean that the topology shown (Figure 7.22, The Jigsaw Tree) is not necessarily a most parsimonious solution for the culled ITS data set, this does not mean that it cannot be our best estimate of phylogeny because, although there are alignment difficulties over the entire data set, the individual compartments (which largely represent within-clade variation) are less ambiguous: firstly, it was possible to tidy up the individual alignments, giving more reliable estimates of homology within them, and

secondly, positions which were excluded from the larger analyses due to being highly variable in some part of the data set could be included in these smaller analyses, meaning that more characters are being used in the compartments. In addition, given the smaller size of these compartment data sets, the search algorithms used are more likely to find globally rather than locally optimal trees (within each compartment).

7.4 Corroborating information from ITS sequences - Gap data:

The alignment of sequence data, by its nature, induces gaps into the alignment. Gaps, like point mutations, are phylogenetic events and can therefore offer information about evolutionary history. However, where potentially homologous gaps in different taxa are slightly different lengths, coding them as characters becomes complex. Simmons and Ochoterena (2000) only treat gaps with identical 5' and 3' termini as homologous, because if these are not identical, "at least one indel event must be postulated to turn one gap into another" (p. 371). They propose a simple coding method whereby all gaps with different 5' and/or 3' ends are coded as separate presence/ absence characters; completely overlapping gaps are coded as inapplicable.

McDade et al. (2000) discuss the use of indels in ITS sequence alignments as presence/absence characters. Gap characters have been found informative for ITS by several authors, including Jeandroz, Roy and Bousqet (1997) and Manos (1997). However, McDade et al. did not use an indel matrix because they had serious difficulties in aligning ITS across the Acanthaceae. From their aligned matrix they consider that "some [indels] may be informative in more narrowly circumscribed studies where alignments, and thus identification of indels, would be unambiguous" (p. 115).

However, even when gaps have not been coded as part of the matrix and used in the analysis of data, they can still be used to support (or possibly undermine) clades in the form of mapped characters. Prather and Jansen (1998), for example, mapped indel events onto an ITS phylogeny of *Cobaea* Cav. They found a high degree of congruence between the evidence from point mutations (the phylogeny) and information provided by indels (the mapped characters). Out of the 14 phylogenetically informative indels from their ITS matrix, only one was homoplastic.

The situation within the Begoniaceae is similar, in that the high levels of divergence across the family render gaps extremely problematic to code. However, there are some gaps within the matrix which appear to have strong phylogenetic signal, for example a 38 base pair gap near the end of the ITS 2 region is shared by nine American taxa. These taxa are resolved into a clade on the basis of ITS point mutations; therefore while the use of the gap as a character is not necessary to obtain this clade, it can help reinforce our faith in it. Similarly, there is a gap shared by all the Madagascan species at the start of the ITS 2 region.

Looking at the ITS alignment produced for this thesis (see CD-ROM), very few gaps are clear-cut; although it would be possible to deal with them in the manner Simmons and Ochoterena (2000) suggest, this would be time consuming for relatively little reward. Instead eight gaps without ragged edges were isolated by eye and coded as presence/absence. Gaps which are very similar but have slightly different 3' and/or 5' ends are named 'A' and 'B' (see Table 7.7).

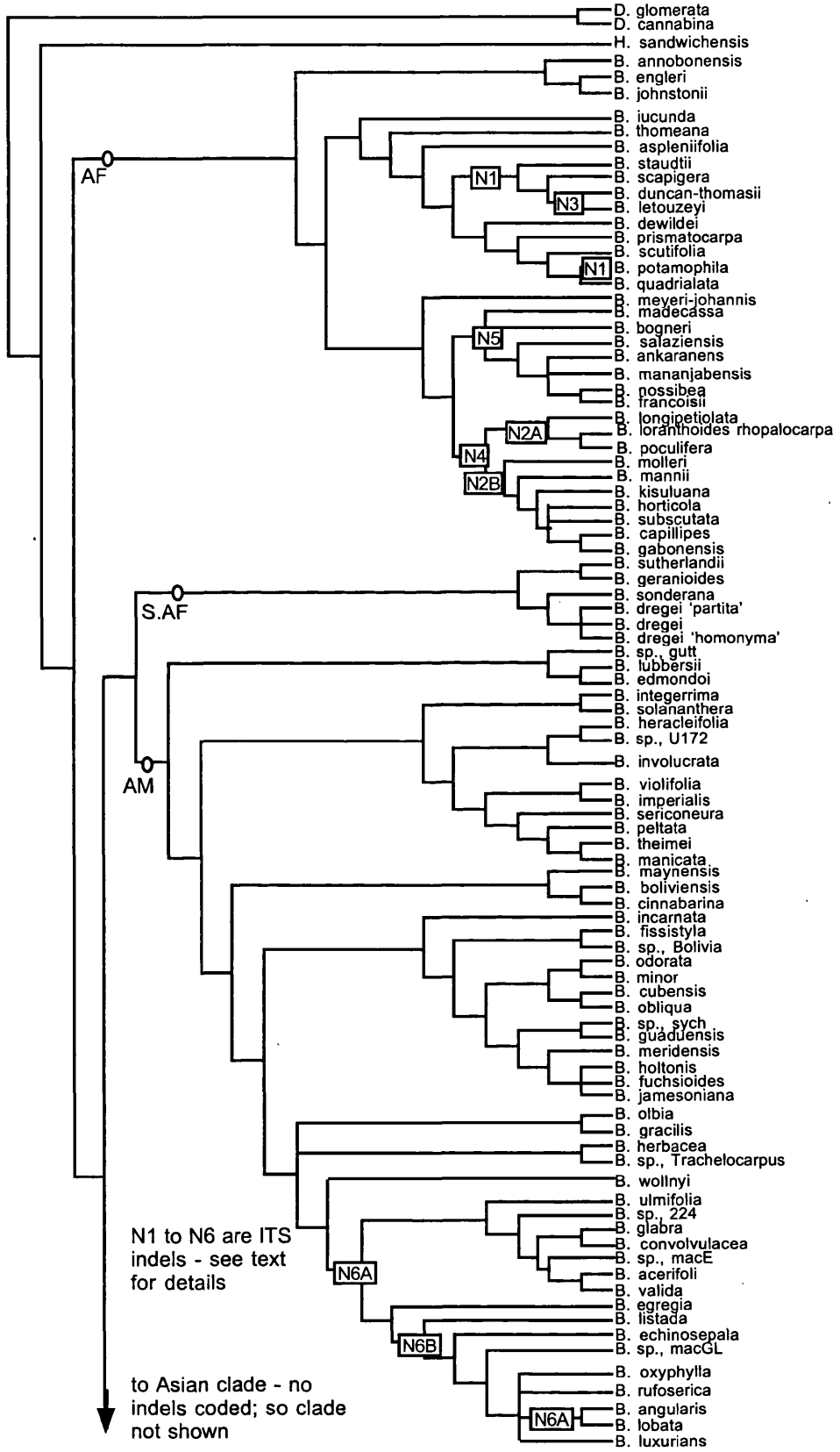
Table 7.7 Unambiguous gaps in ITS manual alignment

GAP	POSITION	TAXA
N1	172-180 G	*B. duncan-thomasii*, *B. staudtii*, *B. letouzeyi*, *B. potamophila* (uncertain taxa: *B. scutifolia*; *B. quadrialata*)
N2A	275-319 G	*B. longipetiolata*, *B. poculifera*, *B. loranthoides* ssp *rhopalocarpa*
N2B	275-324 G	*B. kisuluana*, *B. capillipes*, *B. mannii*, *B. molleri*, *B. horticola*, *B. gabonensis*, *B. subscutata*
N3	311-332 G	*B. duncan-thomasii*, *B. letouzeyi*
N4	354-359 S	*B. kisuluana*, *B. capillipes*, *B. mannii*, *B. molleri*, *B. horticola*, *B. gabonensis*, *B. subscutata*, *B. longipetiolata*, *B. loranthoides* ssp *rhopalocarpa*, *B. poculifera*
N5	693-695 G	*B. ankaranensis*, *B. madecassa*, *B. mananjabensis*, *B. salaziensis*, *B. nossibea*, *B. francoisii*, *B. bogneri*
N6A	1042-1100 G	*B. listada*, *B. echinosepala*, *B. rufosericae*, *B.* sp. macG, *B. oxyphylla*, *B. luxurians*
N6B	1061-1098 G	*B. glabra*, *B. convolvulacea*, *B. ulmifolia*, *B. acerifolia*, *B.* sp. macE, *B. valida*, *B. angularis*, *B. egregia*, *B. lobata*

(G = all the taxa cited share a gap; S = all the taxa cited share sequence)

The gaps coded in Table 7.7 were mapped across the 'Jigsaw' ITS phylogeny (Figure 7.22) (see Figure 7.23).

Figure 7.23: ITS phylogeny (the Jigsaw Tree) for African and American taxa, with coded ITS gaps mapped on



Gaps N2A, N2B, N3, N4 and N5 fit this topology perfectly. Gap N1 shows some homoplasy on this topology; gaps N6A also shows homoplasy (although this may be due to a problem with coding - if gaps N6A and N6B were homologous there would be no homoplasy). It certainly seems as if these ITS indels are in general agreement with the ITS point mutation data; clades in Africa and one clade in America ('*Pritzelia*') are supported by them.

7.5 General discussion and conclusions - Testing different alignments

The majority rule tree produced from automated alignment 4 is the most markedly different from those produced from other alignments and from the elision data set. It disperses several morphologically robust-seeming sections into clades which have no apparent geographic or morphological basis. The tree statistics for this data set (see Table 7.2) are by no means the worst for any data set - the consistency index is one of the lower values, but the retention index is among the higher end of the range, and the rescaled consistency index appears somewhere around the middle. Thus confidence measures themselves would not necessarily lead to the rejection of this hypothesis. The gap and extension penalties used in different alignments ranged from gap penalties of 3.3 to 45, and extension penalties of 3.3 to 15; alignment 4 is at neither end of these ranges, with a gap penalty of 30 and an extension penalty of 15. Extremes of the penalties range are represented in alignments 10 (3.3/3.3) and 8 (45/15). For alignment 8, the consistency index is low but the retention index is high; alignment 10 has a high consistency index and rescaled consistency index although a relatively low retention index.

Two things are worth noting here, firstly that values for these statistics are renowned to be negatively correlated with the number of characters and of taxa (Siebert, 1992) (the numbers of taxa are constant in each alignment; the difference in number of informative characters i.e. characters involved in the algorithms, particularly with autapomorphies excluded, is not nearly as large as the difference in the total number of characters). Secondly, the direct opposition of these numbers is not unexpected because with low penalties for creating and extending gaps, the bases will have been aligned in a way which maximises sequence similarity; thus there will be many characters which have (only) one or two states; where the penalties are higher and insertion of gaps

is discouraged, there are likely to be more multistate characters and so more potential for multiple state changes over a cladogram.

One surprising result of looking at different alignments is that, although the total numbers of characters in the data sets vary greatly (as discussed), the numbers of invariant characters vary from 43 (in alignment 15, although the value is also low for alignment 8, with 58 characters) to 405 (alignment 10) (range 362), the total numbers of parsimony informative characters only vary from 688 (alignments 13) to 756 (alignment 9) (therefore by only 68 characters) and do not show the inverse correlation with the gap opening and extension penalties that the other character numbers do. This means that the real size of each data set is similar, which increases the comparability of the tree statistics among the 16 alignments.

As the gap opening and extension penalties relax, the numbers of characters of all sorts increase, but the numbers of parsimony-informative characters fall **in proportion** to other sorts of characters. When the gap opening and extension penalties are higher, however, the numbers of multistate characters increase (as more data are shoe-horned into a smaller space) and thus the numbers of characters which are informative can also increase relative to other sorts of characters.

Although these methods can in many ways be considered purely data manipulation (even where a range of statistics are applied, to chose between solutions, it is difficult to be confident of the information provided by ambiguously aligned regions - and choices based on ranges of statistics have been shown for these Begoniaceae alignments to produce unconventional topologies) an awareness of the effects of altering alignments gives an indication of the amount of support for certain clades; I can see no reason not to feel more confident of clades which can withstand such data manipulation.

One important point is the difficulty and importance of rooting; due to the divergence between outgroup and ingroup, the OG/IG relationship is highly sensitive to alignment and can dramatically alter the evolutionary hypothesis. There is a weight of evidence for a sister group relationship between the outgroup and African taxa (26S sequences, *trnL* (Plana, unpublished data, 2001), unambiguously aligned ITS regions, the elision data and Badcock's

(1998) *trnC-trnD* intron data); in some alignments, however, this is not the favoured hypothesis. The position of *Hillebrandia* in many alignments is more reliable (according to conventional hypotheses) as a root; however rooting on *Hillebrandia* is not advisable if one is interested in testing its position within Begoniaceae. The sister group relationship between *Hillebrandia* and *Begonia* is geographically unusual in that it necessitates acceptance of an Hawaii/Africa disjunction; Hawaii has always been isolated from the continental land-masses and has been populated through long-distance dispersal events (Kim et al., 1998). Africa has never been proximal to Hawaii (and is currently c. 15,000 km away), which makes it a surprising place to find *Begonia*'s closest relative.

A remarkably shorter search time for swapping the elision data set to completion when compared to TBR swapping of the individual alignments appears to be due to the decreased probability of hitting islands of equally parsimonious trees as the amount of data increases (although comparisons of starting and final tree lengths were not made; it is possible that shorter distances between these lengths could be responsible for shorter analysis times - see Savolainen et al., 2000). Instead the search descends rapidly to some local minimum. With fewer character differences in the individual automated alignment data sets, many trees can share the same length (MaxTrees was hit in eight out of 16 searches). Island structure of the elision data set was revealed by running 1000 random addition replicates. No island contained more than 24 trees, while most contained less than 10. The lengths of equally parsimonious trees found on different islands ranged by 138 steps, from 175227 to 175464.

Of course, the difference between the elision and the individual alignment data sets is more subtle than 'character differences', as these different data sets do not represent different character sets. Instead the elision represents a special form of character weighting. Successive weighting of characters, for example by their rescaled consistency index, can also be used to analyse data where large numbers of equally parsimonious trees are found by analyses using Fitch analysis. Successive weighting favours characters which perform consistently over an hypothesis (cladogram); elision (weighting by alignability) is not dependent on an initial parsimony stage (although will be influenced by the distance measure used in CLUSTAL for alignment generation).

There is no *a priori* reason to accept or dismiss one set of trees on the basis of similarity or difference from another set of trees (such as the elision trees) although a more parsimonious interpretation of data would seek congruence with other data sets. Comparisons with the elision tree are more complex, in that elision does not represent another data set but a manipulation of the sets under study; however, in that it weights conserved regions, it should be more robust than the individual automated alignments.

The position of the root is critical for the interpretation of evolutionary events within *Begonia*; several alignments give unconventional rooting. However these (badly) rooted trees may more accurately reconstruct relationships within the ingroup; a major problem with ITS sequence data in *Begonia* is the difference in levels of sequence divergence between different clades. Parameters which perform badly across wide sequence divergences may well reconstruct more accurately the relationships within (and between closely related) clades.

The different alignments contain two classes of characters which could be loosely described as 'homoplasy'. Some of the observed homoplasy is due to taxa sharing derived characters which are not due to common ancestry but to convergence (e.g. multiple hits); this sort of homoplasy is relevant, for instance, if we are interested in mutation rates, and may be locally informative phylogenetically. However, the other class of homoplasy is due to nonsense characters ('hybrids' of true characters, created by data misalignment) and has nothing to say about evolution. There is no way to separate these except through careful culling of our data. However, in these analyses the consistency index is actually lower for the culled tree than for the unculled tree - the removal of 336 parsimony-informative characters caused a fall in consistency index (uninformative characters excluded) from 0.271 to 0.267.

Despite the arguments of some of the more transformed cladists (such as Wenzel & Siddall, 1999) that the exclusion of data is undesirable regardless of the presumed levels of homoplasy, it is apparent that because differences in alignment do not create straightforward homoplasy but 'nonsense' characters, the exclusion of data which align ambiguously across the taxa being considered is the most reliable option. However, elision, although including

'nonsense' characters, should downweight these characters to an extent where they do not have a huge influence on the overall topology.

Compartmentalization (Mishler, 1994) as an alternative method of using all (or almost all) the data offers the benefit of avoiding any use of 'nonsense' characters but still obtaining the fine resolution that is more likely to be obtained from more variable regions of sequence; in practice, however, it is time-consuming and has offered little overall advantage in terms of support and resolution with this data set.

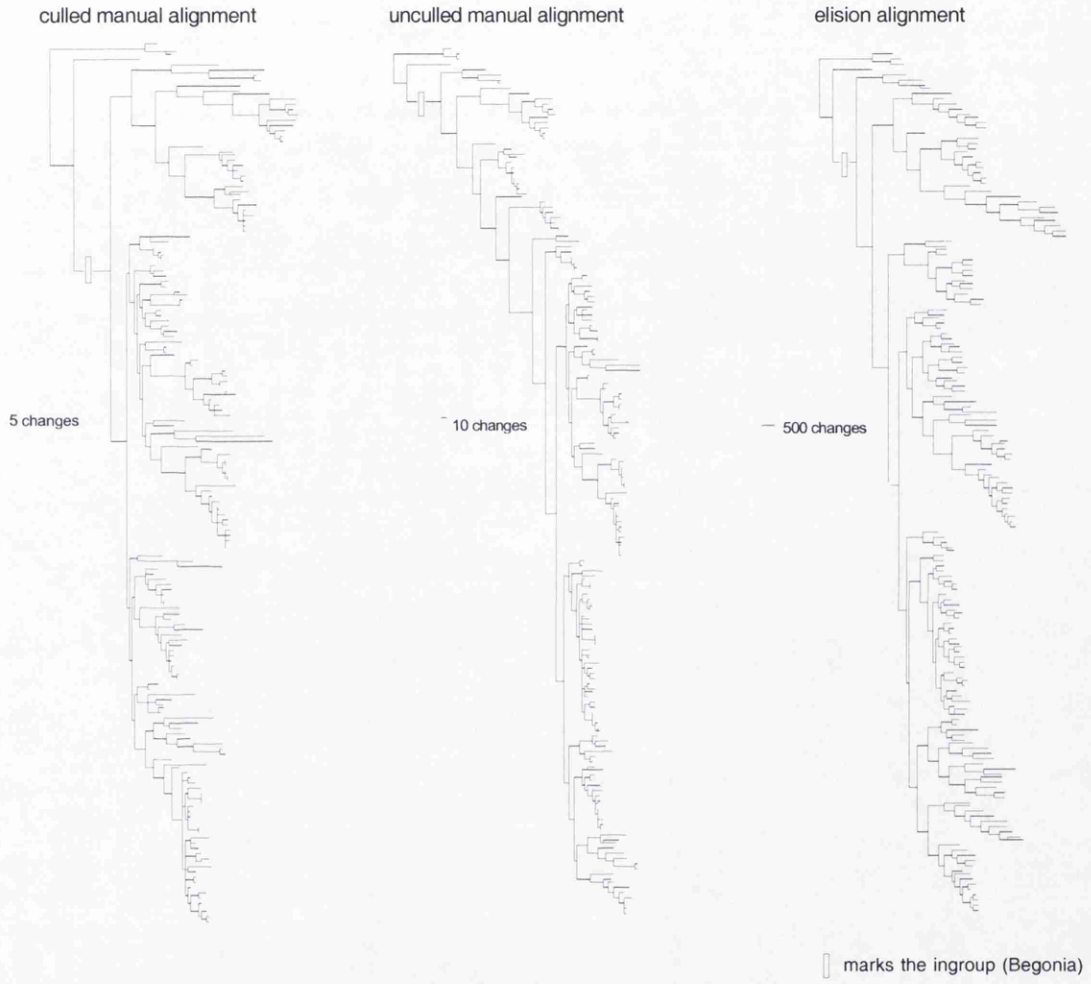
Manual alignment is more sensitive (than using an algorithm which has set parameters) to the real situation wherein parts of a sequence have diverged more or less than other parts. The conserved regions will reconstruct higher order phylogeny, while the less conserved regions can reconstruct the hierarchy of closely related taxa, and so the probability of indel events in different parts of the sequence will vary. Another reason to support manual over automated alignment is that different mutational events (e.g. duplication, inversion, repetitive DNA) all require different alignment decisions.

The manual alignment, even with the misaligned parts of sequence included ("unculled"), although providing less parsimony-informative characters than any of the automated alignments, has better consistency, retention and rescaled consistency index values, suggesting that there are fewer homoplastic characters in the matrix. Thus if one accepts the alignment which optimises values for the consistency, retention and rescaled consistency indices, the manual alignment outperforms the automated alignments.

The three data sets considered here (elision, manual, unculled and manual, culled) give basically similar tree shapes (see Figure 7.24), with less species-rich clades on longer branches at the base of the trees. The overall tree shapes are unbalanced, and many of the clades on the trees are also unbalanced. The unculled, manual alignment gives the most unbalanced topology, this is due to African taxa resolving as highly paraphyletic; analyses of the culled manual alignment and the elision alignment resolve many of the African species within one clade. The African taxa suffer most alignment ambiguity, with blocks of sequence alignable within but not between clades. This artifact of alignment is most probably responsible for the patterns

observed here, with the differences between clades overriding their shared characters in the unculled manual alignment.

Figure 7.24: Tree shape (phylograms) for manual (culled and unculled) and elision data sets



7.6 Summary

Different alignment methods (automated, elision and manual) were tested using an ITS sequence matrix. Further, different ways of analysing large data sets were tested, by heuristic parsimony analysis of the manually aligned 177-sequence matrix, followed by compartmentalization of that matrix. Analyses were run for eight compartments isolated from the total matrix; these were used to test the performance, in parsimony analyses, of the elision and manual alignments from the beginning of the chapter. The manual alignment for ITS, with ambiguous regions of sequence excluded, performed best (i.e. produced topologies most similar to those produced in the compartment analyses). Three ways of dealing with the taxa which were not included in the compartments were also tested (constraint trees, hypothetical ancestors and exemplar taxa). Of these, using exemplar taxa was preferred, although analysis of the complete 177-sequence matrix gave better results. A topology was constructed using the 177-sequence manual alignment, culled of ambiguous regions, and adjusted to reflect the topologies of the compartment analyses (and was nicknamed the 'Jigsaw' tree, Figure 7.22). Finally, some of the indel events in the ITS matrix were coded and mapped across this 'Jigsaw' topology. The indel data were shown to be phylogenetically informative in this case, and reinforced some of the clades identified using nucleotide substitutions. The 'Jigsaw' topology will be used in further chapters, to discuss evolution in the Begoniaceae.

8. Secondary structure

8.1 Introduction

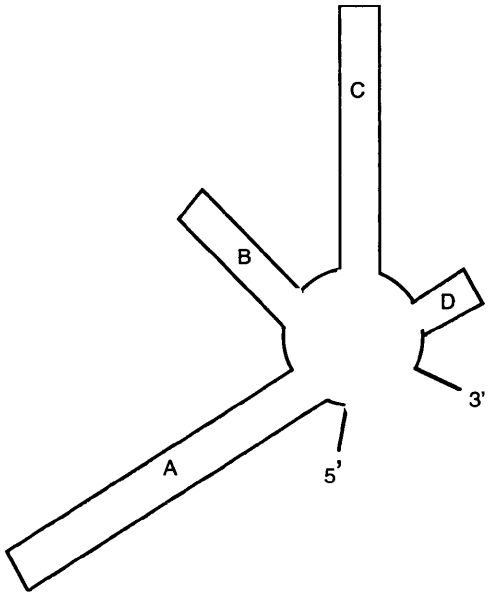
8.1.1 Length of ITS regions: Baldwin et al. (1995) reported that all flowering plants examined had less than 300 base pairs in both ITS 1 and ITS 2. The longest regions they report are in Malvaceae, which has up to 298 base pairs for ITS 1, and Cucurbitaceae, up to 252 bases for ITS 2. ITS 1 is usually 200 - 300 bases long; ITS 2 is usually 180 - 240 bases and 5.8S, 163 - 165 bases (Hershkovitz, Zimmer & Hahn, 1999).

There are also a few reported lengths for taxa in the Cucurbitales (i.e. taxa phylogenetically close to *Begonia* according to recent classifications, such as Savolainen et al., 2000). In *Corynocarpus* ITS 1 varies between 234 - 258 bases, while ITS 2 is shorter, between 197 and 220 bases. The 5.8S region is 165 bases (Wagstaff & Dawson, 2000). Jobst, King and Hemleben (1998) found that the ITS 1 spacer varies from 186 to 233 bases in Cucurbitaceae, while ITS 2 varies from 224 to 263 bases.

8.1.2 Secondary structure: Hershkovitz and Zimmer (1996) produced secondary structures for ITS 2 for nine angiosperm genera. All but one consist of a central loop with four radiating stems (three stems in *Arabidopsis*). Mai and Coleman (1997), who reconstructed similar structures for angiosperms and algae, describe ITS 2 as a "self-contained folding complex, usually consisting of four distinct hairpin loops" (p. 262). They found a lot of sequence conservation near the bases of each of the four stems, and note that the considerable length variation found in ITS 2 does not impede the formation of these four conserved structures.

Mai and Coleman (1997) found that the first stem (which I will call 'A') can be highly variable in length and sequence; the second ('B') shows extensive nucleotide covariation; the third ('C') is generally the longest region, and also shows the highest degree of structural conservation; the fourth (D) is the most variable region within their study (see Figure 8.1 for schematic representation of ITS 2 secondary structure).

Figure 8.1: Schematic summary diagram of ITS 2 secondary structure



Stem lengths in the angiosperms: Stem A is between 31 (*Sinapsis*) and 55 (*Cucurbita*) bases long; B, from 16 (*Arabidopsis*) to 39 (*Vicia*); C, from 84 (*Oryza*) to 108 (*Canella*) and the fourth stem is absent in *Arabidopsis*, up to 34 bases in *Cucurbita* (values from Hershkovitz & Zimmer, 1996; Mai & Coleman, 1997).

8.2 Material and methods

The lengths of ITS 1 and ITS 2, and of the 5.8S region between them, were estimated using the start and end points from Hershkovitz and Zimmer (1996). Secondary structure reconstruction was undertaken for a selection of sequences (sequences were obtained as described previously). One of the reasons for reconstructing secondary structure in *Begonia* was to assist with difficulties in manual alignment of sequences across the genus. Taxa were selected from an initial manual alignment, focusing on taxa which were problematic to align, and where possible, more than one taxon from groups of similar sequences were selected in order to check that similar structures were obtained. Taxa with the fewest ambiguous base calls were preferred.

Secondary structure was determined using MulFold (Zuker, 1989; Jaeger, Turner & Zuker, 1989a; Jaeger, Zuker & Turner, 1989b). Foldings were done at 20°c, saving up to 15 folds within a 10% range from the optimal free energy

value. The folds were viewed in LoopDloop (Gilbert, 1995) and compared to ITS 2 secondary structures published by Hershkovitz and Zimmer (1996) and by Mai and Coleman (1997). Foldings were also made to determine the secondary structure of ITS 1, but this was abandoned as a common pattern could not be determined from the range of structural variants found. No parts of 18S, 5.8S or 26S were included in the foldings, partly to follow Hershkovitz and Zimmer (1996), also because including up to 50 bases from each of the flanking coding regions would have made many of the sequences over 300 base pairs long (creating analytically difficulties in MulFold), and also because trial folds made with parts of the coding regions included were less comparable with the secondary structures found by previous authors (Hershkovitz & Zimmer, 1996; Mai & Coleman, 1997). Sequences were not constrained in any way.

8.3 Results

The length of ITS 2 is very variable in *Begonia*, ranging by 149 bases, from 212 in *B. angularis* to 360 in *B. masoniana* var. *maculata* (see Table 8.1); in several species it is over 300 base pairs, which is unusually long for the angiosperms (see Baldwin et al., 1995). ITS 1 is less variable. It ranges from 223 in *B. thomeana* to over 270 in *B. prismatocarpa*[5], a difference of over 47 base pairs. The 5.8S region, on the other hand, is largely invariant at 144 base pairs.

[5] part of the start of ITS 1 is missing for *B. prismatocarpa* and for *B. salaziensis*, so exact values cannot be given

Table 8.1: The length of ITS 1, 5.8S and ITS 2 for representative taxa

TAXON	ITS 1	5.8S	ITS 2
D. cannabina	257	159	222
D. glomerata	263	157	222
B. molleri	231	146	267
B. gabonensis	232	146	269
B. nossibea	247	144	304
B. salaziensis	>232	144	304
B. bogneri	249	144	240
B. thomeana	223	144	306
B. iucunda	251	144	293
B. engleri	258	144	?
B. madecassa	248	144	273
B. duncan-thomasii	240	144	298
B. prismatocarpa	>270	144	306
B. aspleniifolia	244	144	298
B. socotrana	258	144	317
B. samhahensis	268	144	330
B. dregei	257	144	274
B. sonderana	258	144	275
B. annobonensis	258	144	225
B. hemsleyana	250	144	306
B. palmata	254	144	306
B. masoniana	260	144	360
B. kingiana	256	144	308
B. tayabensis	260	134	317
B. aequata	262	144	300
Symbegonia	258	144	301
B. cubensis	244	144	296
B. fissistyla	255	144	295
B. oxyphylla	268	144	267
B. valida	256	144	281
B. angularis	257	144	212

Secondary structure was reconstructed for 21 taxa (see Table 8.2 for taxon names; accession details etc. are given elsewhere). Examples of the secondary structure of *Datisca glomerata* (the outgroup) (Figure 8.2), *B. nossibea* (Figure 8.3) and *B. gabonensis* (Figure 8.4) (African) *B. socotrana* (Figure 8.5) (Socotra) *B. hemsleyana* (Figure 8.6), *B. aequata* (Figure 8.7), *B. masoniana* (Figure 8.8) and *Symbegonia* sp. (Figure 8.9) (Asian) and *B. fissistyla* (Figure 8.10) and *B. oxyphylla* (Figure 8.11) (American) are given.

Figure 8.2: *Datisca glomerata* ITS 2 secondary structure (free energy -102.8)

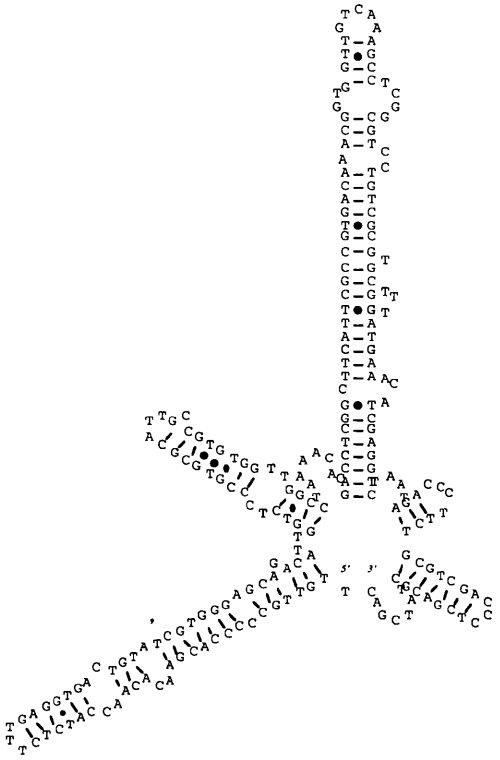


Figure 8.3: *B. nossibea* ITS 2 secondary structure (free energy -103.9)

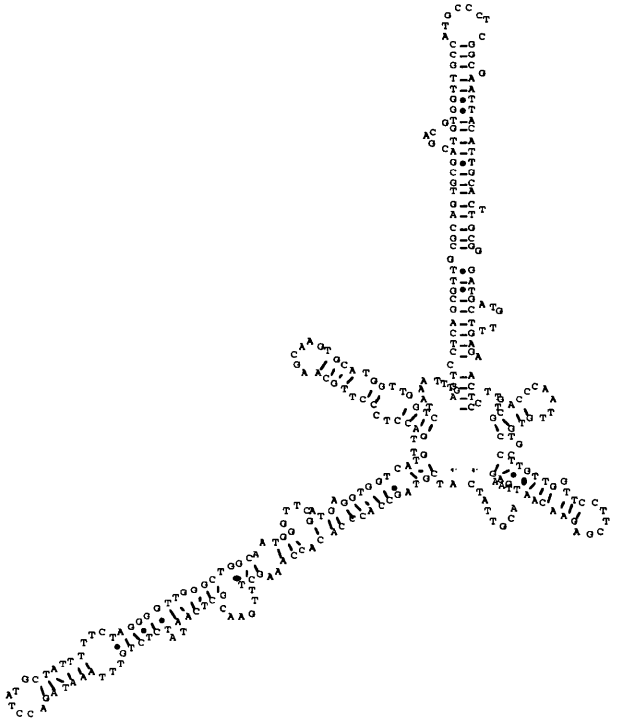


Figure 8.4: *B. gabonensis* ITS 2 secondary structure (free energy -145.5)

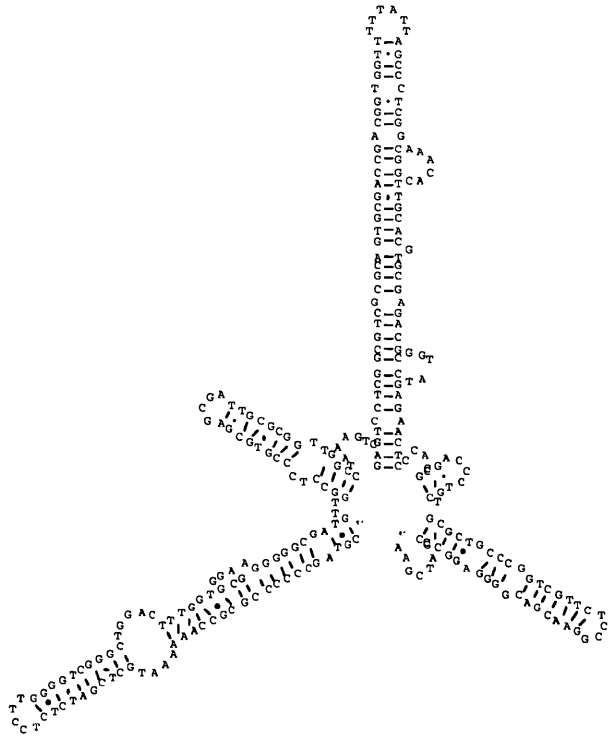


Figure 8.5: *B. socotrana* ITS 2 secondary structure (free energy -175.3)

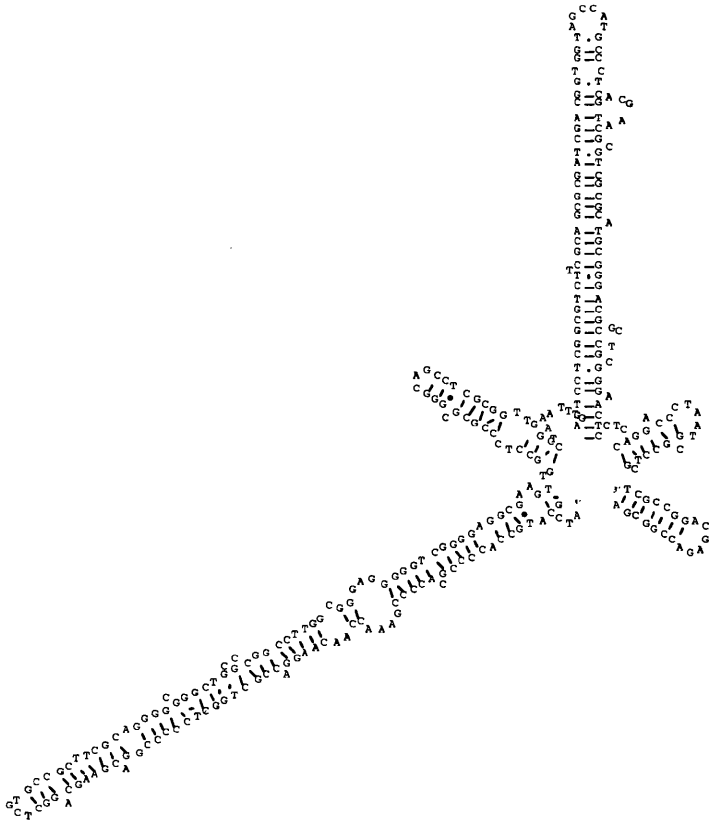


Figure 8.6: *B. hemsleyana* ITS 2 secondary structure (free energy -140.0)

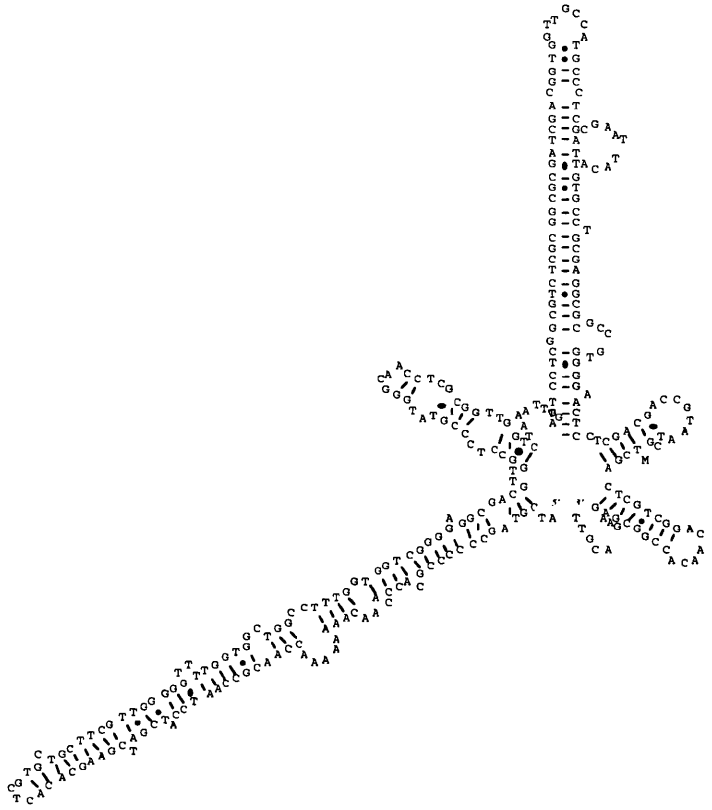


Figure 8.7: *B. aequata* secondary structure (free energy -119.5)

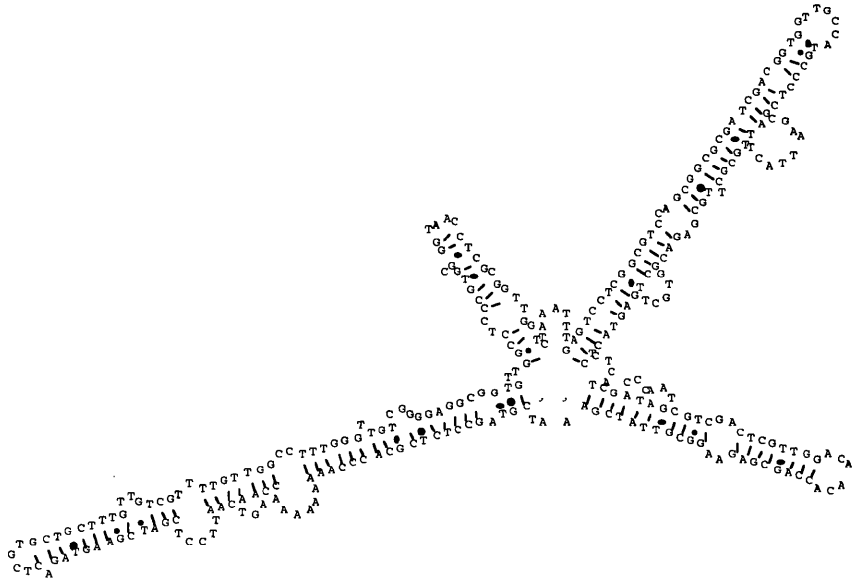


Figure 8.8: *B. masoniana* ITS 2 secondary structure (3' end cut short) (free energy -130.1)

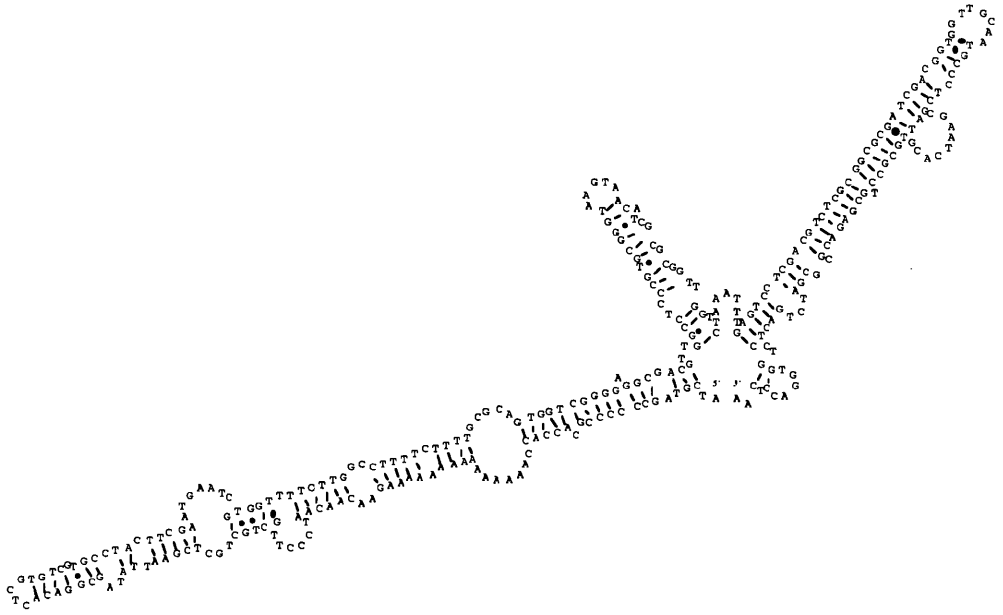


Figure 8.9: *Symbegonia* sp. 136, ITS 2 secondary structure (free energy -131.8)

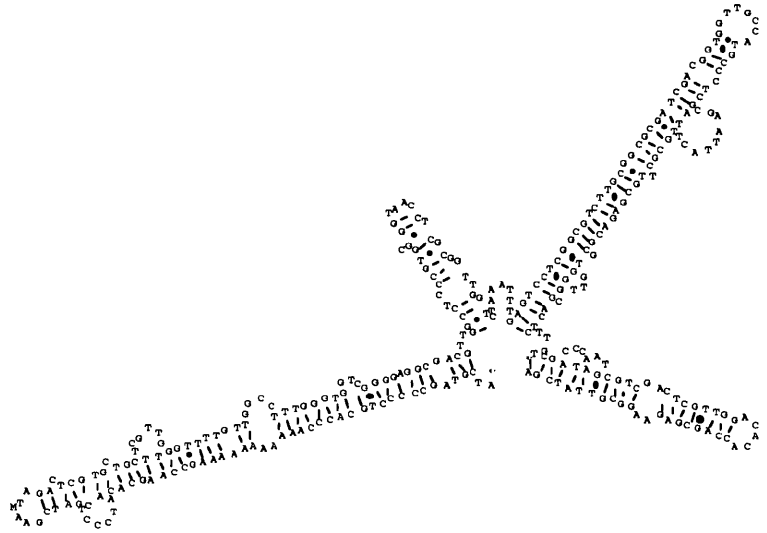


Figure 8.10: *B. fissistyla* ITS 2 secondary structure (free energy -120.3)

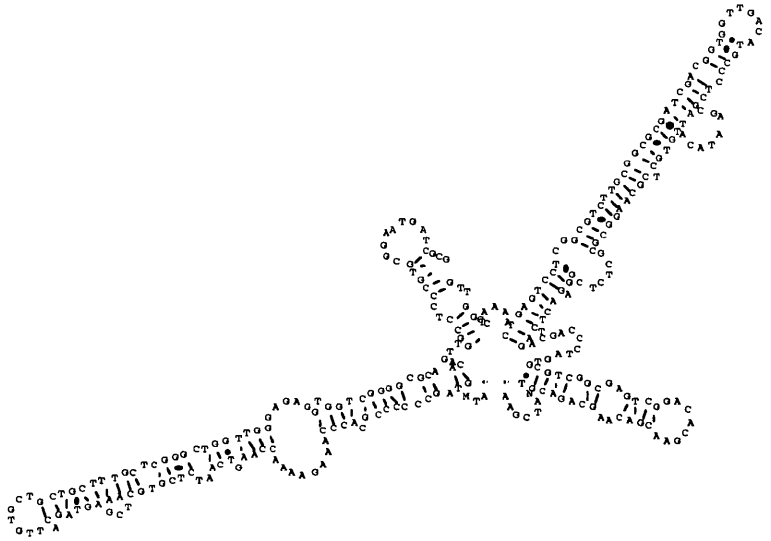


Figure 8.11: *B. oxyphylla* ITS 2 secondary structure (free energy - 115.8)

Structures with four to five stems were obtained for most of the taxa analysed. All taxa share a highly conserved second (B) and third (C) stem; most of the variability is in the first stem (A). The reconstruction of the fourth stem was often ambiguous (with a fifth stem sometimes present); therefore the lengths of the fourth stem are not considered here.

From Table 8.2, it can be seen that, of the three stems considered, stem B is almost always invariant at 32 bases; stem C is slightly more variable, between 88 and 91 bases in *Begonia* (83 to 84 in *Datisca*), and stem A is the most highly variable, between 42 (*B. angularis*) and 148 (*B. masoniana*) bases long.

Table 8.2: Lengths of the first, second and third stems from secondary structure reconstructions of ITS 2

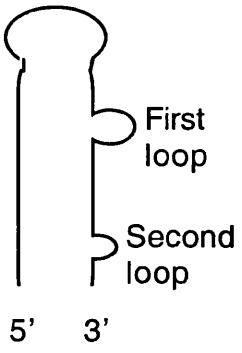
TAXON	A	B	C
C. melo [6]	55	32	90
D. cannabina	52	32	83
D. glomerata	52	32	84
B. molleri	82	32	88
B. gabonensis	76	32	91
B. nossibea	107	32	90
B. salaziensis	107	32	88
B. bogneri	?	32	90
B. iucunda	103	32	90
B. prismatocarpa	?	32	90
B. socotrana	123	32	90
B. dregei	83	32	90
B. sonderana	84	32	90
B. hemsleyana	115	32	90
B. palmata	115	32	90
B. masoniana	148	38	?
B. aequata	108	32	90
S. sp 136	111	32	90
B. fissistyla	105	31	89
B. oxyphylla	105	32	91
B. valida	108	32	90
B. angularis	42	32	91

[6] Values for *Cucurbita melo* are based on the secondary structure depicted by Hershkovitz and Zimmer (1996, p. 2864).

8.4 Discussion

The highly conserved structure of stem B may suggest some sort of functional constraint; stem C also has some highly conserved secondary structure including a loop quite near to the top of the 3' side of the stem, and often a second loop nearer to the base of the stem on the same side (see Figure 8.12) (although the first loop is not evident in the reconstruction shown for *B. nossibea,* Figure 8.3).

Figure 8.12: Schematic diagram of Stem C showing conserved secondary structure



B. angularis (not shown) has distinctly shorter ITS 2 than other *Begonia* species sampled; this appears to be due to deletion of the end of stem A; there is also a deletion around the region of the fourth stem. *B. oxyphylla*, which also has a short ITS 2, has lost bases from the region where the 4th stems occurs (see Figure 8.11). The very long ITS 2 sequences found in *B. socotrana* (Figure 8.5) and *B. masoniana* (Figure 8.8) are mainly caused by the extended length of stem A. The most ambiguities in the sequence alignments for ITS 2 are around the region of stem A (see chapter 7); although similar taxa have similar sequences in this region, disparate taxa have very low levels of sequence similarity.

The sequence for *Begonia masoniana* was obtained from a clone of *B. masoniana* var. *maculata*, due to difficulties in obtaining readable sequence from consensus PCR of *Begonia masoniana* or *B. masoniana* var. *maculata*. The difference in the second ITS 2 stem (B), which is 34 bases rather than the 32 bases it is in the rest of the *Begonia* species examined, suggests that we could have amplified a non-functional paralogue. There are also a few mutations in the 5.8S sequence for other cloned *B. masoniana* var. *maculata* (which also have 34-base-long B stems) One clone has a one base pair

deletion at character 555; one has a T to C mutation at character 578, and another has a T to C mutation at character 591. However, not only do the sequenced clones nest together (and with the parts of consensus sequence which could be read) in parsimony analyses (see chapter 7, Figures 7.1, 7.3, 7.5), but *B. masoniana* also resolves with the consectional *B. porteri* and *B. morsei*, which were not cloned (i.e. the sequences give the 'expected' topology, and no obviously atypical relationships are inferred). Unfortunately neither *B. porteri* nor *B. morsei* gave complete sequences; ITS 1 for *B. morsei* is 255 bases, while ITS 2 is over 287 bases (the first 15 or more bases are missing); *B. porteri* was less complete. Due to these problems with incomplete sequences, it was not possible to reconstruct secondary structure for these other species.

Further studies involving cloned ITS *Begonia* sequences, particularly for section *Coelocentrum*, are needed to see whether some sequences exist which form secondary structures more within the range shown by other *Begonia* species (and so, whether it is likely that we have recovered some paralogous and potentially non-functional members of a gene family).

9. *trnC* to *trnD*; separate and combined analyses with ITS

9.1 Introduction

The ITS region has been discussed in earlier chapters. The non-coding, non-transcribed region *trnC* - *trnD* is located in the large single-copy region of the chloroplast. It varies in length from c. 3000 to 4000 bases in *Begonia*; it is AT rich, with a large number of simple sequence repeats (Badcock, 1998). Badcock did not sequence the entire region, but obtained sequence data inwards from universal primers located in the tRNA genes (sequencing in from the *trnC* region at the 5' end, and the *trnD* region at the 3' end).

Because of differences in inheritance, phylogenies reconstructed from nuclear and chloroplast regions can vary, particularly where there have been evolutionary reticulations. Using DNA data from two genomes (e.g. McDade et al., 2000 - ITS and *trnL-trnF*, Acanthaceae) potentially allows the tracking of these different evolutionary histories, with biparental inheritance of nuclear DNA and predominantly uniparental inheritance of chloroplast DNA (in most angiosperms).

9.2 Material and Methods

The *trnC* - *trnD* matrix was taken from Badcock (1998); ITS sequences were taken from a larger matrix compiled for this thesis (chapter 7). Voucher details are the same as in previous chapters (and are on the CD-ROM). A list of the taxa included in this study is presented in Table 9.1.

Although the *trnC* - *trnD* data has already been analysed by Badcock (1998), analyses were rerun with *B. oacacana* A.DC. excluded, as it did not prove possible to amplify the ITS region for this taxon. Badcock (1998) found *Datisca* species amplified poorly for *trnC* - *trnD*; therefore she used a consensus sequence for the two species. There was no problem getting sequence for ITS for *Datisca*; *D. cannabina* was used in place of a consensus sequence in the ITS analyses.

Badcock (1998) provides an indel matrix for *trnC - trnD*, which she included in her analyses and found to be phylogenetically informative. However, there is not an indel matrix for the ITS data set, due to the ragged nature of many of the indels; because the purpose of this chapter is to compare phylogeny reconstruction from *trnC - trnD* and ITS, the gap matrix for *trnC - trnD* was consequently not used. However, several unambiguous gaps from the *trnC - trnD* matrix were coded and mapped across MPTs produced from MP analysis of both the *trnC - trnD* and the ITS data sets to see whether there is any conflict in their signal.

9.2.1 Taxa included in this study

Table 9.1: Summary of taxa included in molecular analyses.

SPECIES	SECTIONAL PLACEMENT	GEOGRAPHIC DISTRIBUTION	SOURCE AND ACCESSION No
B. acerifolia H.B.K.	Knesbeckia	America: Ecuador	GL 001 057 96
B. acutifolia Jaq.	Begonia	America: West Indies	GL 002 1147 66
B. convolvulacea (Klotzsch) A.DC.	Wageneria	America: Brazil	GL 001 093 79
B. dipetala Graham	Haagia	Asia: S. India & Sri Lanka	GL 003 018 96
B. dregei Otto & Dietrich	Augustia	Africa: S. Africa	GL 004 026 94
B. dregei Otto & D. non 'partita'	Augustia	Africa: S. Africa	GL 002 036 89
B. floccifera Bedd.	Reichenheimia	Asia: S. India & Sumatra	GL 030 099 89
B. goegoensis N.E.Br.	Reichenheimia	Asia: Sumatra	GL 011 125 57
B. gracilis Humb., Bonpl. & Kunth	Quadriperigonia	America: Mexico	Z. Badcock no. 9
B. grandis Dryand. ssp grandis	Diploclinium	Asia: China	GL 004 085 80
B. grandis Dryand. ssp holostyla	Diploclinium	Asia: China	E 1998 0035
B. heracleifolia Schltdl. & Cham.	Gireoudia	America: Mexico	GL 001 126 83
B. incarnata Link & Otto	Knesbeckia	America: Mexico	GL 011 089 95
B. malachosticta Sands	Petermannia	Asia: Malesia, Sabah	GL 010 117 94
B. mannii Hook.f.	Tetraphila	Africa: Nigeria, Eq. Guinea, Cameroon	GL 008 067 80
B. masoniana Irmsch.	Coelocentrum	Asia	GL 001 007 56
B. maynensis A.DC.	Knesbeckia	America: Peru, Ecuador	GL 001 107 92
B. obliqua L.	Begonia	America: Martinique	GL 005 105 91
B. olbia Kerch.	Knesbeckia	America: Brazil	GL 002117 94
B. palmata D.Don	Platycentrum	Asia: China	E 1998 0048
B. peltata Otto & Dietrich	Knesbeckia	America: Mexico, Central America	GL 308 000 xx
B. rajah Ridl.	Reichenheimia	Asia: Malaya	GL 003 081 96
B. ravenii C.I.Peng & Y.K.Chen	Diploclinium	Asia: Taiwan	E 1993 3938
B. roxburghii A.DC.	Sphenanthera	Asia: N.E. India to Burma	GL 004 093 79
B. rubella Buch.-Ham. ex D.Don	Diploclinium	Asia: Nepal	GL 005 094 94
B. salaziensis (Gaud.) Warb.	Mezieria	Africa: Reunion, Mauritius	K 1986 412
B. sutherlandii Hook.f.	Augustia	Africa: S.Africa & Tanzania	E 1971 1552
B. tayabensis Merr.	Reichenheimia	Asia: Philippines	GL 006 035 89
B. ulmifolia Willd.	Donaldia	America: Venezuela	GL 014 125 57
B. wollnyi Herzog	Knesbeckia	America: Brazil, Bolivia	GL 003 057 96
Datisca cannabina L.		Asia: S.W. Asia to Himalaya	E 1969 4093
Datisca glomerata (Presl) Baill.		America: California, USA	S.Swensen 767
Symbegonia sanguinea Warb.		Asia: New Guinea	GL 003 127 93

9.2.2 Analyses

MP analyses were run on three data sets: *trnC - trnD* (from Badcock, 1998); the corresponding species for ITS; the two regions combined. ME and ML analyses were run only on the *trnC - trnD* data set.

To look at cladistic structure within the data matrices, PTP was estimated with the outgroup (*Datisca*) excluded (although the outgroup in this case is only one taxon so should not influence character covariance), 1000 PTP replicates, 10 random addition replicates, TBR, steepest descent, five trees saved per step. G1 was estimated using 10,000 random trees. Uncorrected pairwise distances and the base composition of the matrix were obtained from PAUP* 4.0b2a (Swofford, 2000).

a. MP (Maximum parsimony): MP searches were performed with 1000 random addition sequence replicates, using TBR, saving no more than 10 MPT at any step. The trees from the initial search were input as starting trees for a further heuristic search, with TBR, and no limit on the number of MPTs saved.

Bootstrap support was estimated with 1000 replicates, heuristic search, 10 replicates random addition per bootstrap replicate, no more than 10 trees of any length held, TBR, steepest descent. Bremer support was estimated using AutoDecay (Eriksson, 1998), with 10 random addition replicates per constraint tree, TBR.

b. ML (Maximum likelihood): Likelihood was only used for the *trn*C - *trn*D data set, because of time limitations. The tree was constructed using the methodology described in Chapter 5, Section 5.2.4.2.

c. ME (Minimum evolution): Again, ME was only used for the *trnC - trnD* data set, as a comparison to the trees produced by MP and ML. The tree was constructed using the methodology described in Chapter 5, Section 5.2.4.3.

9.3 Results

All the trees presented have geography marked onto the clades. AF = Africa, S.AF. = southern Africa, AM = America and AS = Asia.

9.3.1 *trnC - trnD*:

a. Data matrix: There were 324 characters excluded; 1569 constant, 369 parsimony-uninformative, and 188 parsimony informative characters were included. Excluded characters are 348-362, 497, 748-901, 1166-1172, 1330-1332, 1424-1436, 1786-1792, 1971, 2021-2035, 2117 and 2344-2450 from the matrix in Badcock (1998).

The PTP probability is 0.002; the skewedness statistic g1 is -1.364.

Uncorrected pairwise distances vary from 0.000 between *B. dregei* and *B. dregei* '*partita*' (conspecifically), 0.011 between *B. acerifolia* and *B. convolvulacea*, to 0.087, between *B. mannii* and *B. convolvulacea* within the ingroup, and 0.203 between the outgroup and ingroup *(Datisca* and *B. acerifolia)*.

The mean base frequencies for taxa are:

A = 0.340
C = 0.141
G = 0.159
T = 0.360
(GC = 0.301).

b. Trees

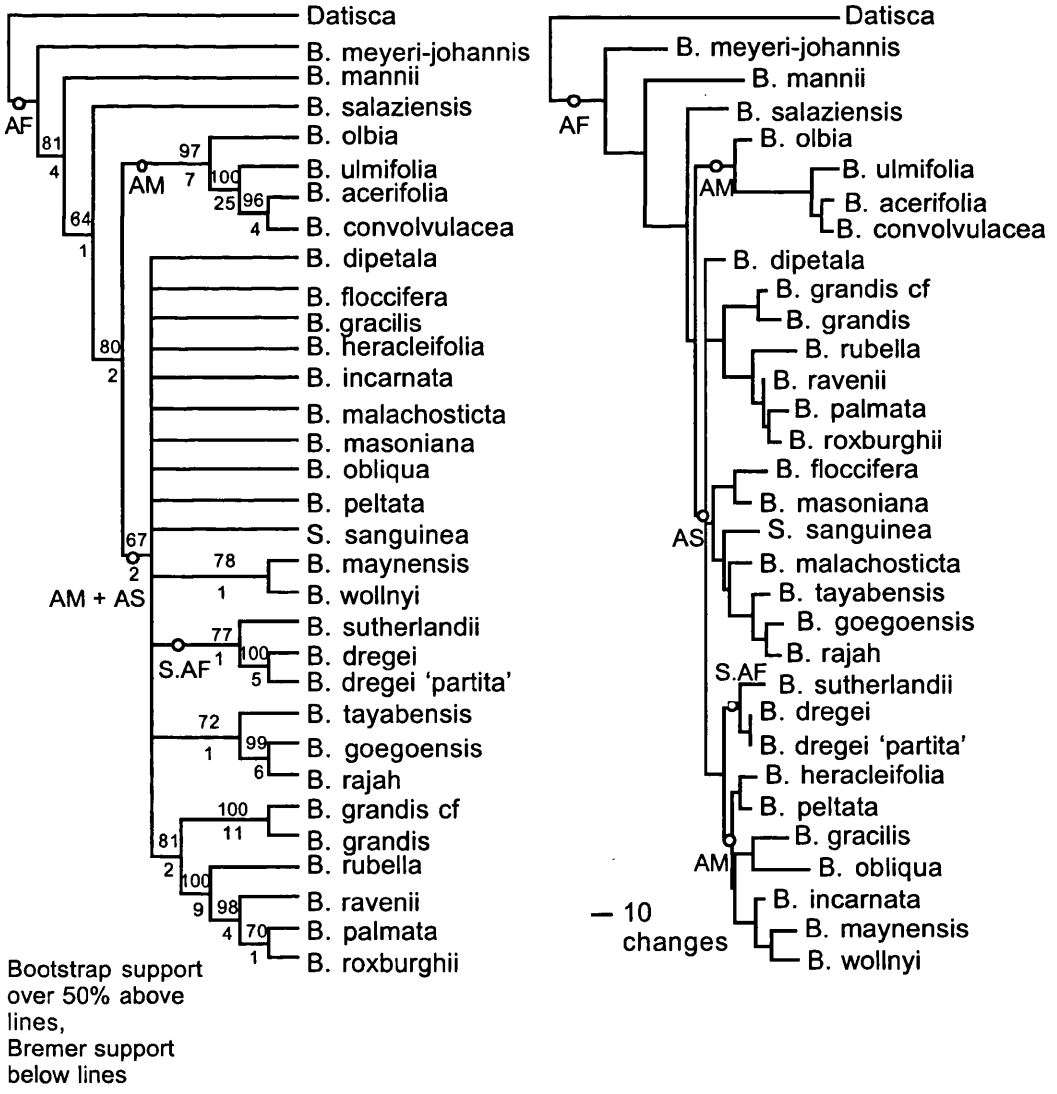
i. MP: There were186 MPTs found, length 807 steps, with a consistency index of 0.82 (0.63 excluding uninformative characters); retention index 0.68. Seventeen clades have over 50% bootstrap support, while there are 18 resolved nodes in the strict consensus tree. See Figure 9.1 for the strict consensus and for a phylogram.

From the topology presented in Figure 9.1, African taxa are basal in *Begonia*, with *B. salaziensis* as sister to an American/Asian clade. Most of the taxa are unresolved in the strict consensus, although one clade of six Asian taxa has 81% bootstrap support, and one of four American taxa has

97% bootstrap support.

Despite different exclusion sets and the removal of one taxon, the topology is consistent with Badcock's (1998) substitutions-only analysis (her Figure 3.4), although her analysis is slightly more resolved, with an American clade of *B. wollnyi*, *B. maynensis*, *B. peltata*, *B. heracleifolia*, *B. acutifolia*, *B. gracilis*, *B. incarnata* and *B. oacacana* sister to *B. sutherlandii* and *B. dregei*. Badcock's analysis with substitutions and coded indels (Figure 3.6 in Badcock, 1998) is again consistent with this strict consensus, but is considerably more resolved.

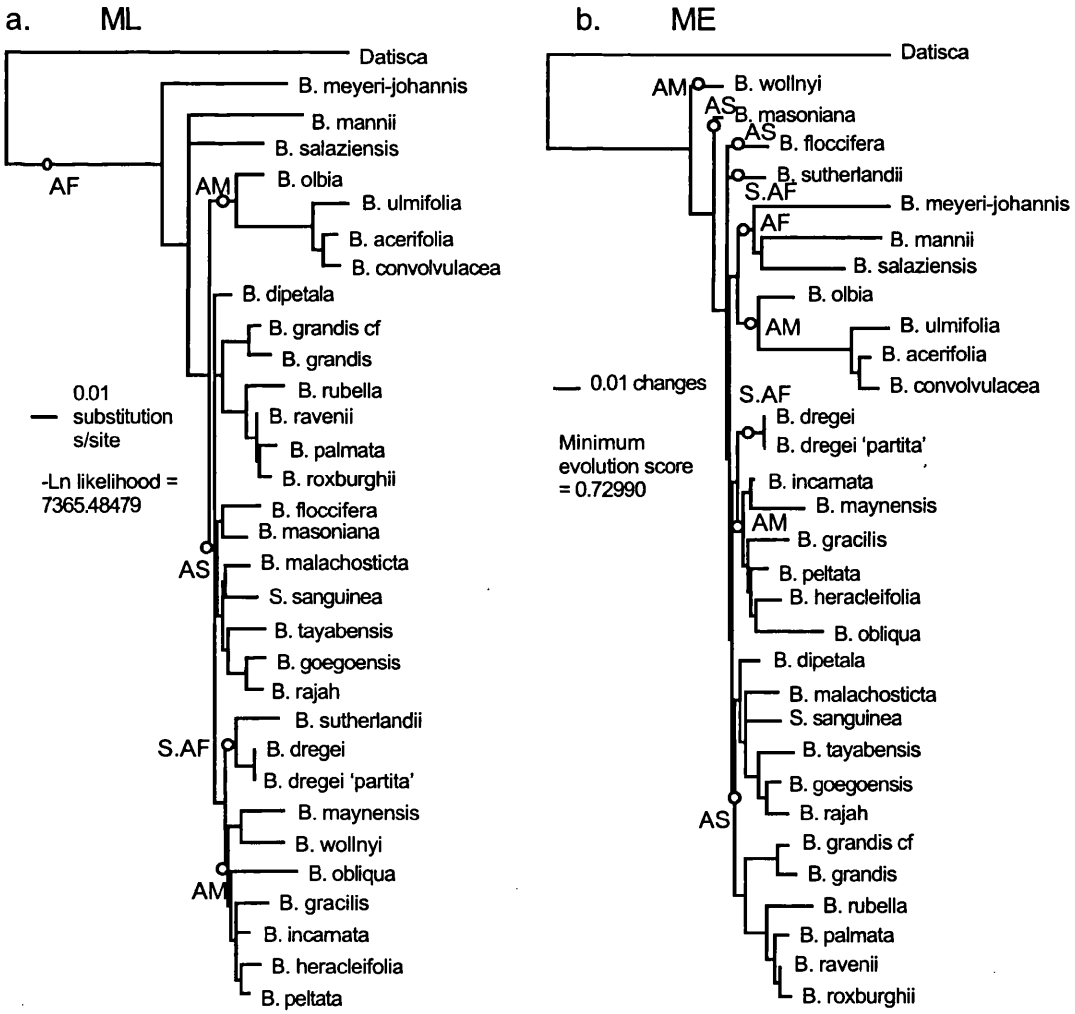
Figure 9.1 *trnC - trnD*, MP strict consensus of 186 MPTs and phylogram



ii. Maximum Likelihood: The assumed nucleotide frequencies are the mean base frequencies for the data set. Gamma distribution shape parameter α = 0.6895; the transition/transversion ratio is 0.699, κ = 1.667. There are 942 distinct patterns under the model.

All the clades which were resolved by MP are in the ML tree (Figure 9.2 a); both methods put *B. meyeri-johannis* as sister to the rest of *Begonia*. Neither Africa, America or Asia are monophyletic.

Figure 9.2: *trnC - trnD*, ML and ME trees.



iii. **Minimum evolution:** one tree was found, with minimum evolution score = 0.72990 (see Figure 9.2 b).

ME, while recovering several clades in common with MP and ML, produces very different basal relationships, with *B. wollnyi* (American), *B. masoniana* and *B. floccifera* (Asian) and *B. sutherlandii* (Southern African) basal to the rest of *Begonia*. The African taxa which are basal in MP and ML (*B. meyeri-johannis* and *B. salaziensis*, section *Mezieria*, and *B. mannii*, section *Tetraphila*) resolve as sister to an American clade in this tree.

9.3.2 ITS

a. Data: There were 632 characters excluded; 247 constant, 92 parsimony-uninformative and 183 parsimony-informative characters were included. Excluded characters are 1-183, 188, 200, 204, 211-217, 223-225, 230-249, 255-256, 266, 274-329, 340-366, 378, 383-384, 406-407, 415, 419-421, 426-428, 435-437, 444, 449-451, 460, 466-469, 475-483, 493-497, 503-507, 513-514, 539, 571, 577, 603, 606, 615, 649, 686, 688, 693-856, 886-901, 930-931, 944, 957-966, 983-984, 992-993, 1013-1014, 1018, 1023, 1029-1035, 1041-1053, 1064-1093, 1110-1114, 1121-1122 and 1137-1154 from the ITS matrix, see CD-ROM, i.e. the same exclusion set as in Chapters 5 and 7.

The PTP probability is 0.001; the skewedness statistic g1 is -0.993.

Uncorrected pairwise distances range from 0.008 (*B. convolvulacea* and *B. acerifolia*) (within-species values are slightly higher, 0.010 (*B. dregei* and *B. dregei* '*partita*') and 0.019 (*B. grandis* ssp *grandis* and *B. grandis* ssp *holostyla*)) to 0.224 between *B. floccifera* and *B. salaziensis*. Pairwise distances between the outgroup and ingroup range from 0.237 (*D. cannabina* and *B. mannii*) to 0.317 (*D. cannabina* and *B. floccifera*).

Mean base frequencies for taxa are:

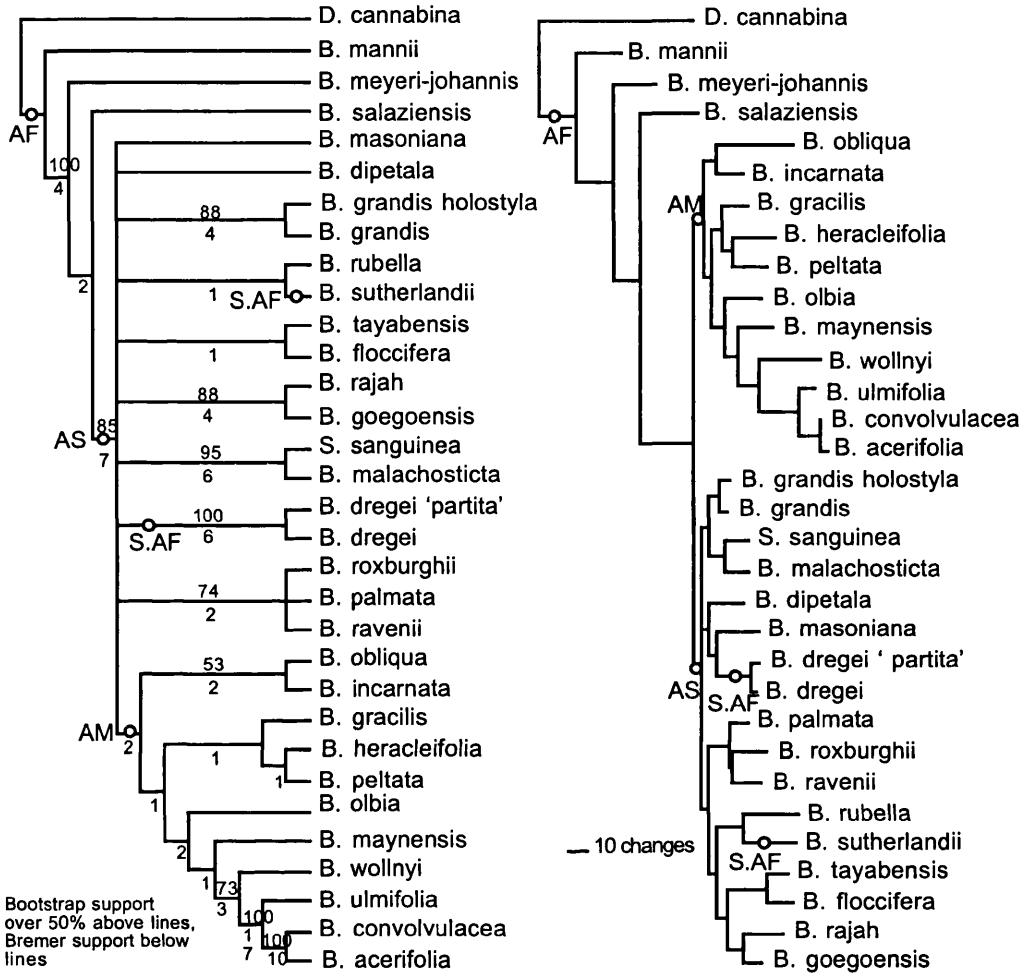
A = 0.209
C = 0.271
G = 0.306
T = 0.214
(GC = 0.577)

b. Trees: Sixteen MPTs were found, of length 862, consistency index 0.53 (0.45 excluding uninformative characters); retention index 0.46. Eleven clades have over 50% bootstrap support, while there are 21 resolved nodes in the strict consensus tree.

On the basis of this topology (Figure 9.3), African taxa are basal in *Begonia*, with *B. salaziensis* sister to a largely unresolved clade which includes American and Asian taxa, as well as the southern African taxa *B. sutherlandii* and *B. dregei*. However, from the phylogram, it can be seen that internal branches are generally short, particularly within the Asian/American clade.

There are some areas of conflict between the *trnC - trnD* strict consensus tree and this ITS strict consensus. The positions of *B. mannii* and *B. meyeri-johannis* are reversed; other changes are the position of *B. maynensis* and *B. wollnyi* (which are within a *B. olbia*/ *B. ulmifolia*/ *B. convolvulacea*/ *B. acerifolia* clade for ITS, but unresolved for *trnC - trnD*) and the separate positions of *B. sutherlandii* and *B. dregei* (which have 77% bootstrap support as a clade in the *trnC - trnD* analysis).

Figure 9.3: ITS, Strict consensus of 16 MPTs and phylogram.



9.3.3 Combined *trn*C - *trn*D and ITS analyses

a. Data: There were 292 characters excluded; of the remaining 2672 included characters: 1816 characters are constant (1569 from *trnC - trnD* and 247 from ITS), 461 variable characters are parsimony-uninformative (369 from *trnC - trnD* and 92 from ITS) and 371 are parsimony-informative (188 from *trnC - trnD* and 183 from ITS).

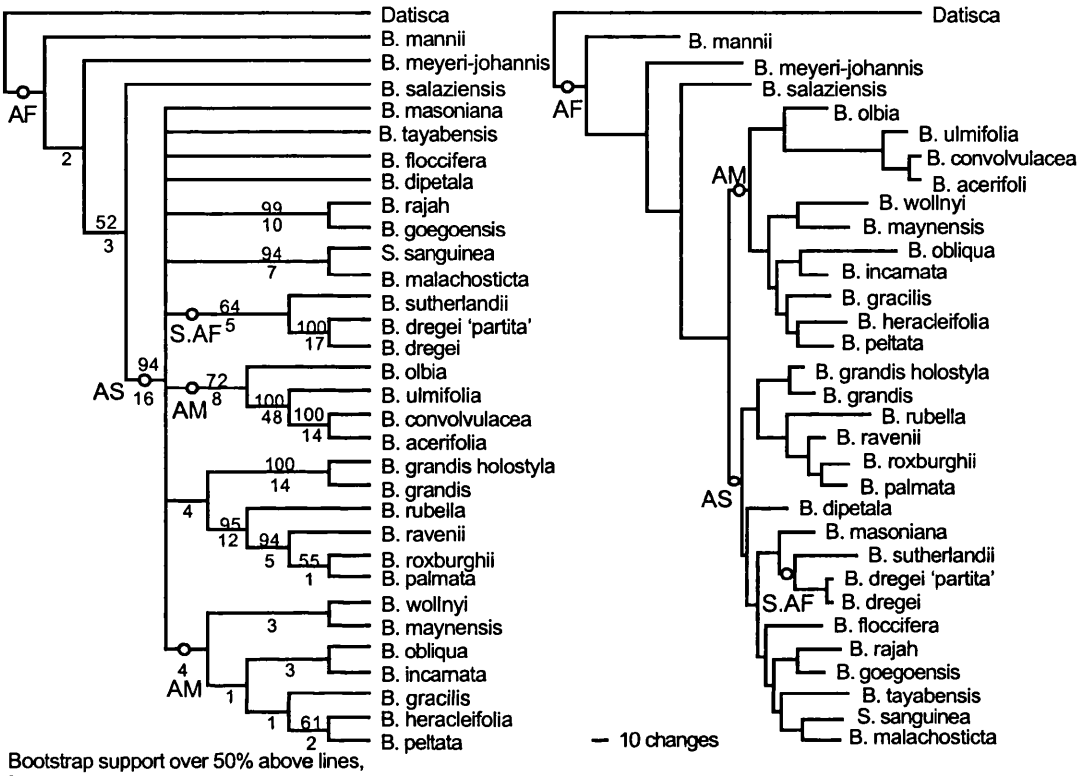
Mean base frequencies for taxa are:
A = 0.303
C = 0.178
G = 0.200
T = 0.319

The PTP probability is 0.001; the skewedness statistic g1 is -1.434.

b. Trees: Four MPTs were found, of length 1699, consistency index 0.65 (0.50 excluding uninformative characters) and retention index 0.52. Fourteen clades had bootstrap support of over 50%, and 22 nodes are resolved in the strict consensus tree.

Like the individual *trnC - trnD* and ITS analyses, this topology (Figure 9.4) shows African taxa basal to a largely unresolved Asian and American clade, which is sister to *B. salaziensis*. Within this largely unresolved clade, however, several smaller clades are resolved (one of southern African taxa, 64% bootstrap support; one of Asian taxa, no bootstrap support; and two of American taxa, one lacking support and one with 72% bootstrap support).

Figure 9.4: Combined *trnC - trnD* and ITS
Strict consensus of 4 MPTs and phylogram.



9.3.4 General Comments: The statistics for the MP trees produced from the three different data sets are summarised in Table 9.2

Table 9.2: Summary: statistics for MP analyses, three different data sets

Data set	No. inform. char.s	g1	PTP	No. MPTs	length	CI	CI ex uninform.	RI	nodes strict consens.	Nodes > 50% bootstrap
trnC - trnD	188	-1.364	0.002	186	807	0.82	0.63	0.68	18	17
ITS	183	-0.993	0.001	16	862	0.53	0.45	0.46	21	11
combined	371	-1.434	0.001	4	1699	0.65	0.5	0.52	22	14

Although the *trnC - trnD* analysis has the most clades with bootstrap support and higher consistency and retention indices, the ITS analysis has less MPTs and (perhaps in consequence) more nodes resolved in the strict consensus of those MPTs.

9.3.5 Gaps

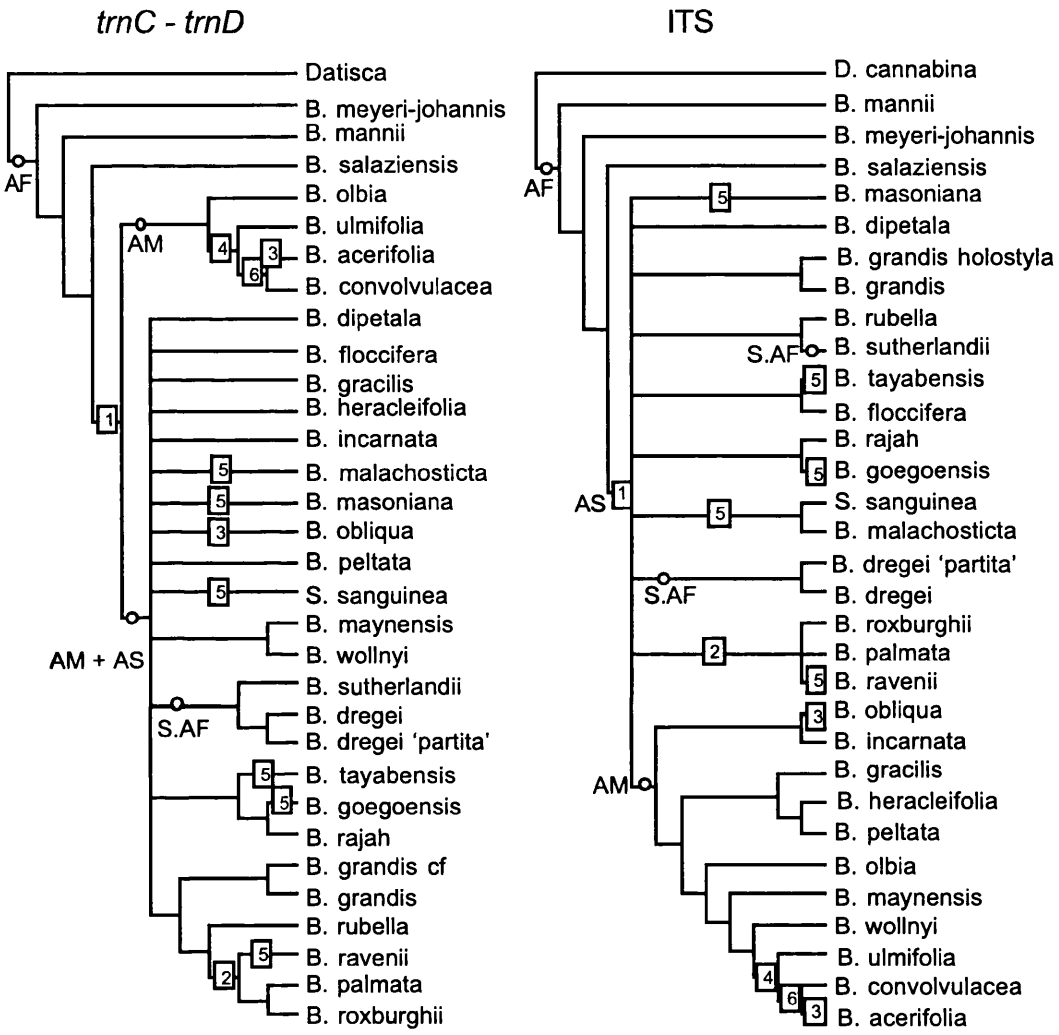
Although Badcock (1998) included a detailed matrix of gaps in *trnC - trnD*, this has been greatly simplified here. Only informative indels with identical sequence at the 5' and 3' ends have been coded (Table 9.3):

Table 9.3 Unambiguous gaps in the *trnC - trnD* alignment

INDEL	SITE	TAXA
C1	349-357 S	*Datisca*; *B. meyeri-johannis*, *B. salaziensis*, *B. mannii*
C2	436-450 S	*B. palmata*, *B. ravenii*, *B. roxburghii* (inapplicable in *B. incarnata*)
C3	616-617 S	*B. obliqua*, *B. acerifolia*
C4	1341-1847 G	*B. acerifolia*, *B. convolvulacea*, *B. ulmifolia*
C5	1411-1423 S	*B. goegoensis*, *B. malachosticta*, *B. masoniana*, *B. ravenii*, *B. tayabensis*, *Symbegonia* (inapplicable in *B. acerifolia*, *B. convolvulacea*, *B. ulmifolia*, *B. meyeri-johannis*)
C6	2137-2143 G	*B. acerifolia*, *B. convolvulacea* (inapplicable in *B. dregei*, *B. dregei* '*partita*', *B. goegoensis*, *B. masoniana*, *B. meyeri-johannis*, *B. olbia*, *B. palmata*, *B. rajah*, *B. salaziensis*, *B. sutherlandii*, *B. wollnyi*)

These indels were then mapped onto the MP strict consensus trees for *trnC - trnD* (Figure 9.1) and for ITS (Figure 9.3), and are presented here as Figure 9.5.

Figure 9.5: *trnC - trnD* indels mapped onto *trnC - trnD* and ITS strict consensus trees



Indel congruence: Indels 1, 2, 4 and 6 map onto both trees without homoplasy. Indels 3 and 5 are homoplastic on both the ITS and the *trnC - trnD* topologies. Part of the difficulty is the lack of resolution (a consequence of mapping onto a consensus tree rather than one of the individual MPTs). However, indel 3 would require two (i.e. its maximum number of) changes even were the backbones of these trees fully resolved. Indel 5 requires a minimum of one loss (*B. rajah*) and one independent gain (*B. ravenii*) in the *trnC - trnD* topology, and at least one more change in the ITS topology (to account for *B. floccifera*).

9.3.6 Molecular Evolution:

Figure. 9.6 *trnC - trnD*: Number of steps per position for one MPT (grey shading on the x-axis represents positions excluded from the analysis).

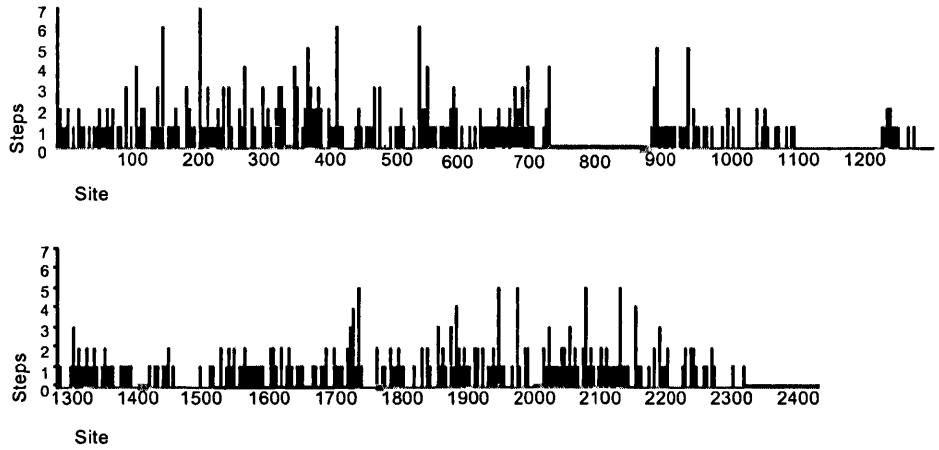
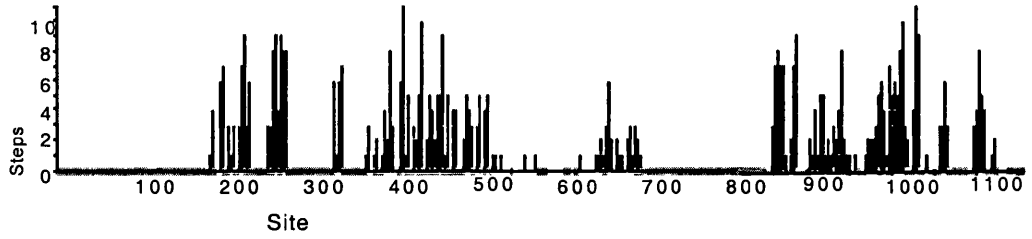
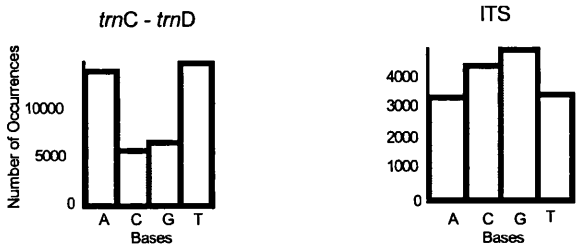


Figure 9.7: ITS: Number of steps per position for one MPT (grey shading on the x axis represents positions excluded from the analysis).



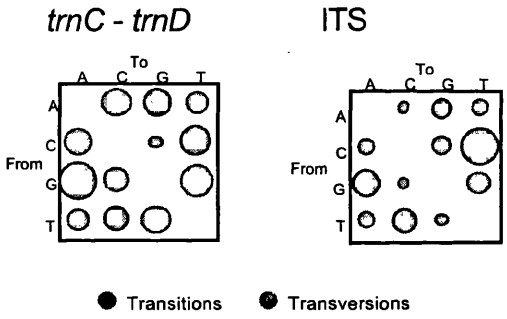
The ITS matrix (Figure 9.7) includes characters which have more steps (11) then the *trnC - trnD* matrix (Figure 9.6) (which has a maximum of 7 steps). There are many more positions in *trnC - trnD* which have one step on the tree (i.e. fit perfectly or are uninformative) than there are in the ITS matrix. While the *trnC - trnD* data set is considerably longer (more than twice the length of ITS) it has fewer changes per variable site; less data from *trnC - trnD* was excluded due to alignment ambiguities.

Figure 9.8: Base composition of the different matrices



As was noted by Badcock (1998), and can be seen in Figure 9.8, *trn*C - *trn*D is appreciably AT rich. ITS has a less skewed base composition, although it is GC rich.

Figure 9.9: Proportions of transitions and transversions in the different matrices, measured over one MPT.



As can be seen in Figure 9.9, *trnC - trnD* has a larger proportion of transversions to transitions than ITS has.

9.4 Discussion

All the topologies agree that African taxa (*B. mannii*, *B. meyeri-johannis* and *B. salaziensis)* are paraphyletic, including a monophyletic clade of Asian, American and southern African taxa.

Gaps which were coded from the *trnC* - *trnD* matrix mapped onto the trees with varying degrees of homoplasy. The four gaps which mapped on without homoplasy mapped equally well onto ITS and *trnC* - *trnD* topologies, while the two gaps which were homoplasious on one topology were also homoplasious on the other. Regarding the two homoplasious gaps, gap C3 is very short, only two base pairs shared by two American taxa; gap C5 is 12 base pairs shared by six Asian taxa.

Although analysis of the combined data set produces topologies with better tree statistics than analysis of the ITS data set does, and with more resolved nodes than the strict consensus from analysis of the *trnC* - *trnD* data set, it is uncertain whether both regions, from different genomes, track the same history. The question of reticulation in *Begonia* evolution will be dealt with in more detail in a further chapter.

9.5: Summary

Sequence data from the chloroplast region *trnC* - *trnD*, the nuclear ribosomal ITS region and a matrix consisting of both regions combined, for 32 taxa (one *Datisca*, one *Symbegonia* and 30 *Begonia* taxa), were analysed using MP. The *trnC* - *trnD* region was also analysed using ML and ME.

The *trnC* - *trnD* topology obtained using ML is fully congruent with the strict consensus tree produced using MP. However, ME produced some highly unconventional groupings; as implemented here it appears unsuitable for this data set.

MP analyses of both ITS and *trnC* - *trnD* support African taxa as basal in *Begonia*; however, *trnC* - *trnD* supports polyphyly/paraphyly of American taxa. The positions of these taxa are not resolved in the ITS tree produced here (although monophyly of American taxa is supported by earlier ITS analyses - see Figure 5.5, Figure 7.3). Partly because the consensus trees from all data sets are not very well resolved, there is no real evidence for conflict between the signal in these regions.

Mapping six indels from the *trnC* - *trnD* data matrix across both the *trnC* - *trnD* and the ITS MP strict consensus trees, four fit both topologies perfectly while two were homoplasious on both topologies; again, both regions appear to be showing similar signal.

10. Morphology

10.1 Introduction

One of the major problems facing systematists who are dealing with *Begonia* is that many of the morphological changes within the genus are continuous - most notably, shape and size do not give the sort of discrete characters useful for delimiting higher taxa. *Begonia* leaves, for example, occur in many shapes and sizes, and these seem to bear no relation to the evolutionary history of the group - strap-like leaves are found in the Madagascan *B. bogneri* and the American *B. herbacea*; compound dissected leaves are found in *B. hemsleyana* (Yunnan, China) and the American *B. luxurians*. Despite the leaf similarities, close relationships of these species have never been suggested - the floral morphology would not warrant it.

Arends (1992, pp. 82-85) observes that the number of styles on some *Tetraphila* species can vary within the same collection locality, and even within the same inflorescence (*B. longipetiolata*). Similarly, the plant *B.* cf. *rubella* in cultivation at RBGE produces inflorescences with both two and three locular fruits. These are the sort of characters traditionally highly weighted in *Begonia* classification, but frequent state reversals within the genus, in the light of such plasticity, cannot be ruled out.

High homoplasy levels have led to suggestions that phylogeny reconstruction using morphology is not viable for plant taxa (Cronk, pers. comm., 2000). Selection pressures for certain morphological features can result in the recurrent evolution of similar morphologies in independent lineages. Given the potential for non-hierarchical patterns of character distribution (e.g. where environment influences phenotype and where mosaic evolution occurs) and problems with character delimitation (e.g. for shapes or numerical ranges) molecular data has come to be perceived as better for phylogenetic purposes.

However, morphological characters are of interest for a number of reasons:

1. phylogeny reconstruction (where sequence data are unavailable, e.g. for fossil taxa);
2. as clade markers for another data set;
3. to trace character evolution over a tree produced largely from another data set, e.g. molecular, e.g. to test hypotheses of homology.

If the purpose of the morphological data gathering is not to reconstruct phylogeny, then the homology of characters is of less importance. Things like leaf shape are unlikely to be quantifiable into homologous character states across a family, but may still be interesting to trace across a phylogeny. Some clades may have a predisposition to some shapes. Many 'problem' characters, such as stomatal density, leaf shape and habit, tie in with ecology. Other problem characters include those with overlapping ranges - like flower number per inflorescence, or anther number per flower. Although these may be deconstructable into characters which could show trends across a phylogeny, it is difficult to see how ranges of numbers can be homologes.

10.1.1 Previous morphological studies: The phylogenetic utility of morphological characters in *Begonia* has been discussed in Badcock (1998) and in Tebbitt (1997). Badcock coded 70 unordered morphological characters for 86 species. She suggested that there were several equally probably evolutionary hypotheses for the data set, thus it was 'difficult to draw many strong conclusions about the evolution of *Begonia*'. However, she found the results useful within sections and between closely related sections.

Tebbitt (1997) coded 40 morphological and anatomical characters for 56 species, which include a detailed study of anther endothecial wall patterns (expanded and published by Tebbitt & McIver, 1999). Low consistency indices, retention indices and resolution in the trees he obtained led him to conclude that there was a large amount of homoplasy, there were few synapomorphies supporting clades, and there was a degree of character conflict within his data.

Doorenbos, Sosef and de Wilde (1998) also examined morphological evolution within *Begonia*, using 63 characters, and all 63 sections of *Begonia* as terminal units. They used this data to produce a phenetic classification of sections (noting that cladistics would not be applicable as some of the sections were paraphyletic or polyphyletic, although one could also question the applicability of phenetics in this situation). They describe the fit of characters to their phenogram as "poor". They were surprised to find that African and Asian sections are dispersed across their tree, while American sections are relatively clustered.

Sosef (1994) conducted a cladistic morphological analysis to assess the monophyly of the African sections *Loasibegonia* and *Scutobegonia*. He chose characters according to two constraints, that they were (preferably) not polymorphic within species, and that both character states occur within more than a single species. Sosef identified 132 characters: 76 macromorphological, 32 from leaf anatomy, 17 from ovary and style anatomy and 7 from seed micromorphology, for 43 taxa. Sosef rejected the most parsimonious cladogram for his data in favour of a less parsimonious tree, which he felt to be closer to the 'truth'. From this tree, Sosef (1994) was able to support the monophyly of sections *Loasibegonia* and *Scutobegonia*.

Morphology is also discussed by Arends (1992), with particular reference to species in the African section *Tetraphila* A. DC.

The rest of Section 10.1 gives a brief overview of some of the morphological diversity within the genus *Begonia*.

10.1.2 Vegetative morphology

10.1.2.1 Perenniating organs: Many *Begonia* species are rhizomatous, e.g. *B. masoniana*, *B. violifolia*, *B. letouzeyi*. Most rhizomatous *Begonia* species are acaulescent, although upright stems can form from rhizomes, particularly when the plant is in flower (e.g. *B. josephii*). Another class of *Begonia* is described in horticultural circles as 'cane'; species possess upright stems which can grow to over one metre. This group includes plants like *B. maculata* Raddi and *B. longifolia*. Some more woody species, like *Begonia luxurians*, can grow to several metres tall. Woody stems are found in several Old and New World *Begonia* species; the wood anatomy is similar to that of *Datisca* (Carlquist, 1985).

Some *Begonia* are tuberous, like *B. boliviensis* and *B. grandis*. Not all tubers are homologous: *B. grandis* has stem tubers, swollen storage roots and axillary tubercles, while American species like *B. boliviensis* have root tubers (Badcock, 1998). *B. socotrana* possesses bulbils, which are produced from axile stem nodes (Irmscher, 1925). The difference between bulbils and tubercles is that in bulbils, the storage organs are reduced leaves, while in tubercils, the stem is the storage organ, and has rudimentary leaves around it. Bulbils and tubers are typically found in species which occur in seasonally

water-limited environments, where the fleshy above-ground parts of the plants die back annually. Another adaptation to this sort of environment is deciduousness, which can occur in some of the thick-stemmed taxa like *B. dregei* and *B. wollnyi*.

B. dregei and related species have a caudex, which is a swollen woody stem base. This may be an adaptation to fire. The caudex only forms on individuals grown from seed, not on plants propagated by cuttings (and so will not be seen on all individuals held in Botanic Garden collections), and is probably derived from the hypocotyl (Hughes, pers. comm., 2000).

10.1.2.2 Stipules: Stipules provide 'characters of considerable diagnostic value to the taxonomist' (Foster & Gifford, 1959, p. 447). Begoniaceae leaves are always stipulate (Doorenbos, Sosef & de Wilde, 1998). These are attached to the node, free from the base of the petiole, and are always formed before the leaves are produced (Arends, 1992). Minute white hair-like stipules are also present at the base of young leaves of *Datisca cannabina,* although Chant (in Heywood, 1978) states that the leaves of Datiscaceae are without stipules.

Burt-Utley describes *Begonia* stipules as 'caducous' or 'fugacious' depending on whether they fall before a new leaf starts expanding or as the leaf is maturing (Burt-Utley, 1985, p. 14). Each of a pair of stipules may be slightly asymmetric, with both members of the pair mirror images. There are distinct differences between the inner and outer stipules in some species, e.g. *B. imperialis*.

The stipules of some species have hairs on the outer surface, e.g. *B. palmata*; other taxa have glabrous stipules, e.g. *B. glabra.* In some taxa, the margins of the stipules are dentate to finely fringed, e.g. *B. sutherlandii*, while in most species, they are entire (e.g. *B. palmata*).

Some species possess a very distinct rib along the back of the stipule, like a keel, while others have little more than a slight thickening over the main vein. Within stipule pairs in sect. *Gireoudia*, one member has the keel excurrent apically, while in the other it is excurrent subapically, and any indumentum tends to only be found on the lamina of the outer member of the stipule pair (Burt-Utley, 1985).

Several taxa have a tooth which projects from the back of the main nerve near the tip of the stipule. This spur may be an extension of the keel; however, in many taxa the spur is found even where a keel is not apparent.

"[T]he major role of most stipules seems to be the protection of young developing leaves" (Foster & Gifford, 1959). In *B. jamesoniana* the young stipules are about one cell thick over most of their surface. Enclosed inside the stipules, the environment seems very damp. After the enfolded leaf has emerged, the stipules soon change from green to brown, losing all moisture from their cells and becoming papery. It is possible that keeping the growing leaf from drying out is one function of the stipules; furthermore, hairy or strongly keeled stipules may offer greater protection from browsers.

10.1.2.3 Leaves: The leaves of *Begonia* are occasionally cauline, e.g. *B. herbacea* (Figure 10.19 e) but usually petiolate. In some species there is a ring of hairs or trichomes at the top of the petiole, e.g. the African species, *B. johnstonii* (Figure 10.18 b), the Asian species *B. tayabensis* (Figure 10.18 e) and the American species *B. manicata*. The homology of these is hard to ascertain, as the hairs may be of different colours, and in some taxa the bases of clusters of hairs are fused together, forming scales. Furthermore, such rings may have an adaptational advantage in reducing access to the lamina to non-flying insects and have arisen several times independently.

Peltate leaves are found in many sections of *Begonia*, in plants which are not in any other way similar, e.g. *B. peltata* (American), *B. tayabensis* (Asian; Figure 10.18 e) and *B. socotrana* (Socotran). However, some individuals of *B. socotrana* form non-peltate leaves on flowering stems (pers. obs., 2000; pers. comm., Hughes, 2000). In some taxa, like those mentioned, the point of insertion of the petiole is more or less central to the lamina; many species in section *Loasibegonia* have highly asymmetric insertion.

There are several reasons why it may be advantageous to be peltate, including:

1. a peltate leaf needs less lignification to support its weight;
2. a peltate leaf should provide the most efficient nerve arrangement to transport water and nutrients (Burt-Utley, 1985);
3. with the functional separation of the petiole and the lamina, leaf (lamina)

orientation is not directly dependent on petiole orientation (pers. comm., Cronk, 2000);

4. the lamina may form a water-catching cup (pers. comm., Cronk, 2000);
5. a peltate leaf may make more efficient use of its leaf meristems, as expansion can occur at equal rates all round the edge of the leaf rather than being mainly limited to the tip (pers. comm., Cronk, 2000).

Only the third point, leaf orientation, applies to the leaf arrangement of some species in sections *Loasibegonia* and *Scutobegonia* (e.g. *B. letouzeyi*), where the point of petiolar insertion is very close to the leaf margin. It may be possible to compare the habit of some of these taxa, e.g. *B. dewildei* (section *Scutobegonia*) with *B. herbacea*, which is non-peltate but has similar overall leaf-shape.

Another possibility is that there is no advantage conferred to plants in these two sections from being peltate, but it is more difficult to switch back from peltate to basifixed.

The lamina is usually asymmetric at the base (e.g. *B. lyman-smithii*, Figure 10.18 c), although it can be difficult to see the asymmetry in species like *B. bogneri*, which has long, strappy leaves, and in some peltate species which have circular (e.g. *B. socotrana*) or pointed (e.g. *B. tayabensis,* Figure 10.18 e) laminas. Other peltate species, e.g. *B. sericoneura* (Figure 10.18 d), may show an asymmetric lamina 'base', despite the point of petiolar insertion being elsewhere.

Most *Begonia* species have simple leaves, although in some species they are highly dissected, like *B. aspleniifolia* (Figure 10.18 a) in Africa, and *B. incisa* in Asia. Other species have truly compound leaves, like *B. luxurians* and *B. theimei* in America, and *B. hemsleyana* in Asia.

The lamina can either be flat, e.g. in species in section *Tetraphila*, or be bullate (raised into many cones on the upper surface, visible as pits on the leaf underside), like the leaves of *B. masoniana* or *B. imperialis*.

a. Leaf colour: Burt-Utley (1985) does not consider leaf patternation to be taxonomically useful within section *Gireoudia*; she has seen populations with

maculate and plain leaved forms growing together, and also suggests that the expression of maculation is environmentally determined. Likewise, plants of *B. palmata* which had plain green leaves when collected from shady under-forest sites in Yunnan developed coloured leaf patterns in cultivation at the Royal Botanic Garden, Edinburgh and Glasgow Botanic Garden. However, some patterning, e.g. that on the leaves of *B. brevirimosa*, *B. serratipetala* and *B.* cf. *brevirimosa* (all in section *Petermannia*), may be homologous, as the leaves are similar in texture, colour and maculation. Leaf pattern is consistent across several accessions of *B. serratipetala*.

b. Leaf venation: Leaves may have palmate (see Figure 10.18 b), palmate/pinnate, or pinnate venation (see Figure 10.19.3 c). Further, the texture of the veins can differ greatly, from raised interconnected networks to barely visible, vanishing veins. There are almost always hairs along the veins on the leaf underside.

c. Stomata: Stomata are found either singly (e.g. section *Coelocentrum*) or in groups (e.g. section *Lepsia*). Stomatal density and group size have been found to vary with environmental factors in *Begonia* (Hoover, 1986) and so are not reliable for phylogeny reconstruction.

10.1.2.4 Hairs: *Begonia* species have glandular and non-glandular hairs, which, on close inspection, can be found on almost every organ of almost every species. Long branching hairs (which give an overall impression of 'fuzziness') are found on certain American species e.g. *B. egregia*. Stellate trichomes are found on many African species, particularly within section *Tetraphila*, like *B. mannii*.

Burt-Utley (1985) found trichome morphology and density useful for species delimitation in section *Gireoudia*. She recognised two basic classes of trichomes - glandular (or capitulate, uniserate) and multiseriate (including villi and 'whiplash' trichomes) (although "in some species these trichomes [villi] often become uniseriate distally" (Burt-Utley, 1985, p. 13)). Burt-Utley (1985) suggests that, given the wide distribution of villi within *Begonia*, they (villi) may represent the primitive condition from which lacerate scales and whiplash trichomes arose. She suggests that scales, which are found in a number of sections, arose repeatedly. Shui, Li and Huang (1999) examined the hairs on 46 species from the Chinese province of Yunnan. They found that epidermal

and hair characters are useful at the specific and varietal levels, but not in distinguishing sections.

Hair presence or absence can be striking on different organs - e.g. some taxa have very hairy leaves (e.g. *B. versicolor*), while the leaves of others are glabrous (e.g. *B. glabra*). However, glandular trichomes are frequently present on both surfaces of the leaf primordium but are not found on mature leaves (McLellan, 1990). Thus presence or absence of hairs is not a simple character, and its determination may require electronmicrograph studies of very young leaves.

10.1.3 Sexual characters

10.1.3.1 Sexual separation and inflorescence architecture: *Begonia* show a wide range of inflorescence structure and of sexual separation. Flowers are (almost always) monoecious[7]; plants, dioecious (e.g. *B. menyangensis*, *B. roxburghii,* Figure 10.19.2 d), protandrous (e.g. *B. chloroneura, B. oxyphylla*) or (rarely) protogynous (e.g. *B. brevirimosa*). Inflorescences are rarely unisexual (e.g. *B. herbacea,* Figure 10.19.1 e) or, more usually, bisexual. Inflorescences are rarely racemes (e.g. section *Petermannia, Symbegonia* - see Figure 10.19.2 c), but usually cymes (e.g. *B. diadema*, Figure 10.19.2 a; *B. luxurians*); a few species have monochasial rather than dichasial branching, e.g. in section *Loasibegonia*, where species frequently have one female and two male flowers per inflorescence. Bisexual inflorescences may have two female flowers and one male in each terminal dichasium, or all the female flowers may be found at the base of the inflorescence (e.g. *B. brevirimosa*) or otherwise separate from the males. Inflorescence architecture is discussed in depth by Goulet, Barabe and Brouillet (1994).

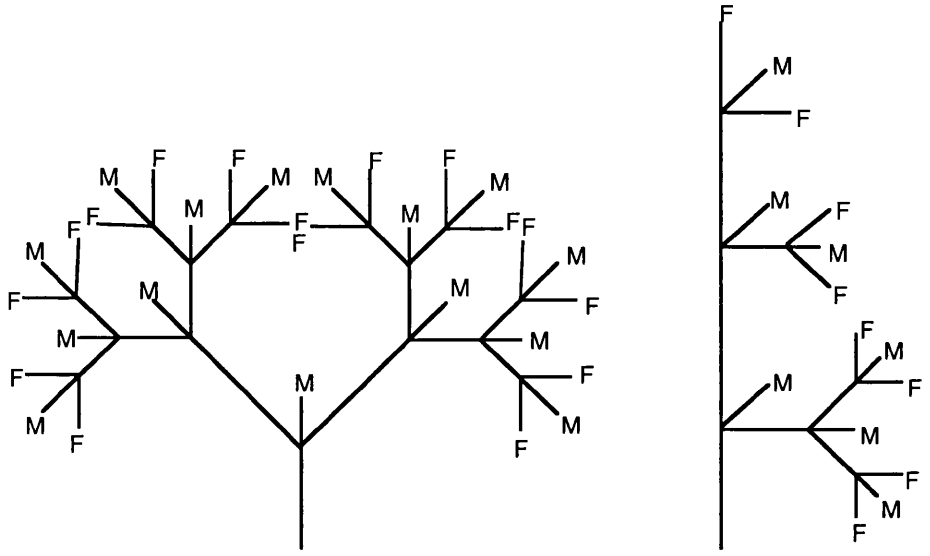
[7] One taxon in cultivation at RGBE (*B.* sp. nov., Philippine) has, for the last 3 years, produced functional female flowers with a few anthers and male flowers with 3 fully developed stigmas (but no ovary). This sort of abberation can be brought on by cultivation, e.g. in *B. samhahensis*.

While the actual numbers of male and female flowers per bisexual inflorescence are often similar, their temporal distribution is not. In protandrous inflorescences (e.g. *B. oxyphylla*), the male flower at the basal branching point is the first to mature; there is then a progression of male flowers maturing at branching points progressively distal to the base. The female flowers are last to open, and may not do so until all the males in the inflorescence have dropped. An individual plant may contain several inflorescences at various stages of maturity. It is often the case that the plant is functionally male for several weeks before any females are mature; furthermore, nearing the end of its flowering period, the plant may retain only female flowers. It is not uncommon to find individuals which look superficially dioecious. This will tend to promote outcrossing.

Having a deciduous (male) flower central to each dichotomy within a dichasial inflorescence makes some sense for resource allocation in large inflorescences, as the male is unlikely to be a major sink for resources after anthesis; a female after fruit set could compete with the rest of the inflorescence. Often in very large inflorescences, e.g. *B. oxyphylla*, *B. luxurians*, the (terminal) females do not develop until most (or all) of the male flowers have fallen.

10.1.3.2 Inflorescence size: The number of flowers per inflorescence range from one (e.g. the female inflorescence for *B. herbacea*) to over 1000 (e.g. *B. luxurians*). Even where the basic structure is dichasial (as is most commonly the case), inflorescences may have symmetrical (e.g. *B. luxurians*; *B. diadema,* Figure 10.19.2 a) or asymmetrical (e.g. *B. heracleifolia,* Figure 10.19.1 b; *B. theimei*, Figure 10.19.1 a) branching (see Figure 10.1).

Figure 10.1: Symmetric and asymmetric inflorescence structure



A rough estimate of the number of flowers in dichotomous inflorescences can be obtained from the number of dichotomies (branching points), as follows:

Table 10.1: Number of flowers in different sized inflorescences

No. branching points	No. male flowers	No. female flowers	Total no. flowers
1	1	2	3
2	3	4	7
3	7	8	15
4	15	16	31
5	31	32	63
6	63	64	127
7	127	128	255
8	255	256	511
9	511	512	1023
10	1023	1024	2047

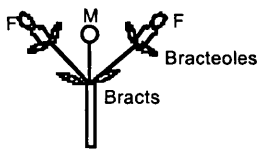
Naturally, the numbers of male and female flowers rely on a 'typical' inflorescence structure where males are central in dichasia. In some taxa, the basal few branching points lack this central flower (e.g. *B. luxurians*). This has only been seen in taxa with large inflorescences (over four branching points), and will slightly reduce the actual numbers of male flowers in these inflorescences. Further, often not all the female flowers develop.

10.1.3.3 Bracts: In some taxa the basal pair of bracts are large and enclose the entire immature inflorescence (e.g. *B. poculifera*; *B. ampla,* Figure 10.20 g), while in other species the bracts are not obvious (e.g. *B. oxyloba* Welw. ex Hook.f.). Bracts are often deciduous (although are reportedly persistent even on the fruit of species from section *Squamibegonia*, e.g. *B. poculifera*), and shapes, sizes and colours vary between species. Bract size and shape also changes between different nodes on an inflorescence (Burt-Utley, 1985) thus complicating the use of any bract characters cladistically. It may be unreasonable to assume homology between bracts from different species based on position without a very clear understanding of inflorescence branching patterns. Bract colour can also vary within species (commonly green to deep pink) and Burt-Utley (1985, p. 21) has found "more intense coloration often developing in bracts on inflorescences in exposed locations".

10.1.3.4 Bracteoles: The staminate flowers in section *Gireoudia* are ebracteolate, and in over half the species in the section, the female flowers are also ebracteolate (Burt-Utley, 1985). However, other *Gireoudia* species have rudimentary or minute bracteoles on occasional flowers. Because the female flowers are terminal and solitary, these bracteoles may be homologous to the bracts which enclose the dichasium (i.e. indicative of an unformed dichasium). Where there are pairs of well developed bracteoles at the base of the ovary, these may be inserted directly below the ovary (e.g. *B. convolvulacea*, *B. peltata*), or at some distance down the pedicel. Variations in shape appear to have little taxonomic significance (Burt-Utley, 1985), but the presence or absence of bracteoles "can be useful in distinguishing among morphologically otherwise similar taxa and in evaluating putative hybrids" (Burt-Utley, 1985, p. 22). Similar bracteoles occur, either in pairs or in threes, on many other taxa of *Begonia*, including *B. annobonensis* and *B. cubensis* (two bracteoles) and *B. fissistyla* (three bracteoles).

See Figure 10.2 for the difference between bracts and bracteoles; both can be seen on the inflorescence of *B. hercleifolia* in Figure 10.19.1 b.

Figure 10.2: Bracts and bracteoles



10.1.3.5 Flowers

a. Tepal colour: Flower colour in *Begonia* is most commonly white (e.g. *B. involucrata*, Figure 10.19.1 c) and/or pink (e.g. *B. socotrana*, Figure 10.19.3 a) (many species possess both forms, e.g. *B. grandis*). Yellow flowers are predominantly found in African taxa (e.g. *B. letouzeyi*, Figure 10.20 d), while red (*Symbegonia sanguinea*, Figure 10.19.2 c; *B. fuchsioides*) and orange (*B. oxysperma*, *B. boliviensis*) are found in some Asian and American taxa (respectively). Most taxa have only one colour in the flower, but some African (e.g. *B. ampla*, Figure 10.20 f; *B. aspleniifolia*) and American (e.g. *B. solananthera*) species have pink markings on otherwise white or yellow tepals. These markings can be asymmetric, appearing strongest on the lower tepal. In section *Platycentrum*, it is not uncommon for the outer and inner whorls of tepals (particularly in the male flower) to be different colours (e.g. the unidentified *B.* species, Figure 10.19.2 b).

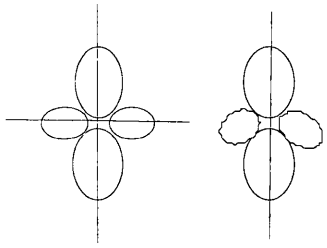
Tepal colour is reported to vary between white to dark pink, in the section *Gireoudia*, according to species, population and light levels (i.e. genetic and environmental factors) (Burt-Utley, 1985).

b. Stigma and anther colour: Yellow colour and strong ultraviolet absorption in stigmas and anthers was found in all insect-pollinated *Begonia* species examined by Schemske, Agren and le Corff (1996). They suggest that this implies mimicry of the anthers by the stigmas ('similarity of nonhomologous organs'). They go on to say: "[b]ecause a phylogeny of *Begonia* is not available, we do not know if characters such as stigma colour and uv absorption in female flowers represent the ancestral condition, or have evolved to increase the resemblance of female to male flowers" (Schemske, Agren & le Corff, 1996, p. 313).

c. Tepals: The commonest tepal numbers for male flowers are two (e.g. *B. brevirimosa*; *B. letouzeyi,* Figure 10.20 d; B. ampla, Figure 10.20 f, g; *B. herbacea*) and four (e.g. *B. luxurians*; *B. handelii,* Figure 10.20 b; *B. loranthoides*, Figure 10.20 e; *B. socotrana*, Figure 10.19.3 e). I have not seen plants with other numbers, although they are reported in the literature. In four-tepalled flowers, two tepals are usually smaller. The four tepals of four-tepalled flowers are usually arranged with two planes of symmetry, although occasionally the smaller pair point downwards, giving bilateral symmetry (see

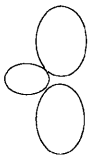
Figure 10.4; see also Figure 10.20 e, *B. loranthoides*).

Figure 10.3: tepal symmetry planes in male flowers



Female tepal number is more variable: two (e.g. *B. oxyloba*, *B. prismatocarpa*), three (e.g. *B. amphioxis*, *B. fallax*, *B. masoniana, B. herbacea,* Figure 10.21 g), four (e.g. *B. molleri,* Figure 10.21 f), five (e.g. *B. brevirimosa*, Figure 10.21 b; *B. listada*; *B. palmata*), six (e.g. *B. socotrana*, Figure 10.19.3 a; *B. crassirostris*) and occasionally more. Three-tepalled flowers may have strong bilateral symmetry, with two larger tepals arranged opposite each other, and one smaller tepal at 90° to them (e.g. *B. masoniana*) - see Figure 10.4.

Figure 10.4: Tepal arrangement in *B. masoniana* female flowers



Several authors have distinguished two whorls of tepals in the male flower into sepals and petals (e.g. de Candolle, 1859, 1864; Irmscher, 1925). Endress (1994) described the difference between sepals and petals being that sepals have broad bases with three vascular traces, and petals have narrow bases with just one vascular trace. All *Begonia* species are thought to have two sepals in the male flower (Badcock, 1998), therefore flowers with two tepals lack petals. It has proved less easy to distinguish the tepals of the female flower in such a way. Barabe (1980) looked at the vascularisation of the tepals in pistillate flowers of *B. handelii*, and found that the flower has two whorls of perianth parts, differentiated into calyx and corolla. However, this differentiation of petals and sepals in the female flower has not been widely followed, given perhaps the greater variation in tepal number in the female.

Tepal fusion is uncommon in *Begonia*, although it characterises *Symbegonia*. The male flowers of the species grown by Glasgow Botanic Garden have two tepals which are fused to a degree - in *S. sanguinea* they are fused along most

of their edge (this can be seen in Figure 10.19.2 c), while in the *Symbegonia* species accessioned 004 137 91 (GL) they are fused very shortly. *B. brevirimosa* similarly has two shortly fused tepals in the male flower. The female flowers in *Symbegonia* have five tepals fused into a long tube, with only the tips free (and pointed). This affords some evidence that the five tepals in the female *Symbegonia* flowers all represent the same organ (i.e. petals or sepals), because they are able to fuse; male flowers with tepal fusion all have two tepals, which may represent either sepals or petals (and are generally interpreted as sepals).

Female flowers of species in section *Squamibegonia,* e.g. *B. ampla*, *B. poculifera*, have a perianth tube between the top of the ovary and the tepals.

Some other species of *Begonia* show some tepal fusion - *B. chloroneura*, for example, has partial fusion of two of the five tepals of the female flower (or, alternatively, four tepals, one deeply divided). This is also seen in *B. tayabensis*. *B. brevirimosa* often produces flowers with tepal fusion in the female flowers, e.g. Figure 10.21 b.

d. Scent: Some *Begonia* species possess perceptibly scented flowers (e.g. flowers of *B. menyangensis*, *B. handelii*, *B. roxburghii*, *B. hatacoa* and *B. diadema* are sweet-scented; *B. herbacea* flowers are slightly almond-scented).

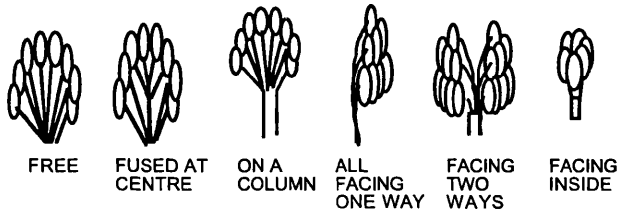
e. Size: Male and female flowers are often superficially similar, at least being approximately of equal sizes. However in a few taxa the female is many times larger than the male (e.g. *B. chlorosticta*; *B. incisa*; *B. maynensis,* Figure 10.19.1 d). While the actual tepal sizes may not be hugely different (both have two male and five female tepals; male tepal length is c. 7 mm and female tepal, c. 10 mm in *B. incisa* and c. 6 mm and c. 14 mm in *B. cholorosticta*) the females have very large and showy ovaries with obvious, coloured wings.

f. Male flowers

i. Androecium: The androecium of the male flowers can be very varied. All the stamens can be free (e.g. *B. handelii*, Figure 10.20 b), or the filaments can be fused to varying degrees into a column. This fusion can be along most of the filament length of all the stamens, creating a long, prominent column (e.g. *B. palmata, B. grandis*), or just among the central stamens with

anthers leaving the column at different heights (e.g. *B. annulata, B. fallax*). The anthers in these sorts of androecia are usually actinomorphic, forming a dome and facing all directions. The fusion of filaments can also be zygomorphic, to only one side of the androecium, creating an effect which has been compared to a hand of bananas (e.g. *B. letouzeyi*, Figure 10.20 d; *B. ampla*, Figure 10.20 f). The anthers in androecia which have this sort of fusion usually all face the same direction (the upper tepal), although occasionally half may face the upper, and half the lower, tepal. In a few taxa, all the anthers dehisce inwards, e.g. *B. subscutata*, section *Tetraphila*. See Figure 10.5 for an illustration of anther arrangements.

Figure 10.5: Anther arrangement



The anthers themselves may dehisce laterally, or the slits can open down the front of the anther. The anther may be longer than (e.g. *B. cubensis*, *B. dietrichiana* Irmsch.) or shorter than (e.g. *B. holtonis*, *B. ulmifolia*) the filament. All anthers may be the same length in one androecium, or they may vary. The connective can be the same length as the anthers or can be extended into a rounded tip (e.g. *B. wollnyi*) or a point (e.g. *B. menyangensis*) beyond them. The top of the connective may form a hood (e.g. *B. letouzeyi*). There are reports of dehiscence via pores rather than slits in some taxa, although the difference between a short slit and a pore may be marginal.

The numbers of stamens per flower varies from about five (e.g. *B. herbacea*) to well over 100 (e.g. *B. palmata*; *B.* sp., Yunnan 25, Figure 10.20 a). Lower anther number may be associated with increasing reliability of pollinators. The flowers of *B. herbacea* are quite strongly scented, in monosexual inflorescences on short pedicels - the inflorescence is hidden in the leaves. This may tie in with a specific pollinator. In other cases, stamen number may relate inversely to the number of flowers on an individual plant, with lower individual anther numbers per flower in species with bigger inflorescences.

The stamen colour is usually yellow (occasionally orange in some taxa, e.g. *B.*

wollnyi); however species of *Symbegonia* in cultivation in Glasgow have white filaments and red anthers.

ii. **Bud and sepal shape:** There are distinct differences in the shapes of male flower buds between taxa in *Begonia*. Many species have flat buds (e.g. *B. involucrata,* Figure 10.19.1 c; *B. maynensis,* Figure 10.19.1 d), while others have more or less spherical buds (e.g. *B. roxburghii,* Figure 10.19.2 d). These differences do not correlate exactly with the size of the androecium; although no taxa with a small number of stamens have spherical buds, not all taxa with many stamens do either. I do not know of any taxa with two tepals in the male flower which have spherical buds.

Sepal shape is reported (Burt-Utley, 1985) as varying not just on different flowers on an individual, but also over time - with differences before and during anthesis. Thus this is not a reliable character for phylogenetic analyses without some detailed quantification of variability.

g. Female flower

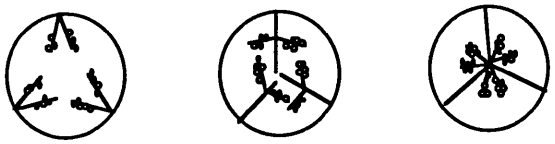
i. **Styles:** Style number varies, occasionally two (e.g. *B. goegoensis*) or four (e.g. *B. molleri,* Figure 10.21 f), but most commonly three (e.g. *B. convolvulacea*, Figure 10.21 h; *B. chlorosticta,* Figure 10.21 d). The styles may be free (e.g. *B. convolvulacea,* Figure 10.21 h) or fused to a varying degree (e.g. *B. boisiana*, Figure 10.21 e). They are usually bifid (e.g. *B. convolvulacea*, Figure 10.21 h), but may be entire (e.g. *B. gabonensis*; *B. mannii,* Figure 10.19.3 c), kidney-shaped (e.g. *B. letouzeyi*) or three- or four-fid (e.g. *B. fissistyla*). Stigmatic papillae usually occur in a spiralling band, but can be confined to the tips of the styles (e.g. *B. quadrialata*, *B. gabonensis*) or more widely across the surface (e.g. *B. annobonensis*, *B. fissistyla*).

Style colour is usually yellow, although may tend towards green or orange in some taxa.

ii. **Ovary:** The female flowers always have an inferior ovary (except in *Hillebrandia*, where the ovary is semi-inferior). Within the ovary, the most common state is three locules (e.g. *B. malachosticta*), although there are species with one (e.g. *B. masoniana*), two (B. *goegoensis*, *B. annulata*, *B. kingiana*, *B. imperialis*) and four (e.g. *B. handelii*, *B. letouzeyi*, *B. molleri*)

locules. Species with one locule have parietal placentation. The placentation in the other species is either axile (e.g. *B. palmata*) or septal (e.g. *B. gabonensis*). Septal placentation may occur when the multilocular condition is caused by the inward growth and fusion (or partial fusion) of parietal placentae, and so is really homologous with the parietal condition (Reitsma, 1984). It can be difficult to determine. See Figure 10.6 for an illustration of placentation types.

Figure 10.6: Placentation types



The placentae may be unbranched (e.g. *B. kingiana*) or bifid (e.g. *B. malachosticta*, *B. masoniana*), or less commonly many-fid. In the case of bifid placentae, ovules may be present on all surfaces or may be absent from the two facing surfaces within each locule (e.g. *B. solananthera*, *B. lubbersii*). Placentae may be green or white, and can be more or less fleshy (e.g. *B. princeps* A.DC. has quite fleshy placentae). In some species the ovules may sometimes appear pink (e.g. *B. dewildei*; although this is not consistent in the same individuals over time); ovules are normally white.

In some two-locular species, there has almost certainly been a secondary loss of one locule from an ancestral condition with three locules. Occasionally individual ovaries can be found within a tiny locule in the position where the third locule would ordinarily be. One fruit of *B. annulata* was found which had such a locule, with a small placenta and a few ovules. This aberrant fruit also had two normal styles and one highly reduced (aberrant) third style.

The ovary may have a number of wings running along it, from the style to the pedicel. Most *Begonia* species have three wings (e.g. *B. johnstonii,* Figure 10.22 a; *B. maynensis*, Figure 10.19.1 d; *B. herbacea*, Figure 10.21 g), although there are species with more, or where the wings are reduced or absent (e.g. *B. gabonensis* and *B. oxyloba* are wingless (Figure 10.22 b); *B. prismatocarpa* and *B. loranthoides* have ribs (Figure 10.19.3 b)). The wings can be equal in size (e.g. *B. dregei*, *B. brevirimosa*, *B. dietrichiana*), or (more commonly) the upper wing is distinctly larger (e.g. *B. glabra*; the unidentified species with

American *Begonia* Society (ABS) no. U205, Figure 10.22 d). In some species this larger wing is on the underside of the mature fruit; the pedicel is curved so that the whole fruit is bent back on itself (e.g. *B. palmata*; *B. hatacoa,* Figure 10.22 c). The wings, when present, may be the same colour as the rest of the ovary (e.g. *B. solananthera*, *B. dipetala* - wings and body white; *B. boisiana* - wings and body pink, Figure 10.21 e) or may be a different colour (e.g. *B. manicata* - pink wings, green body; *B. valida* - white wings, green body).

Species in the epiphytic section *Trachelocarpus* (e.g. *B. herbacea*) have a very distinctive fruit shape, with a long beak (or throat, as the section name suggests) between the top of the locules and the tepals (see Figure 10.21 g). This appears to have a similar function to the pedicel in other taxa, as the pedicel in these species is very short and the fruit sits more or less on the rhizome. The long beak lifts the tepals and stigmas out from among the leaves.

iii. Fruit: The fruit, when matured, is most commonly dry and papery (e.g. *B. johnstonii*, Figure 10.22 a; *B. glabra*), although in some taxa it is fleshy. Fleshy fruit are more frequent in wingless taxa (e.g. *B. oxyloba*, Figure 10.22 b; *B. gabonensis*), although some fleshy fruits have wings or ribs (e.g. *B.* sp. nov., Philippine; *B. bogneri*). The style (and less commonly, the tepals (e.g. *B. tomentosa* Schott)) can remain on the mature fruit (e.g. *B. annobonensis*, *B. socotrana*, *B. ulmifolia*), or can be deciduous (e.g. *B. brevirimosa*, *B. chlorosticta*). Fruits may be indehiscent (which often correlates with fleshiness) or form short splits near the pedicel, or right along the edges of the wings (e.g. *B. palmata*, *B. brevirimosa*). Fruits dehiscing through the wings have been reported in the literature. The fruit may be erect (e.g. *B. herbacea*), pendant (e.g. *B. chlorosticta*) or recurved (e.g. *B. palmata*).

10.2 Material and methods

10.2.1 Plant material: Taxa in this analysis are the same as those included in the ITS analysis; accession and voucher details are the same, and can be found on the CD-ROM.

10.2.2 Non-DNA character coding: Characters which refer to shape were avoided as far as possible, due to complications with scoring indiscrete characters. Because this matrix was intended to compliment an ITS DNA matrix, wherein the sampled taxa are individuals not species, a similar approach was taken with morphology: only the individual plant from which DNA was extracted was scored for the selected characters. This avoids any problems with plant identification which could occur were characters to be taken from literature, and of species delimitation which may complicate scoring characters from herbarium sheets. However, this approach does generate rather more 'missing data', and is particularly problematic where plants are dioecious or have not been known to flower in cultivation. For some taxa, herbarium sheets of the same accession were made at the time of collection (or after introduction to cultivation) and could be used to obtain floral characters for plants which did not flower within the time-frame of this study (*B. aequata*, Wilkie et al. 1997 2515, E; *B. formosana*, ETE 24, E; *B. oxysperma*, Wilkie et al. 29142, E; *B. rufo-sericae*, C11195, E; *B. serratipetala*, Reeves 588, E; *B.* sp. 'exotica', Reeves 142, E; *B.* sp., Sulawesi 252, Argent et al., 00116, E; *B.* sp., Sulawesi 253, Argent et al., 00151, E; *B.* sp., Sulawesi 254, Argent et al., 00152, E - see Appendix 14.5 for further details). The problem is less retractable for dioecious plants; in the case of *B. handelii*, male and female plants were collected at the same location in Yunnan; for *B. menyangensis*, although there is only a male plant in cultivation in Glasgow, there is field information for female plants from the same locality. For both these taxa it was decided to relax criteria and score characters from both sexes.

See Table 10.2 for a list of the non-DNA characters.

Table 10.2: Summary of non-DNA characters and their states

VEGETATIVE CHARACTERS

1. Stem tubers
   - 0: absent
   - 1: present

2. Root tubers
   - 0: absent
   - 1: present

3. Bulbils
   - 0: absent
   - 1: present

4. Tubercils
   - 0: absent
   - 1: present

5. Caudex
   - 0: absent
   - 1: present

6. Leaf shape
   - 0: simple
   - 1: compound

7. Peltateness
   - 0: basifixed
   - 1: peltate

8. Leaf maculation
   - 0: colour same all over
   - 1: with patterning

9. Petiole transverse section
   - 0: circular
   - 1: crescent
   - 2: square

10. Trichome ring at top of petiole
    - 0: absent
    - 1: present

11. Stipule persistence
    - 0: persistent
    - 1: caducous

12. Stipule pair
    - 0: both the same
    - 1: different

13. Stipule keeling
    - 0: indistinct
    - 1: strongly keeled

14. Stipule spur
    - 0: imperceptible
    - 1: distinctly spurred

15. Stipule edge
    - 0: entire
    - 1: fringed

16. Stipule back
    - 0: glabrous
    - 1: hairy

17. 'Fuzzy' hairs [8]
    - 0: absent
    - 1: present

18. Stellate hairs [9]
    - 0: absent
    - 1: present

SEXUAL CHARACTERS - INFLORESCENCE

19. Lifestyle
    0: perennial
    1: monocarpic

20. Inflorescence position
    0: axile
    1: terminal

21. Inflorescences per axil
    0: one
    1: more

22. Sexual separation
    0: dioecious
    1: monoecious

23 Inflorescence type
    0: cyme
    1: raceme

24. Inflorescence branching at base
    0: dichasial
    1: monochasial

25. Inflorescence symmetry
    0: symmetric
    1: asymmetric

26. Dichasial inflorescence: basal dichotomies
    0: with central flower
    1: without central flower

27. Flower number per inflorescence
    0: less than 70
    1: over 100

28. Sexual separation
    0: male and female in same inflorescence, interspersed
    1: male and female in same inflorescence, female basal
    2: male and female on separate inflorescences

29. Flower sizes
    0: similar in male and female
    1: distinctly larger female than male

30. Flower colour (most prevalent)
    0: white or pink
    1: yellow
    2: red
    3: orange

30: Flower pattern
    0: tepals all one colour
    1: both tepals with similar red veins or patches
    2: red veins or patches only on one tepal

32. Scent
    0: imperceptible
    1: strong

33. Perianth tube
    0: absent
    1: present

SEXUAL CHARACTERS - MALE FLOWER

34. Male tepal number
    0: 2 tepals
    1: 4 tepals
    2: absent

35. Male flower symmetry
    0: radial symmetry of tepals
    1: bilateral symmetry of tepals

36. Male tepal fusion
    0: free
    1: partly fused

37. Male tepal hairiness
    0: glabrous
    1: with hairs

38: Male tepal edge
    0: entire
    1: lobed

39. Male bud shape
    0: flat
    1: spherical

40. Androecium
    0: anthers face all directions
    1: anthers face upper and lower tepals
    2: anthers face upper tepal

41. Stamen number
    0: less than 10
    1: 10 or more

42. Stamen colour
    0: yellow
    1: orange
    2: red

43. Anther dehiscence
    0: via slits
    1: via pores

44. Stamen fusion
    0: free
    1: fused only in the centre
    2: fused only at one side
    3: on a column (all fused)

45. Anther connective extension
    0: not extended
    1: extended

46. Anther connective hooding
    0: not hooded
    1: hooded

SEXUAL CHARACTERS - FEMALE FLOWER

47. Female tepal number
    0: 2 tepals
    1: 3 tepals
    2: 4 tepals
    3: 5 tepals
    4: 6 tepals
    7: absent

48. Female tepal fusion
    0: free
    1: two tepals partly fused
    2: all tepals partly fused

49. Female tepal hairiness
    0: glabrous
    1: hairy

50. Female tepal edge
    0: entire
    1: lobed or serrate

51. Style number
    0: 2 styles
    1: 3 styles
    2: 4 styles
    3: (5-) 6 (-7) styles

52. Style colour
    0: yellow
    1: greenish
    2: white
    3: pink
    4: red

53. Style fusion
    0: free
    1: fused

54. Style branching
    0: unbranched
    1: kidney-shaped
    2: bifid
    3: 3-fid - 4-fid

55. Style persistence on fruit
    0: persistent
    1: caducous

56. Ovary position
    0: inferior
    1: semi-inferior

57. Locule number
    0: 1 locular
    1: 2 locular
    2: 3 locular
    3: 4 locular
    4: (5-) 6 (-7) locular

58. Placentation
    0: parietal
    1: septal
    2. axile

59. Placentation
    0: one-fid
    1: bifid, with ovules on inner and outer surfaces of placentae
    2: bifid, with ovules only on outer surfaces of placentae

60. Fruit wing number
    0: absent
    1: 2 wings
    2: 3 wings
    3: 4 wings
    4: c. 6 wings (coronate)
    5: 1 wing

61. Fruit wing symmetry
    0: equal to subequal
    1: one distinctly larger

62. Fruit dry or fleshy
    0: dry
    1: fleshy

63. Fruit orientation
    0: upright
    1: pendant to nodding
    2: recurved

64: Fruit hair
    0: glabrous
    1: with hairs

65: Beaked fruit
    0: absent
    1: present

66: Dehiscence
    0: not between styles
    1: between styles

67. Bracteole subtending ovary
    0: absent
    1: 2 bracteoles
    2: 3 bracteoles

The data matrix for these non-DNA characters is included on the CD-ROM.

10.2.3 Cladistic Analyses

Several characters were missing for many taxa. For example, the differences between the outer and inner stipules in some taxa were not observed until most species had been scored. This character is not easily observed from herbarium material, and it was not possible to revisit all the living plants to add in this data. Several other characters were felt to be obviously homoplastic, e.g. hairiness on the backs of stipules. However, *a priori* exclusion of data on such grounds involves the assumption that taxa which share this character are unrelated; this should be a deduction from the analysis, not an assumption of it.

10.2.3.1 Data sets: The taxa in the morphological matrix were sorted into the same order as the taxa in the ITS matrix; duplicate sequences (e.g. clones, different primers) were removed and the two matrices were combined by interleaving.

Thus there are three data sets to be analysed in this chapter:

1. The non--DNA matrix
2. The corresponding ITS matrix
3. The combined non-DNA and ITS matrix.

10.2.3.2 Analyses: Analyses were run using PAUP* 4.0 (Swofford, 2000). For each matrix, g1 was estimated using 10,000 random trees. PTP was estimated with the outgroup (the two *Datisca* species) excluded, with 100 replicates, simple addition, saving no more than five trees per step. An heuristic search was run, 1000 random additions, saving no more than five trees at each step, steepest descent, TBR swapping. MaxTrees was set to 1000. The strict consensus topology from the resulting trees was input as a constraint file, and a further heuristic search with 1000 random additions, TBR, saving only five trees at any step, was performed to see if any other equally parsimonious topologies were supported. Bootstrapping was performed with the fast heuristic option in PAUP, 5000 replicates. Bremer support was estimated using AutoDecay (Erikkson, 1998) (10 random addition replicates, TBR, steepest descent, maximum of five trees held per step).

10.3 Results

10.3.1. Non-DNA data set: The skewedness statistic g1 is -0.1799. PTP probability is 0.010. One thousand MPTs were found, of length 499. Searching with topological constraints found a shortest tree length of 501.

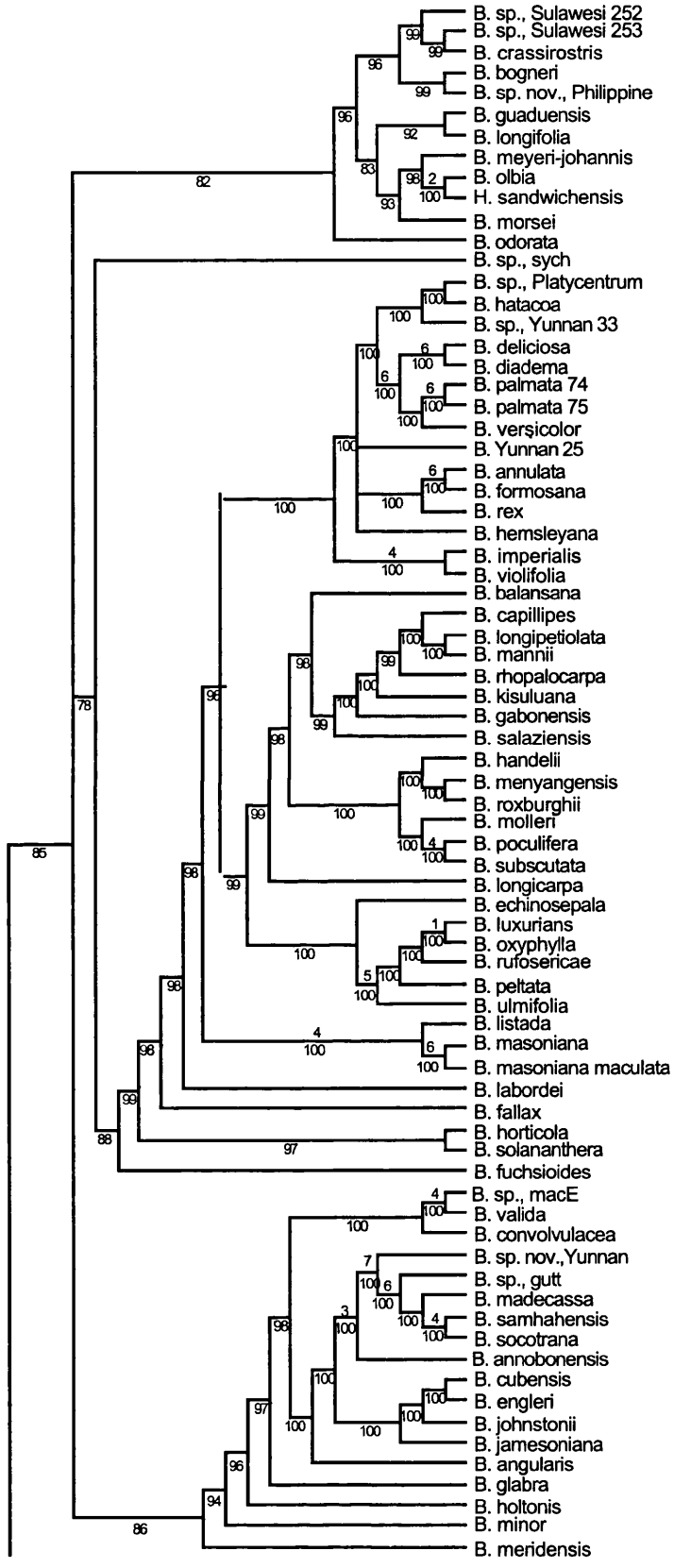
There was little resolution in the strict consensus tree; a majority rule tree is presented (Figure 10.7). Nodes are annotated with the percentage of trees they appear in and Bremer support values on the relevant clades. A phylogram is also presented (Figure 10.8).

There are 33 nodes in the strict consensus tree; five nodes have over 50% bootstrap support. Clades with bootstrap support over 50% are listed below:

55%, *B. grandis* ssp. *grandis* and *B. grandis* ssp. *holostyla*;
64%, *B. letouzeyi* and *B. quadrialata*;
61%, *B. masoniana* and *B. masoniana* var. *maculata*;
74%, *B. samhahensis* and *B. socotrana*;
97%, *D. cannabina* and *D. glomerata*.

The consistency index is 0.21 (0.20 excluding uninformative characters) and the retention index is 0.65.

Figure 10.7: Majority rule cladogram from 1000 MPTs, non-DNA data set



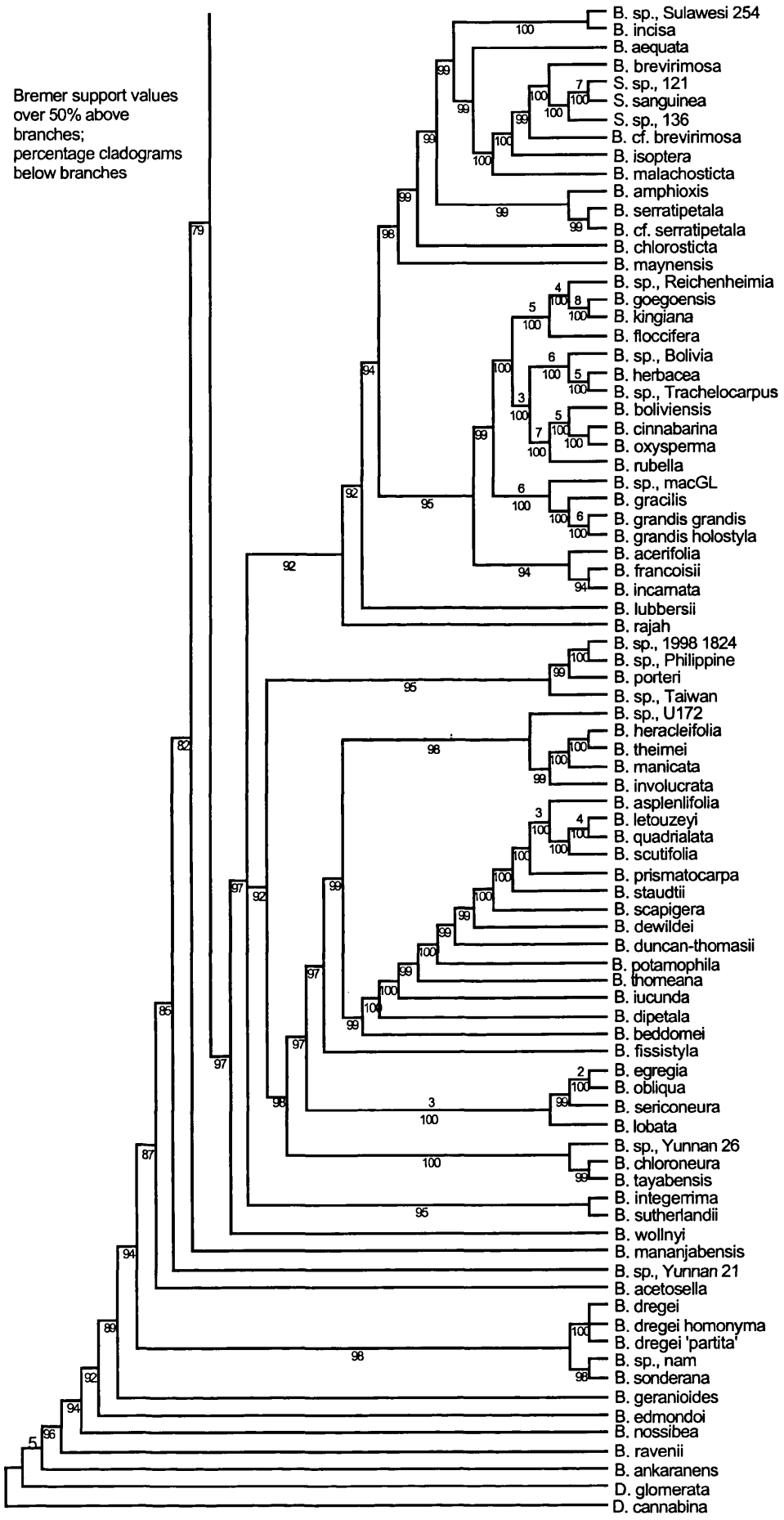
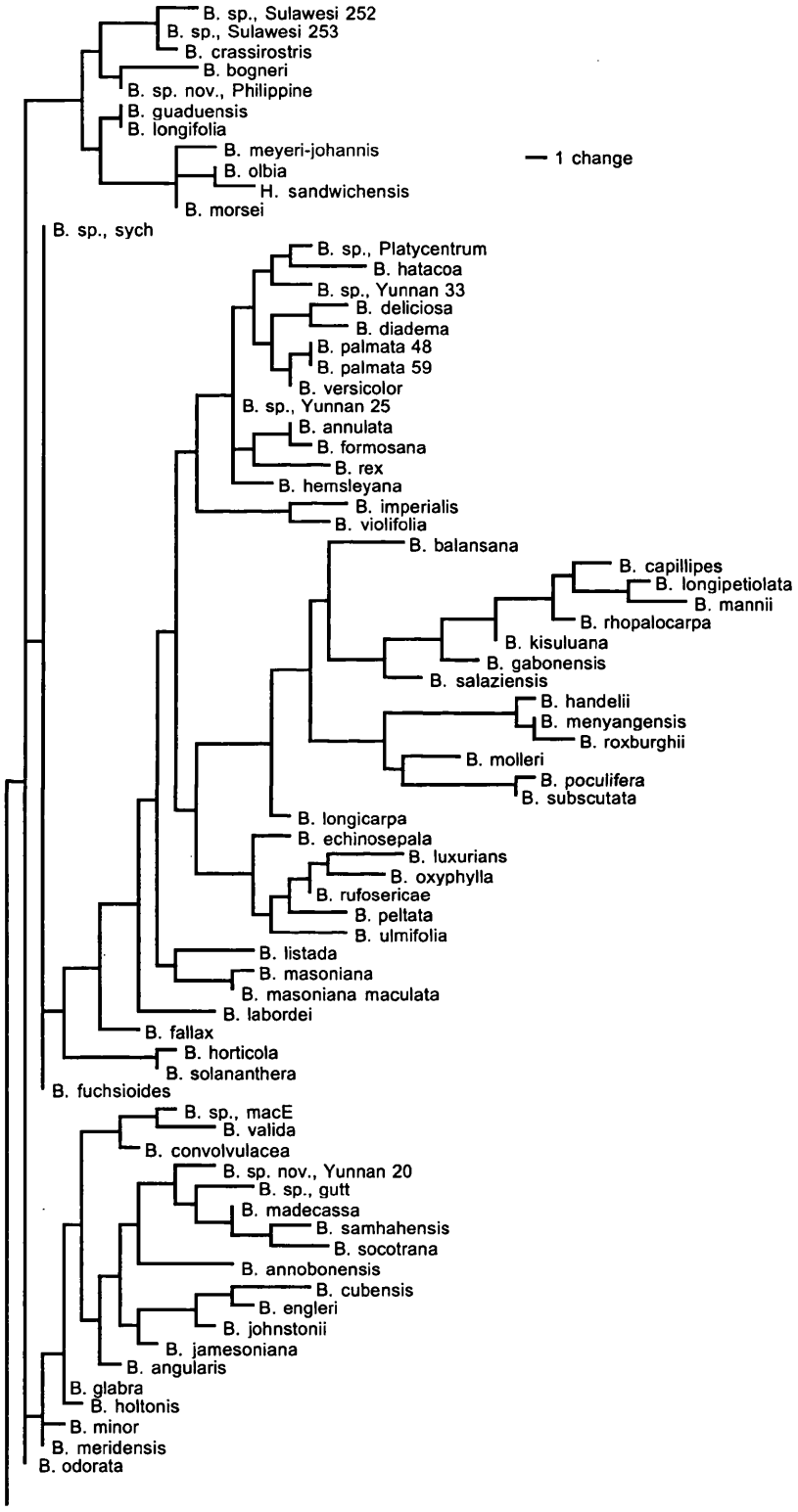
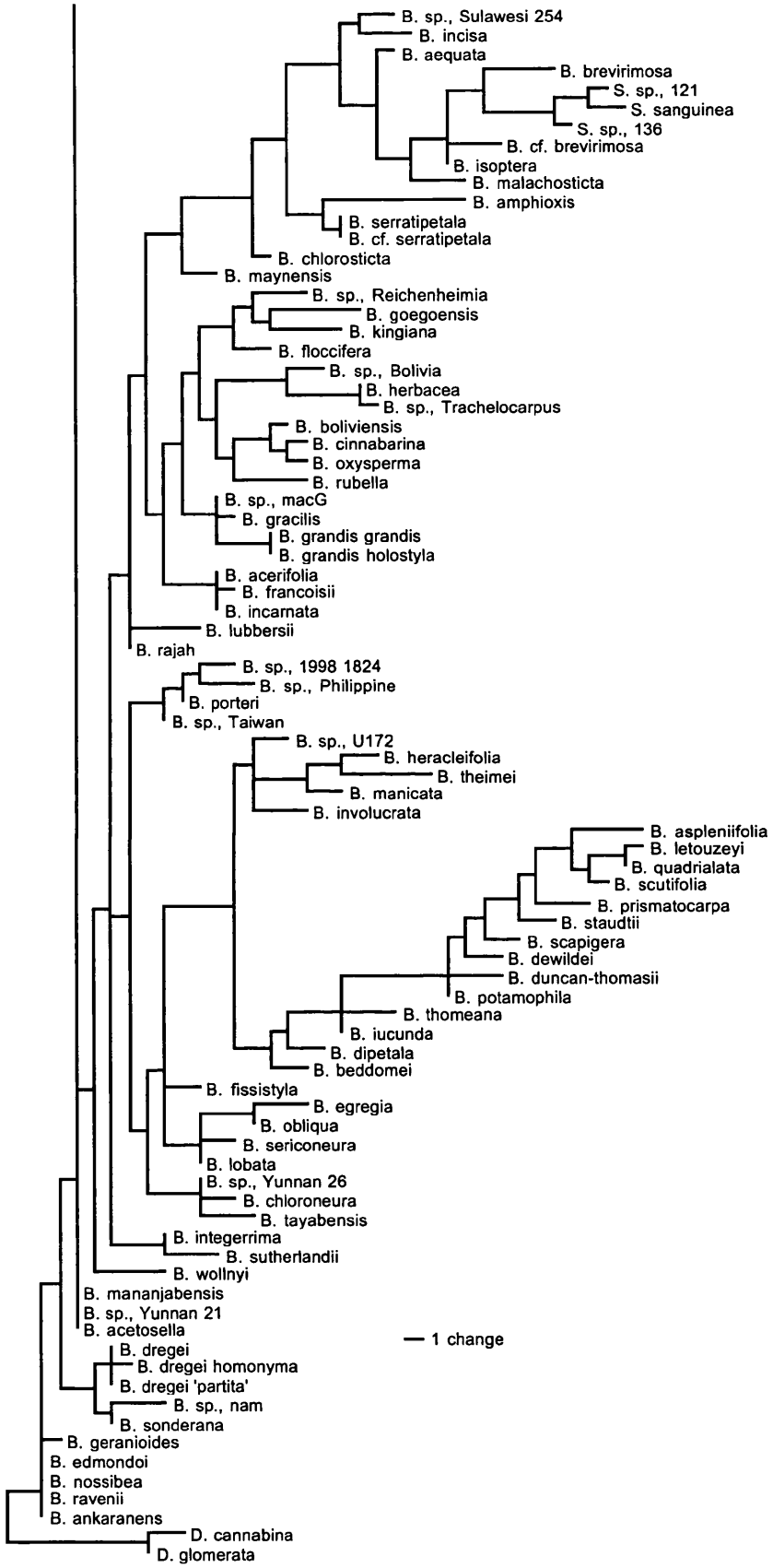

Bremer support values over 50% above branches; percentage cladograms below branches
B. sp., Sulawesi 254
B. incisa
B. aequata
B. brevirimosa
S. sp., 121
S. sanguinea
S. sp., 136
B. cf. brevirimosa
B. isoptera
B. malachosticta
B. amphioxis
B. serratipetala
B. cf. serratipetala
B. chlorosticta
B. maynensis
B. sp., Reichenheimia
B. goegoensis
B. kingiana
B. floccifera
B. sp., Bolivia
B. herbacea
B. sp., Trachelocarpus
B. boliviensis
B. cinnabarina
B. oxysperma
B. rubella
B. sp., macGL
B. gracilis
B. grandis grandis
B. grandis holostyla
B. acerifolia
B. francoisii
B. incarnata
B. lubbersii
B. rajah
B. sp., 1998 1824
B. sp., Philippine
B. porteri
B. sp., Taiwan
B. sp., U172
B. heracleifolia
B. theimei
B. manicata
B. involucrata
B. asplenlifolia
B. letouzeyi
B. quadrialata
B. scutifolia
B. prismatocarpa
B. staudtii
B. scapigera
B. dewildei
B. duncan-thomasii
B. potamophila
B. thomeana
B. iucunda
B. dipetala
B. beddomei
B. fissistyla
B. egregia
B. obliqua
B. sericoneura
B. lobata
B. sp., Yunnan 26
B. chloroneura
B. tayabensis
B. integerrima
B. sutherlandii
B. wollnyi
B. mananjabensis
B. sp., Yunnan 21
B. acetosella
B. dregei
B. dregei homonyma
B. dregei 'partita'
B. sp., nam
B. sonderana
B. geranioides
B. edmondoi
B. nossibea
B. ravenii
B. ankaranens
D. glomerata
D. cannabina

Figure 10.8: Phylogram of one of 1000 MPTs, non-DNA characters




B. sp., Sulawesi 254
B. incisa
B. aequata
B. brevirimosa
S. sp., 121
S. sanguinea
S. sp., 136
B. cf. brevirimosa
B. isoptera
B. malachosticta
B. amphioxis
B. serratipetala
B. cf. serratipetala
B. chlorosticta
B. maynensis
B. sp., Reichenheimia
B. goegoensis
B. kingiana
B. floccifera
B. sp., Bolivia
B. herbacea
B. sp., Trachelocarpus
B. boliviensis
B. cinnabarina
B. oxysperma
B. rubella
B. sp., macG
B. gracilis
B. grandis grandis
B. grandis holostyla
B. acerifolia
B. francoisii
B. incarnata
B. lubbersii
B. rajah
B. sp., 1998 1824
B. sp., Philippine
B. porteri
B. sp., Taiwan
B. sp., U172
B. heracleifolia
B. theimei
B. manicata
B. involucrata
B. aspleniifolia
B. letouzeyi
B. quadrialata
B. scutifolia
B. prismatocarpa
B. staudtii
B. scapigera
B. dewildei
B. duncan-thomasii
B. potamophila
B. thomeana
B. iucunda
B. dipetala
B. beddomei
B. fissistyla
B. egregia
B. obliqua
B. sericoneura
B. lobata
B. sp., Yunnan 26
B. chloroneura
B. tayabensis
B. integerrima
B. sutherlandii
B. wollnyi
B. mananjabensis
B. sp., Yunnan 21
B. acetosella
B. dregei
B. dregei homonyma
B. dregei 'partita'
B. sp., nam
B. sonderana
B. geranioides
B. edmondoi
B. nossibea
B. ravenii
B. ankaranens
D. cannabina
D. glomerata
— 1 change

There are no clear correlations between this tree and the ITS trees. The clear geographical structuring apparent in analyses of ITS data is lost here. A few clades survive - the *Loasibegonia* group of species appear to be held together by several morphological characters, and *Petermannia*/*Symbegonia* also hold together. In other clades, it is possible to guess which characters are responsible for the unconventional groupings - there is a clade which includes some fleshy-fruited African species from section *Tetraphila* and some fleshy fruited Asian species. Elsewhere there is a clade of orange-flowered taxa (*B. boliviensis*, *B. cinnabarina* and *B. oxysperma*). In general, however, this tree makes little sense in the light either of previous taxonomic treatments or of geographical distributions, or in the light of the ITS, 26S and *trnC - trnD* cladograms discussed previously.

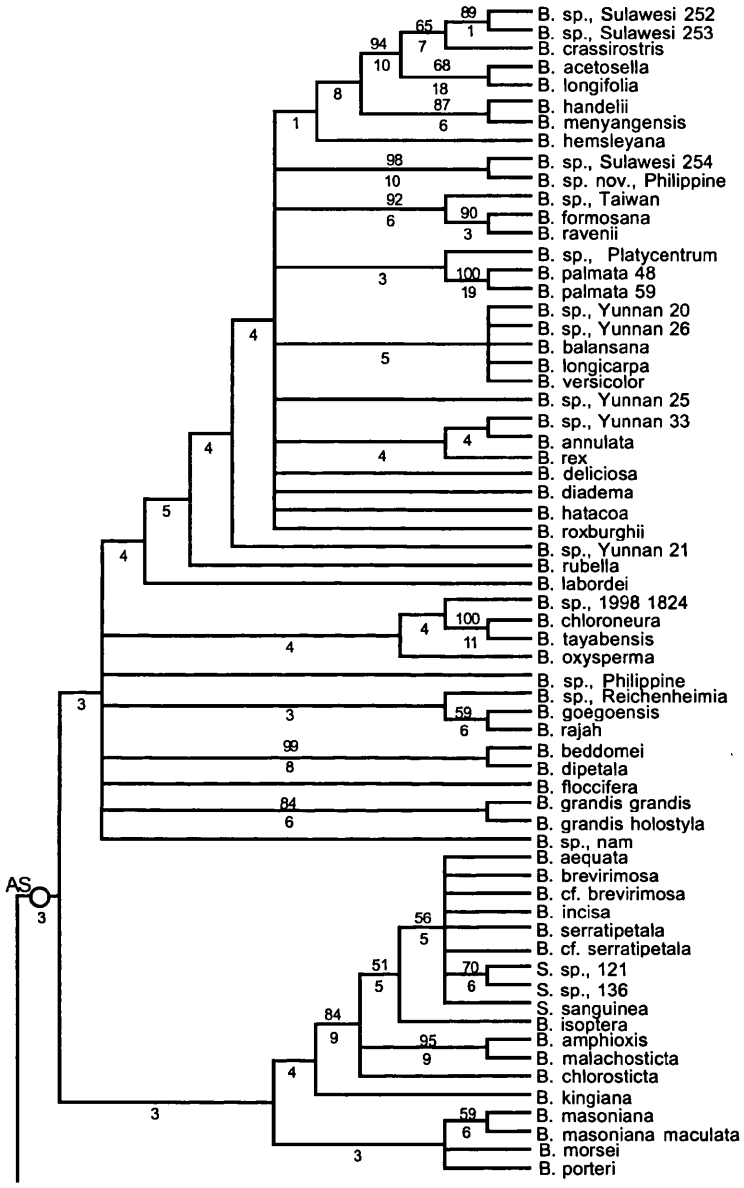
Using majority rule to summarise a group of alternative topologies is one thing; using it as an estimator of phylogeny would be very different, and not justifiable under a criterion of parsimony. There will be other equally parsimonious topologies which are not congruent with this topology. Given that the strict consensus tree for these data is highly unresolved (N.B. the 100% branches on the majority rule tree do not all appear in the strict consensus of 1000 MPTs; presumably a grouping found in 999, i.e. 99.9% of the trees, is rounded up to 100%) there is little that can be said about *Begonia* evolution based on this analysis.

10.3.2. ITS sequence data analysis:

There were 122 constant, 92 uninformative, and 311 informative characters included. The skewedness statistic g1 for this data set is -0.453.

One thousand MPTs were found, of length 2702. Searching with topological constraints found trees of length 2703. The consistency index is 0.30; with uninformative characters excluded it is 0.27. Retention index is 0.68. The strict consensus tree is presented as Figure 10.9; one of the MPTs is presented as a phylogram, Figure 10.10. There are 108 nodes in the strict consensus tree; 71 nodes have over 50% bootstrap support.

Figure 10.9: Strict consensus of 1000 MPTs, ITS data set



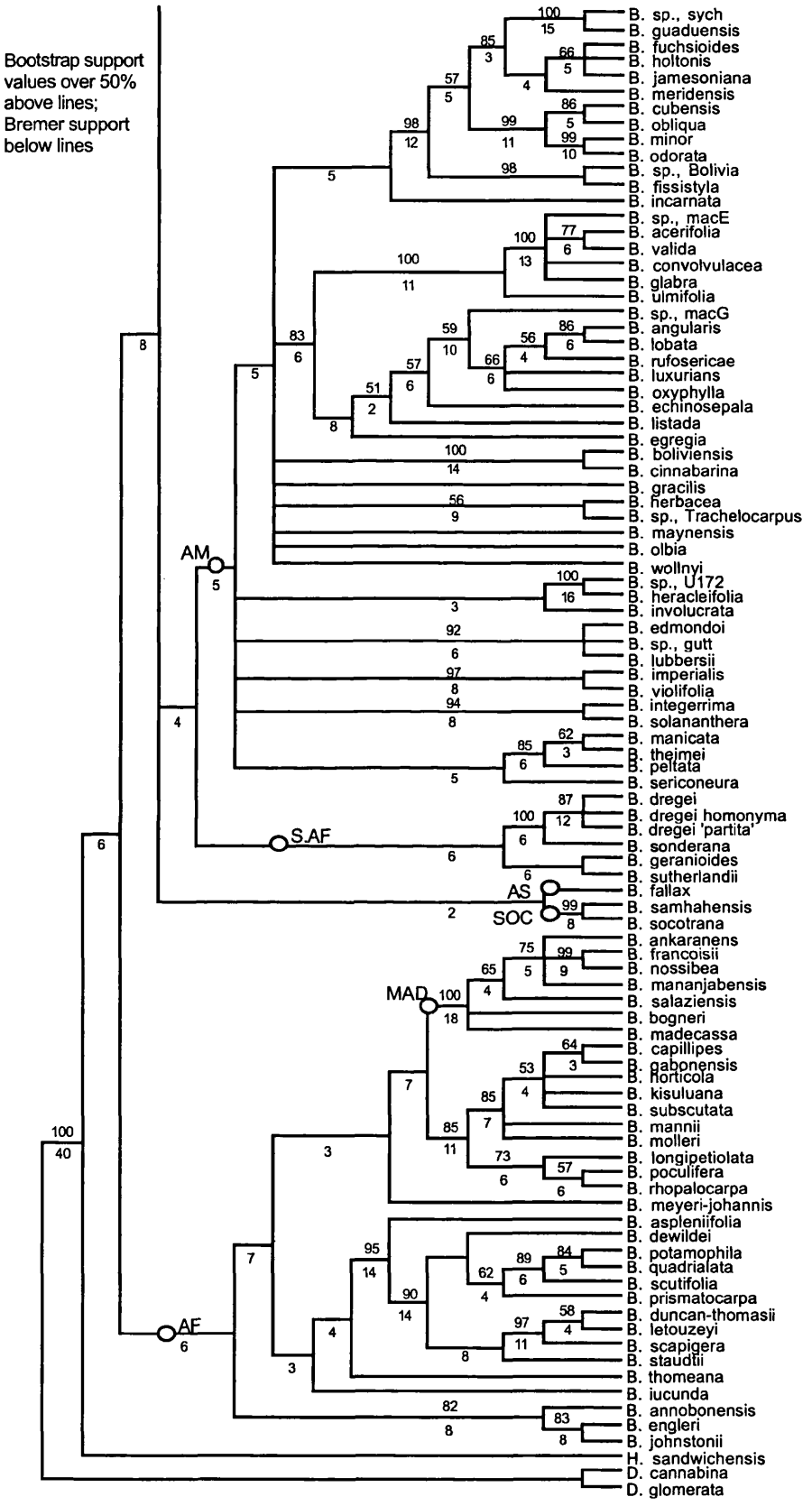
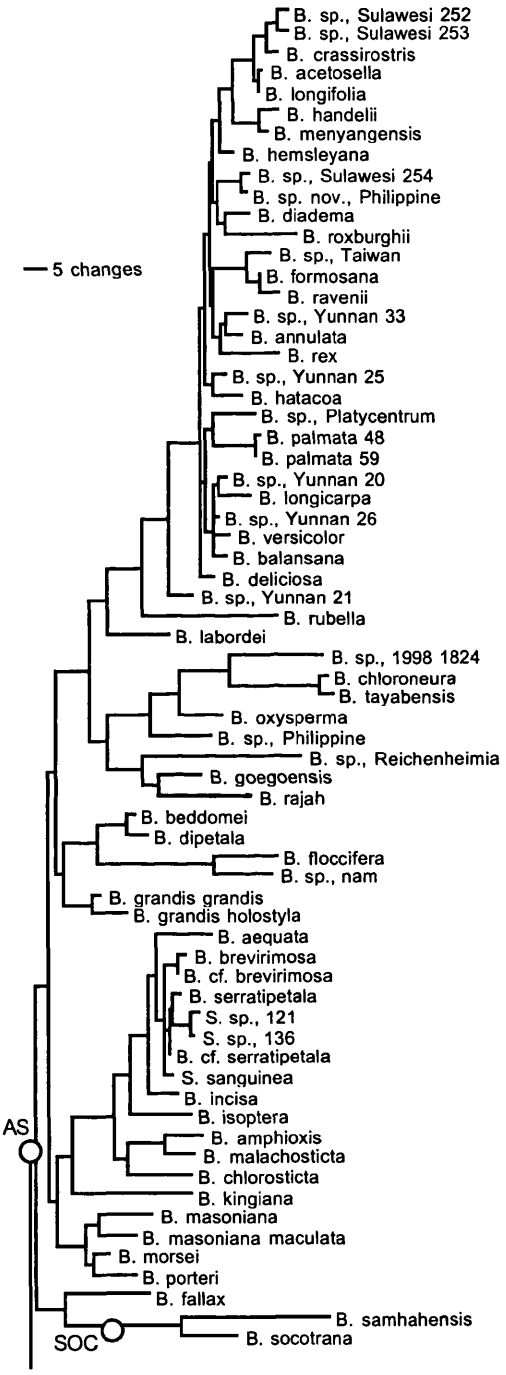
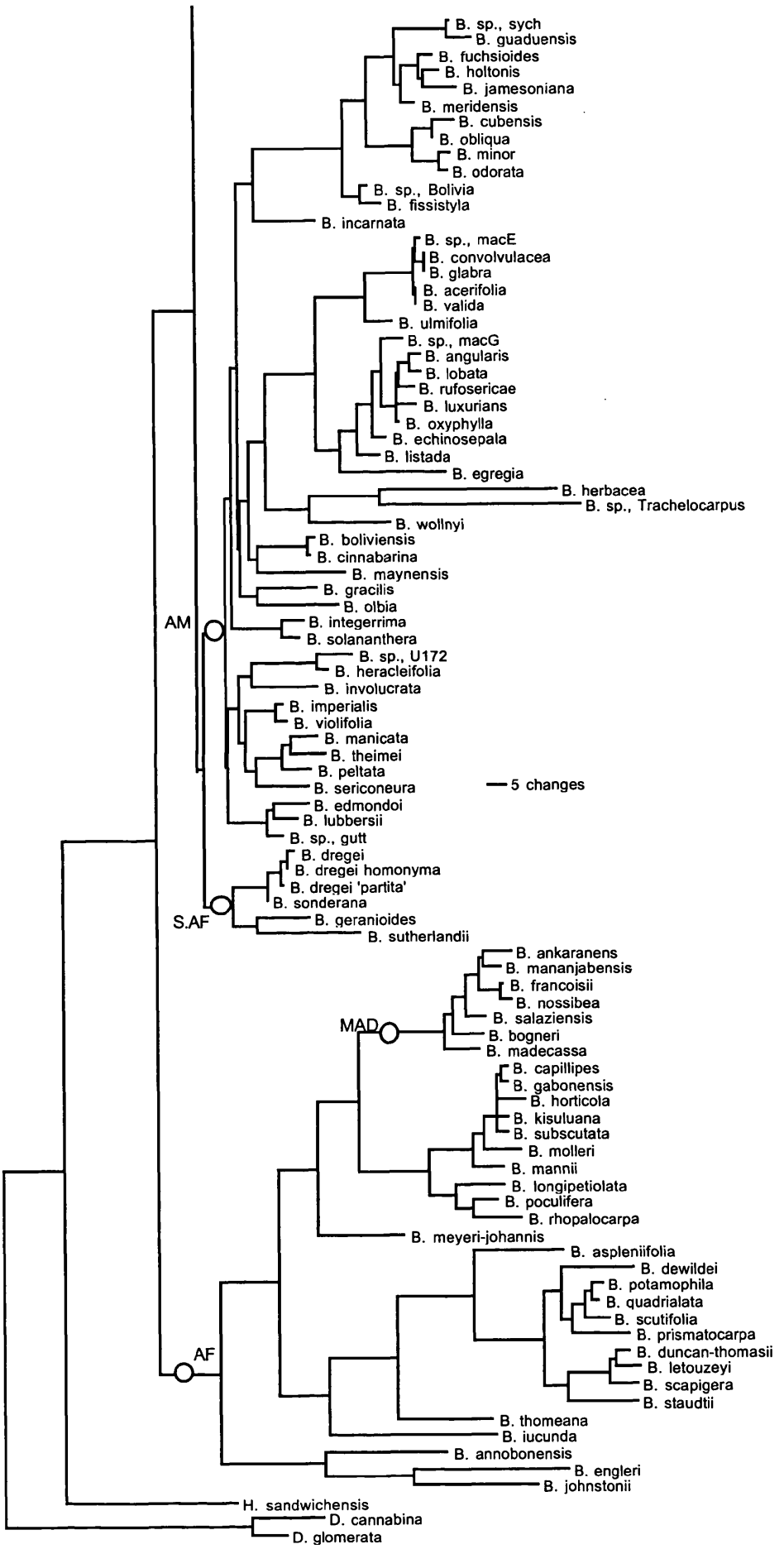

Bootstrap support values over 50% above lines; Bremer support below lines
B. sp., sych
B. guaduensis
B. fuchsioides
B. holtonis
B. jamesoniana
B. meridensis
B. cubensis
B. obliqua
B. minor
B. odorata
B. sp., Bolivia
B. fissistyla
B. incarnata
B. sp., macE
B. acerifolia
B. valida
B. convolvulacea
B. glabra
B. ulmifolia
B. sp., macG
B. angularis
B. lobata
B. rufosericae
B. luxurians
B. oxyphylla
B. echinosepala
B. listada
B. egregia
B. boliviensis
B. cinnabarina
B. gracilis
B. herbacea
B. sp., Trachelocarpus
B. maynensis
B. olbia
B. wollnyi
B. sp., U172
B. heracleifolia
B. involucrata
B. edmondoi
B. sp., gutt
B. lubbersii
B. imperialis
B. violifolia
B. integerrima
B. solananthera
B. manicata
B. theimei
B. peltata
B. sericoneura
B. dregei
B. dregei homonyma
B. dregei 'partita'
B. sonderana
B. geranioides
B. sutherlandii
B. fallax
B. samhahensis
B. socotrana
B. ankaranens
B. francoisii
B. nossibea
B. mananjabensis
B. salaziensis
B. bogneri
B. madecassa
B. capillipes
B. gabonensis
B. horticola
B. kisuluana
B. subscutata
B. mannii
B. molleri
B. longipetiolata
B. poculifera
B. rhopalocarpa
B. meyeri-johannis
B. aspleniifolia
B. dewildei
B. potamophila
B. quadrialata
B. scutifolia
B. prismatocarpa
B. duncan-thomasii
B. letouzeyi
B. scapigera
B. staudtii
B. thomeana
B. iucunda
B. annobonensis
B. engleri
B. johnstonii
H. sandwichensis
D. cannabina
D. glomerata
AM
S.AF
AS
SOC
MAD
AF

Figure 10.10: Phylogram, ITS data set, one of 1000 MPTs




B. sp., sych
B. guaduensis
B. fuchsioides
B. holtonis
B. jamesoniana
B. meridensis
B. cubensis
B. obliqua
B. minor
B. odorata
B. sp., Bolivia
B. fissistyla
B. incarnata
B. sp., macE
B. convolvulacea
B. glabra
B. acerifolia
B. valida
B. ulmifolia
B. sp., macG
B. angularis
B. lobata
B. rufosericae
B. luxurians
B. oxyphylla
B. echinosepala
B. listada
B. egregia
B. herbacea
B. sp., Trachelocarpus
B. wollnyi
B. boliviensis
B. cinnabarina
B. maynensis
B. gracilis
B. olbia
AM
B. integerrima
B. solananthera
B. sp., U172
B. heracleifolia
B. involucrata
B. imperialis
B. violifolia
B. manicata
B. theimei
B. peltata
B. sericoneura
5 changes
B. edmondoi
B. lubbersii
B. sp., gutt
B. dregei
B. dregei homonyma
B. dregei 'partita'
B. sonderana
S.AF
B. geranioides
B. sutherlandii
B. ankaranens
B. mananjabensis
B. francoisii
B. nossibea
B. salaziensis
MAD
B. bogneri
B. madecassa
B. capillipes
B. gabonensis
B. horticola
B. kisuluana
B. subscutata
B. molleri
B. mannii
B. longipetiolata
B. poculifera
B. rhopalocarpa
B. meyeri-johannis
B. aspleniifolia
B. dewildei
B. potamophila
B. quadrialata
B. scutifolia
B. prismatocarpa
B. duncan-thomasii
B. letouzeyi
B. scapigera
B. staudtii
AF
B. thomeana
B. iucunda
B. annobonensis
B. engleri
B. johnstonii
H. sandwichensis
D. cannabina
D. glomerata

10.3.3. The combined ITS/non-DNA analysis:

Of a total of 1224 characters, 632 are excluded (the ITS exclusion set from previous chapters); 123 constant, 95 uninformative and 374 parsimony informative characters are included.

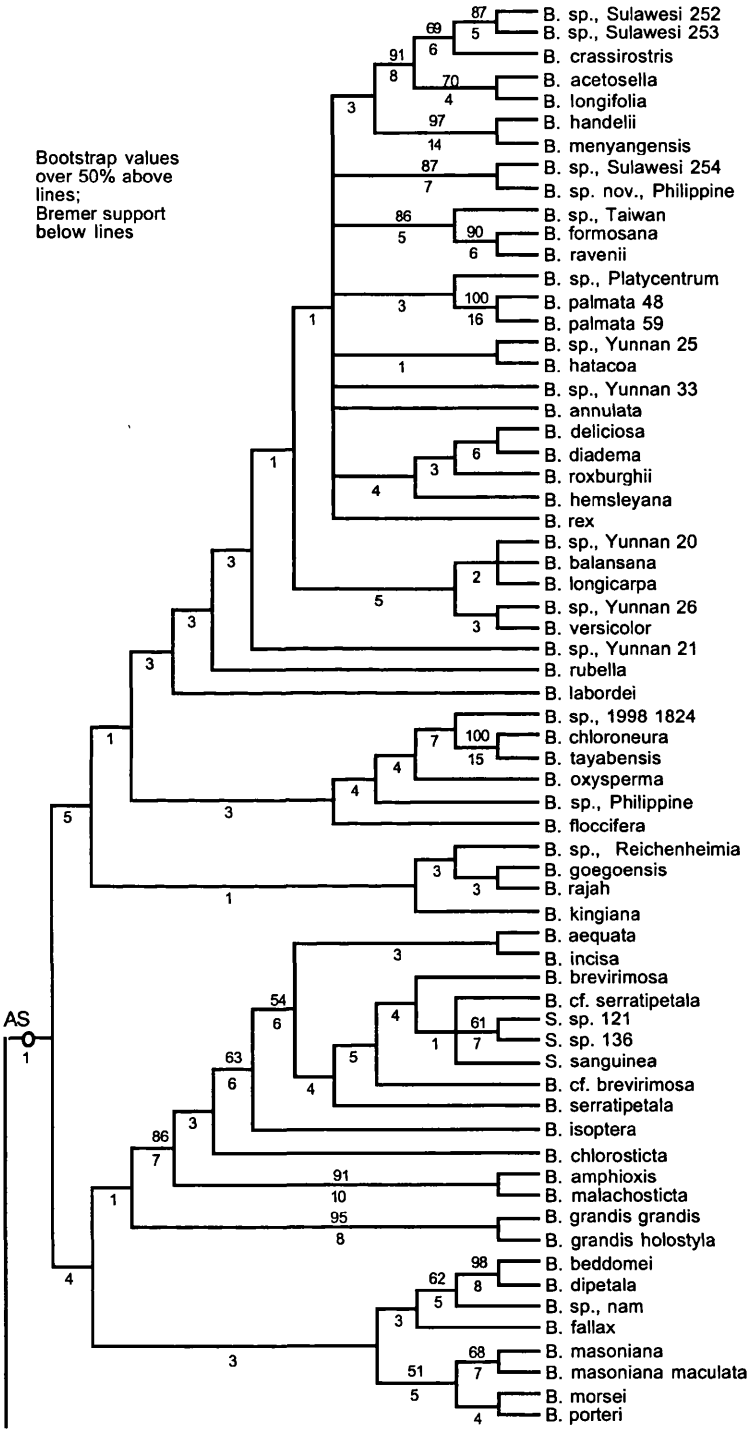
The skewedness statistic g1 is -0.4778.

One thousand MPTs of length 3365 were found; searching with topological constraints found a shortest length of 3367.

Consistency index is 0.27 (0.24 with uninformative characters excluded); retention index is 0.65. There are 131 nodes in the strict consensus tree; 67 nodes have over 50% bootstrap support.

The strict consensus tree is presented as Figure 10.11, and one of the 1000 MPTs is presented as Figure 10.12.

Figure 10.11: Strict consensus of 1000 MPTs, combined non-DNA and ITS



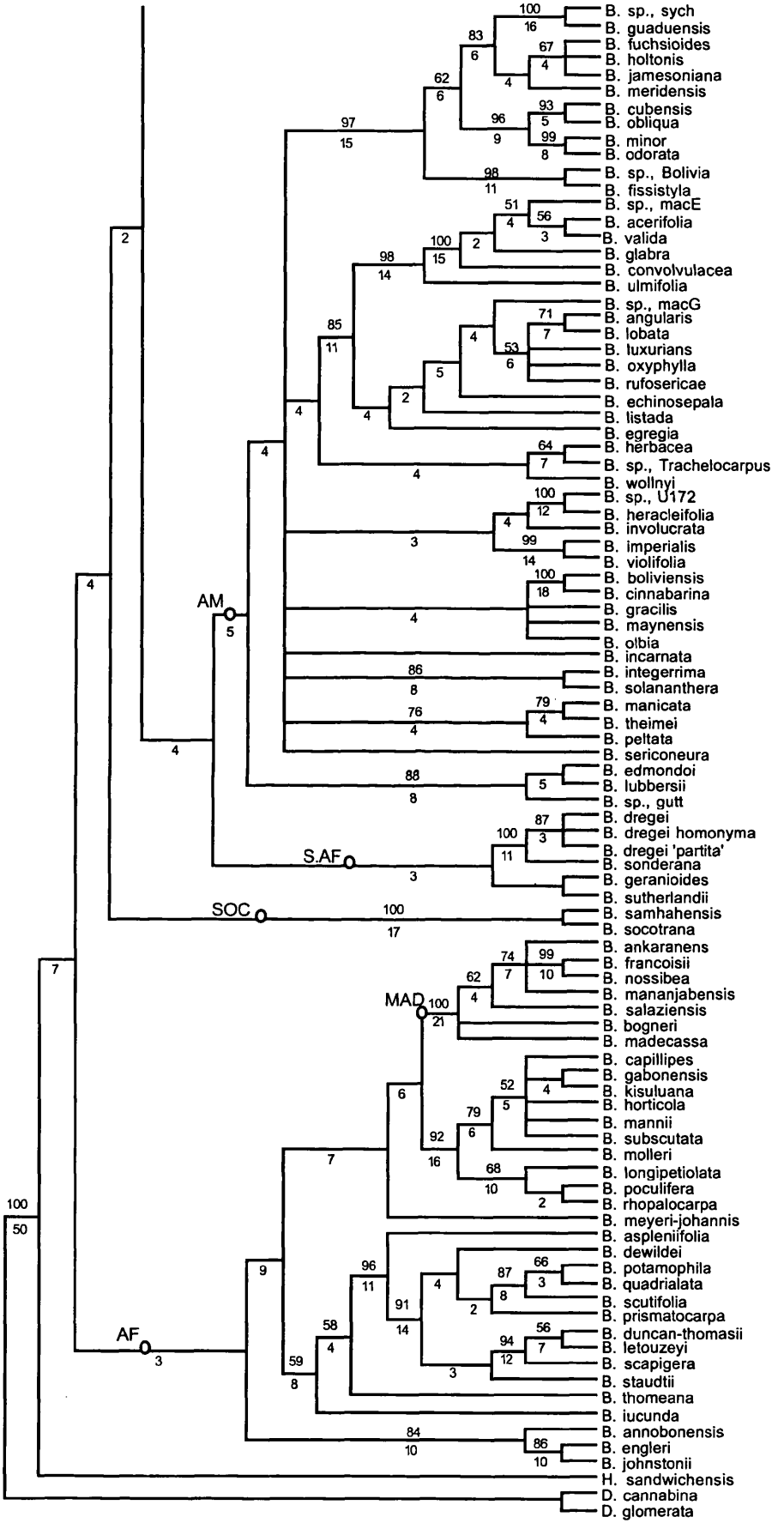
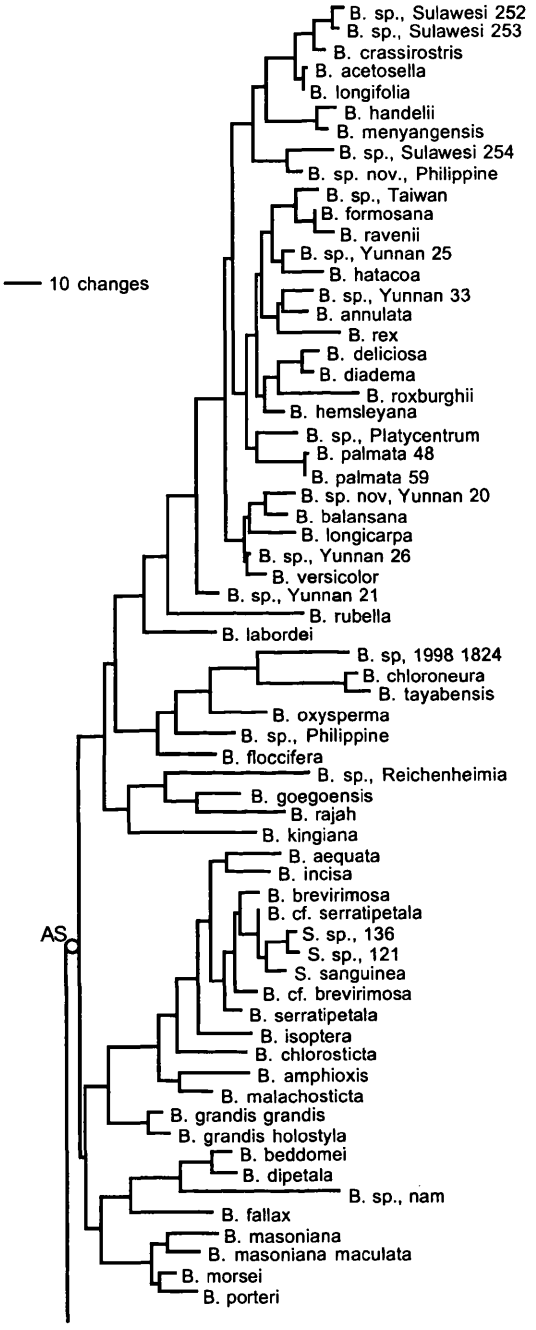
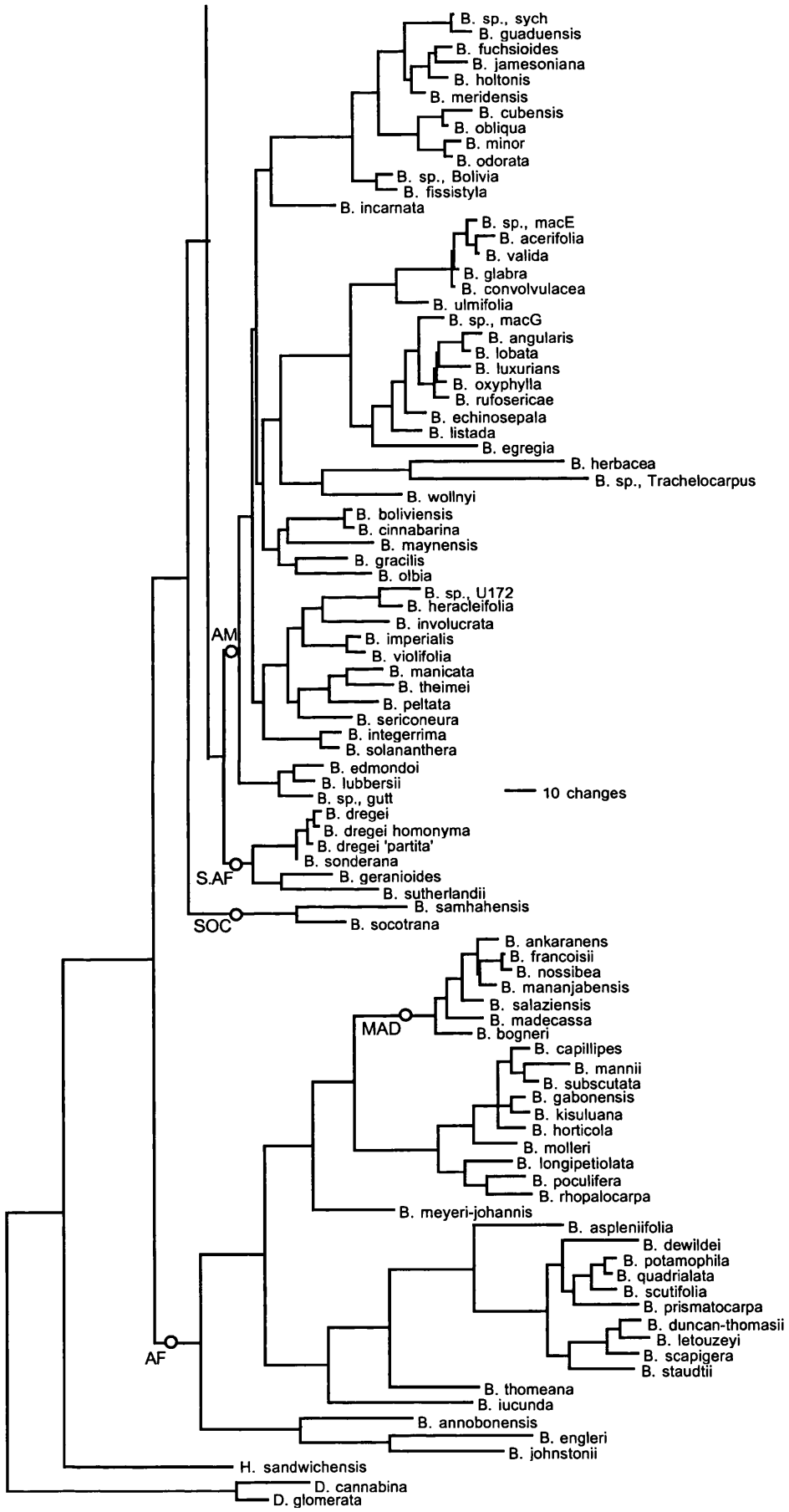

B. sp., sych
B. guaduensis
B. fuchsioides
B. holtonis
B. jamesoniana
B. meridensis
B. cubensis
B. obliqua
B. minor
B. odorata
B. sp., Bolivia
B. fissistyla
B. sp., macE
B. acerifolia
B. valida
B. glabra
B. convolvulacea
B. ulmifolia
B. sp., macG
B. angularis
B. lobata
B. luxurians
B. oxyphylla
B. rufosericae
B. echinosepala
B. listada
B. egregia
B. herbacea
B. sp., Trachelocarpus
B. wollnyi
B. sp., U172
B. heracleifolia
B. involucrata
B. imperialis
B. violifolia
B. boliviensis
B. cinnabarina
B. gracilis
B. maynensis
B. olbia
B. incarnata
B. integerrima
B. solananthera
B. manicata
B. theimei
B. peltata
B. sericoneura
B. edmondoi
B. lubbersii
B. sp., gutt
B. dregei
B. dregei homonyma
B. dregei 'partita'
B. sonderana
B. geranioides
B. sutherlandii
B. samhahensis
B. socotrana
B. ankaranens
B. francoisii
B. nossibea
B. mananjabensis
B. salaziensis
B. bogneri
B. madecassa
B. capillipes
B. gabonensis
B. kisuluana
B. horticola
B. mannii
B. subscutata
B. molleri
B. longipetiolata
B. poculifera
B. rhopalocarpa
B. meyeri-johannis
B. aspleniifolia
B. dewildei
B. potamophila
B. quadrialata
B. scutifolia
B. prismatocarpa
B. duncan-thomasii
B. letouzeyi
B. scapigera
B. staudtii
B. thomeana
B. iucunda
B. annobonensis
B. engleri
B. johnstonii
H. sandwichensis
D. cannabina
D. glomerata
AM
S.AF
SOC
MAD
AF

Figure 10.12: Phylogram, one of 1000 MPTs, combined non-DNA and ITS




B. sp., sych
B. guaduensis
B. fuchsioides
B. jamesoniana
B. holtonis
B. meridensis
B. cubensis
B. obliqua
B. minor
B. odorata
B. sp., Bolivia
B. fissistyla
B. incarnata
B. sp., macE
B. acerifolia
B. valida
B. glabra
B. convolvulacea
B. ulmifolia
B. sp., macG
B. angularis
B. lobata
B. luxurians
B. oxyphylla
B. rufosericae
B. echinosepala
B. listada
B. egregia
B. herbacea
B. sp., Trachelocarpus
B. wollnyi
B. boliviensis
B. cinnabarina
B. maynensis
B. gracilis
B. olbia
B. sp., U172
B. heracleifolia
B. involucrata
AM
B. imperialis
B. violifolia
B. manicata
B. theimei
B. peltata
B. sericoneura
B. integerrima
B. solananthera
B. edmondoi
B. lubbersii
B. sp., gutt
10 changes
B. dregei
B. dregei homonyma
B. dregei 'partita'
B. sonderana
S.AF
B. geranioides
B. sutherlandii
B. samhahensis
SOC
B. socotrana
B. ankaranens
B. francoisii
B. nossibea
B. mananjabensis
B. salaziensis
B. madecassa
MAD
B. bogneri
B. capillipes
B. mannii
B. subscutata
B. gabonensis
B. kisuluana
B. horticola
B. molleri
B. longipetiolata
B. poculifera
B. rhopalocarpa
B. meyeri-johannis
B. aspleniifolia
B. dewildei
B. potamophila
B. quadrialata
B. scutifolia
B. prismatocarpa
B. duncan-thomasii
B. letouzeyi
B. scapigera
B. staudtii
AF
B. thomeana
B. iucunda
B. annobonensis
B. engleri
B. johnstonii
H. sandwichensis
D. cannabina
D. glomerata

10.3.4 Tree comparisons:

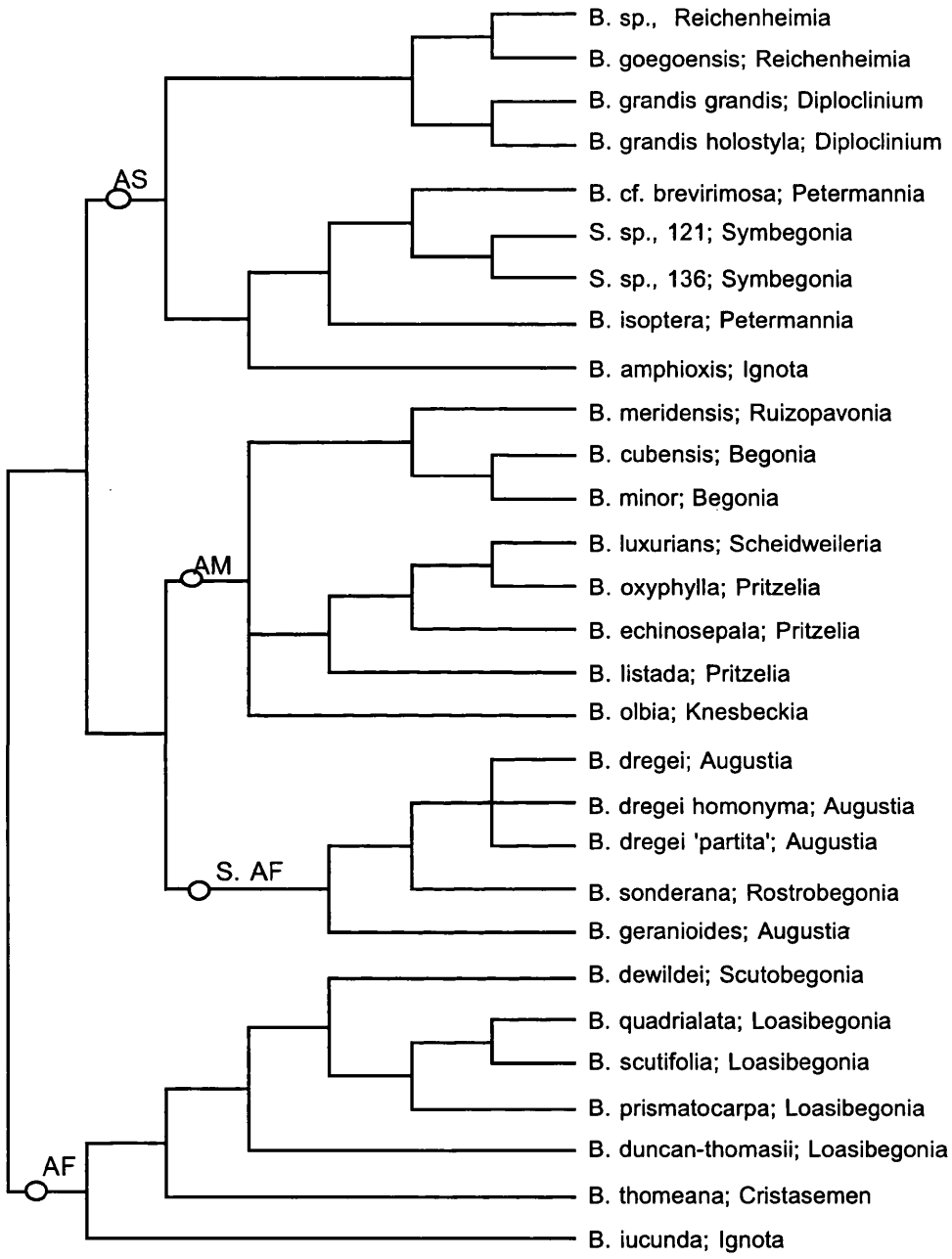
See Table 10.3 for a summary of the statistics for the different analyses. From this, it can be seen that the ITS data set performs better than the other two (non-DNA and combined) in terms of consistency and retention indices and in the bootstrap support for nodes, although more nodes are resolved in the strict consensus tree of the combined data analysis. The combined analysis MPT length is 164 steps longer than the total lengths of the non-DNA and ITS MPTs.

Table 10.3: Data and tree statistics for the non-DNA, ITS and combined analyses.

DATA SET	INF. CHARS	g1	No. MPTs	LENGTH	CI	CI ex UNINF.	RI	No. NODES	NODES > 50% BS
MORPH	63	-0.180	1000	499	0.21	0.20	0.65	33	5
ITS	311	-0.453	1000	2702	0.30	0.27	0.68	108	71
COMBINED	374	-0.478	1000	3365	0.27	0.24	0.65	131	67

The majority rule tree for the non-DNA data was compared with the strict consensus tree for the ITS data. The partition metric is 233 (maximum value is 312, i.e. 74.7% of the maximum); agree D_1 is 130. The agreement subtree tree for the majority-rule non-DNA tree and the strict consensus tree for ITS is as follows (size 29/159, i.e. with 130 taxa pruned) (Figure 10.13):

Figure 10.13: Agreement subtree tree, non-DNA and ITS analyses



It is obvious, on the basis of this tree (Figure 10.13), that while there are areas of agreement between the non-DNA and ITS data sets, there is also a lot of conflict (with 130 taxa in different places). Indeed, tracing morphological characters, across non-DNA, combined and ITS trees, shows one of the major problems with the non-DNA characters chosen in this study. Many of the characters which offer good support for some clades are homoplastic in others. Furthermore, many of the characters which are characteristic of clades

suffer reversals and state changes within those clades. The 'best' characters in terms of fit are those, like character 65, beaked fruit (see Table 10.4), which only occur in a few closely related species; unfortunately these are not informative about the deeper level relationships across *Begonia*.

10.3.5 Character performance: For a breakdown of how different non-DNA characters performed when analysed in combination with the ITS data set, see Table 10.4.

Table 10.4: Statistics for individual morphological characters, over a tree produced by analysis of the combined ITS - non-DNA data set.

Character	Range	Min steps	Tree steps	Max steps	CI	RI	RC	HI	G-fit
1 (Tubers - stem)	1	1	4	5	0.250	0.250	0.062	0.750	0.500
2 (Tubers - root)	1	1	4	5	0.250	0.250	0.062	0.750	0.500
3 (Bulbils)	1	1	1	2	1.000	1.000	1.000	0.000	1.000
4 (Tubercils)	1	1	1	2	1.000	1.000	1.000	0.000	1.000
5 (Caudex)	1	1	1	3	1.000	1.000	1.000	0.000	1.000
6 (Leaf shape)	1	1	4	5	0.250	0.250	0.062	0.750	0.500
7 (Peltateness)	1	1	10	18	0.100	0.471	0.047	0.900	0.250
8 (Leaf colour)	1	1	16	25	0.062	0.375	0.023	0.938	0.167
9 (Petiole TS)	2	2	27	52	0.074	0.500	0.037	0.926	0.107
10 (Trichome ring)	2	2	10	13	0.200	0.273	0.055	0.800	0.273
11 (Stipule persistence)	1	1	19	25	0.053	0.250	0.013	0.947	0.143
12 (Stipule pair)	1	1	6	8	0.167	0.286	0.048	0.833	0.375
13 (Stipule keel)	2	2	9	11	0.222	0.222	0.049	0.778	0.300
14 (Stipule spur)	1	1	20	29	0.050	0.321	0.016	0.950	0.136
15 (Stipule edge)	1	1	11	20	0.091	0.474	0.043	0.909	0.231
16 (Stipule back)	1	1	31	61	0.032	0.500	0.016	0.968	0.091
17 (Fuzzy hair)	1	1	7	9	0.143	0.250	0.036	0.857	0.333
18 (Stellate hair)	1	1	2	11	0.500	0.900	0.450	0.500	0.750
19 (Lifestyle)	1	1	1	1	1.000	0/0	0/0	0.000	1.000
20 (Inflor. position)	1	1	5	6	0.200	0.200	0.040	0.800	0.429
21 (Inflor./axil)	1	1	2	2	0.500	0.000	0.000	0.500	0.750
22 (Sexual separation)	1	1	3	5	0.333	0.500	0.167	0.667	0.600
23 (Inflor. type)	1	1	5	16	0.200	0.733	0.147	0.800	0.429
24 (Cyme type)	1	1	1	9	1.000	1.000	1.000	0.000	1.000
25 (Inflor. symm.)	1	1	7	10	0.143	0.333	0.048	0.857	0.333
26 (Inflor. basal dichotomy)	1	1	7	13	0.143	0.500	0.071	0.857	0.333
27 (Flowers/inflor.)	1	1	5	8	0.200	0.429	0.086	0.800	0.429
28 (Sexual separation)	2	2	10	21	0.200	0.579	0.116	0.800	0.273
29 (Flower size)	1	1	5	8	0.200	0.429	0.086	0.800	0.429
30 (Flower colour)	3	3	8	15	0.375	0.583	0.219	0.625	0.375
31 (Flower pattern)	2	2	13	19	0.154	0.353	0.054	0.846	0.214
32 (Scent)	1	1	7	11	0.143	0.400	0.057	0.857	0.333
33 (Perianth tube)	1	1	1	1	1.000	0/0	0/0	0.000	1.000
34 (Male tepal no.)	3	3	16	45	0.188	0.690	0.129	0.812	0.188
35 (Male flower symm.)	1	1	13	23	0.077	0.455	0.035	0.923	0.200
36 (Male tepal fusion)	1	1	1	3	1.000	1.000	1.000	0.000	1.000
37 (Male tepal hair)	1	1	24	41	0.042	0.425	0.018	0.958	0.115
38 (Male tepal edge)	1	1	2	2	0.500	0.000	0.000	0.500	0.750
40 (Androecium)	3	3	10	22	0.300	0.632	0.189	0.700	0.300
41 (Stamen no.)	1	1	9	9	0.111	0.000	0.000	0.889	0.273
42 (Stamen colour)	2	2	2	4	1.000	1.000	1.000	0.000	1.000
44 (Stamen fusion)	3	3	36	70	0.083	0.507	0.042	0.917	0.083
45 (Anther connective ext.)	1	1	18	46	0.056	0.622	0.035	0.944	0.150
46 (Anther connective hood)	1	1	18	23	0.056	0.227	0.013	0.944	0.150
47 (Female tepal no.)	6	6	28	60	0.214	0.593	0.127	0.786	0.120
48 (Female tepal fusion)	2	2	4	7	0.500	0.600	0.300	0.500	0.600
49 (Female tepal hair)	1	1	18	32	0.056	0.452	0.025	0.944	0.150
50 (Female tepal edge)	1	1	5	5	0.200	0.000	0.000	0.800	0.429
51 (Style no.)	4	4	11	29	0.364	0.720	0.262	0.636	0.300
52 (Style colour)	4	4	6	8	0.667	0.500	0.333	0.333	0.600
53 (Style fusion)	1	1	25	42	0.040	0.415	0.017	0.960	0.111

54 (Style branching)	3	3	16	26	0.188	0.435	0.082	0.812	0.188
55 (Style persistence)	1	1	6	10	0.167	0.444	0.074	0.833	0.375
56 (Ovary position)	1	1	1	1	1.000	0/0	0/0	0.000	1.000
57 (Locule no.)	4	4	14	36	0.286	0.688	0.196	0.714	0.231
58 (Placentation type)	2	2	8	14	0.250	0.500	0.125	0.750	0.333
59 (Placentation no.)	2	2	16	37	0.125	0.600	0.075	0.875	0.176
60 (Fruit wing no.)	4	4	9	26	0.444	0.773	0.343	0.556	0.375
61 (Fruit wing symm.)	1	1	17	43	0.059	0.619	0.036	0.941	0.158
62 (Fruit dry/fleshy)	1	1	6	19	0.167	0.722	0.120	0.833	0.375
63 (Fruit orientation)	2	2	6	16	0.333	0.714	0.238	0.667	0.429
64 (Fruit hair)	1	1	24	44	0.042	0.465	0.019	0.958	0.115
65 (Fruit beaking)	1	1	1	2	1.000	1.000	1.000	0.000	1.000
66 (Dehiscence)	1	1	1	3	1.000	1.000	1.000	0.000	1.000
67 (Bracteole)	2	2	13	21	0.154	0.421	0.065	0.846	0.214

10.3.6 Character evolution, some case studies: A few selected characters have been reconstructed across one of the MPTs from the combined ITS - non-DNA data matrix. (Reconstructions from MacClade, ACCTRAN optimisation).

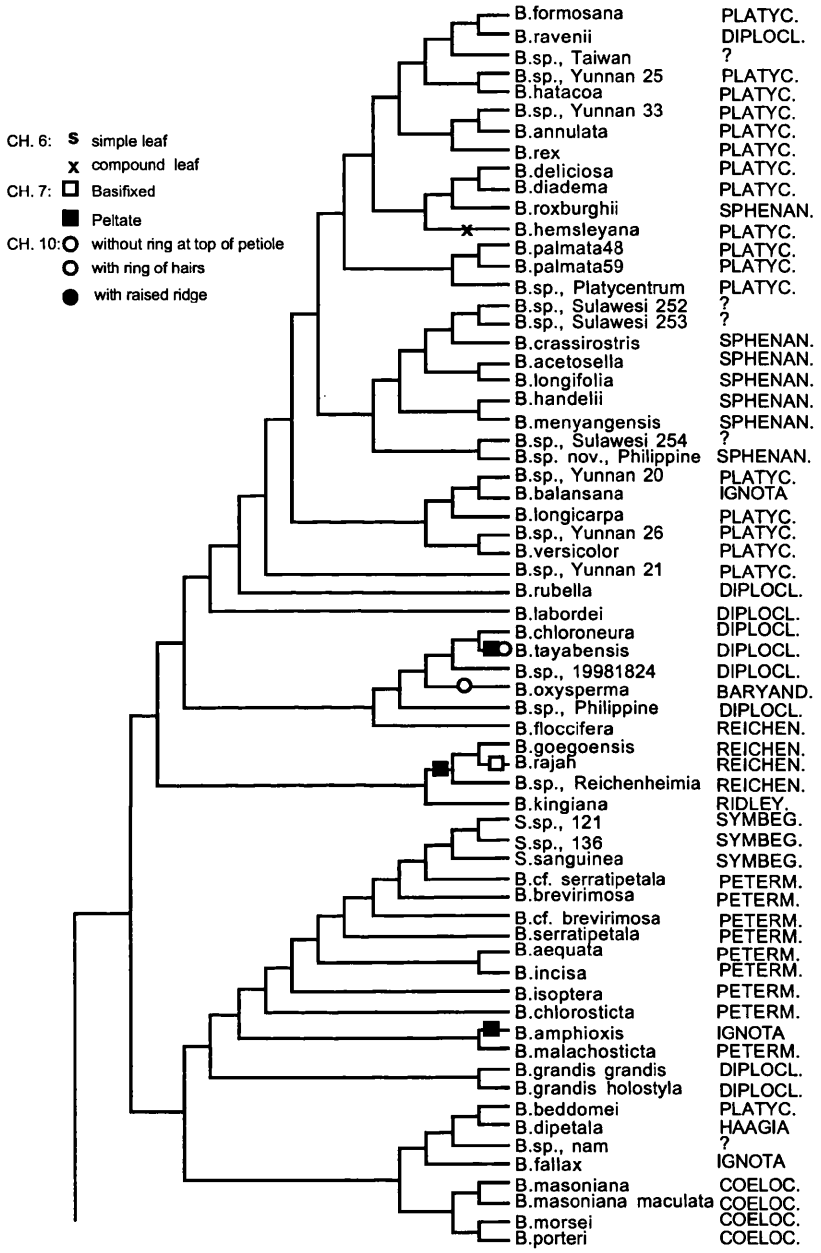
A. Leaf characters: Figure 10.14 shows characters 6 (leaf: compound or simple), 7 (leaf peltateness) and 10 (trichome ring at top of petiole).

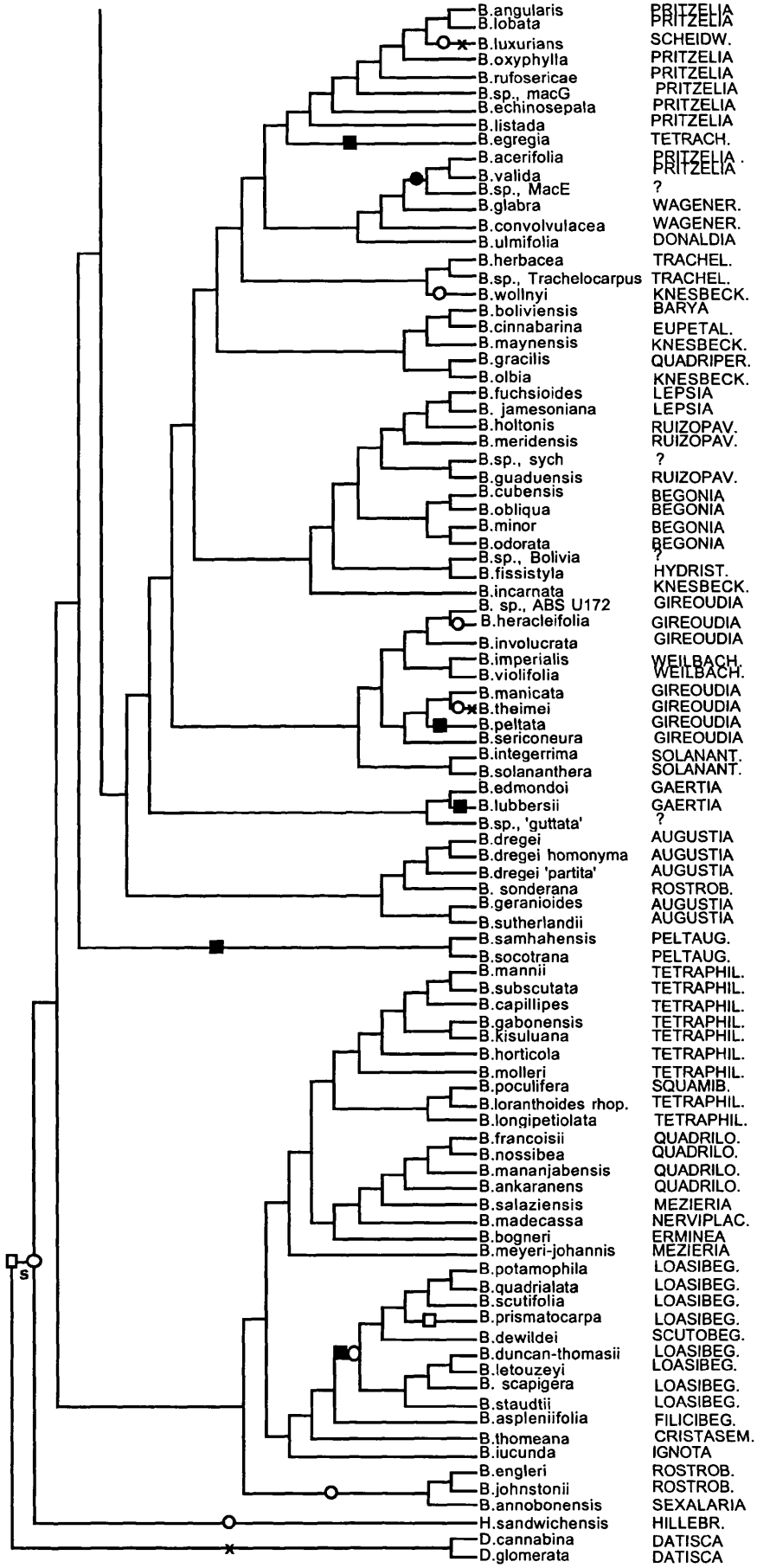
Out of a maximum of five possible steps, character 6 (leaf compound/simple) takes four steps on this tree (one between the outgroup and ingroup) (ci 0.25; ri 0.25). It seems that compound leaves have arisen several times independently within Begoniaceae.

Leaf peltateness is a 'better' character, at least in terms of retention; out of 18 possible steps it takes 10 (ci 0.10; ri 0.47). Peltateness is a good character for the *Loasibegonia*/*Scutobegonia* clade (*B. staudtii* to *B. potamophila*), with only one reversal (*B. prismatocarpa*) and for the *Peltaugustia* clade (*B. socotrana* and *B. samhahensis*). It may also be useful in section *Reichenheimia* (*B.* sp., *Reichenheimia* to *B. goegoensis*); on this tree it resolves as belonging to the ancestor for the section, with a reversal in *B. rajah*.

The presence or absence of a trichome ring is very homoplastic - out of 13 possible steps it takes 10 (ci 0.20; ri 0.27), although it does group *B. annobonensis*, *B. engleri* and *B. johnstonii*.

Figure 10.14: Leaf characters, ACCTRAN optimisation

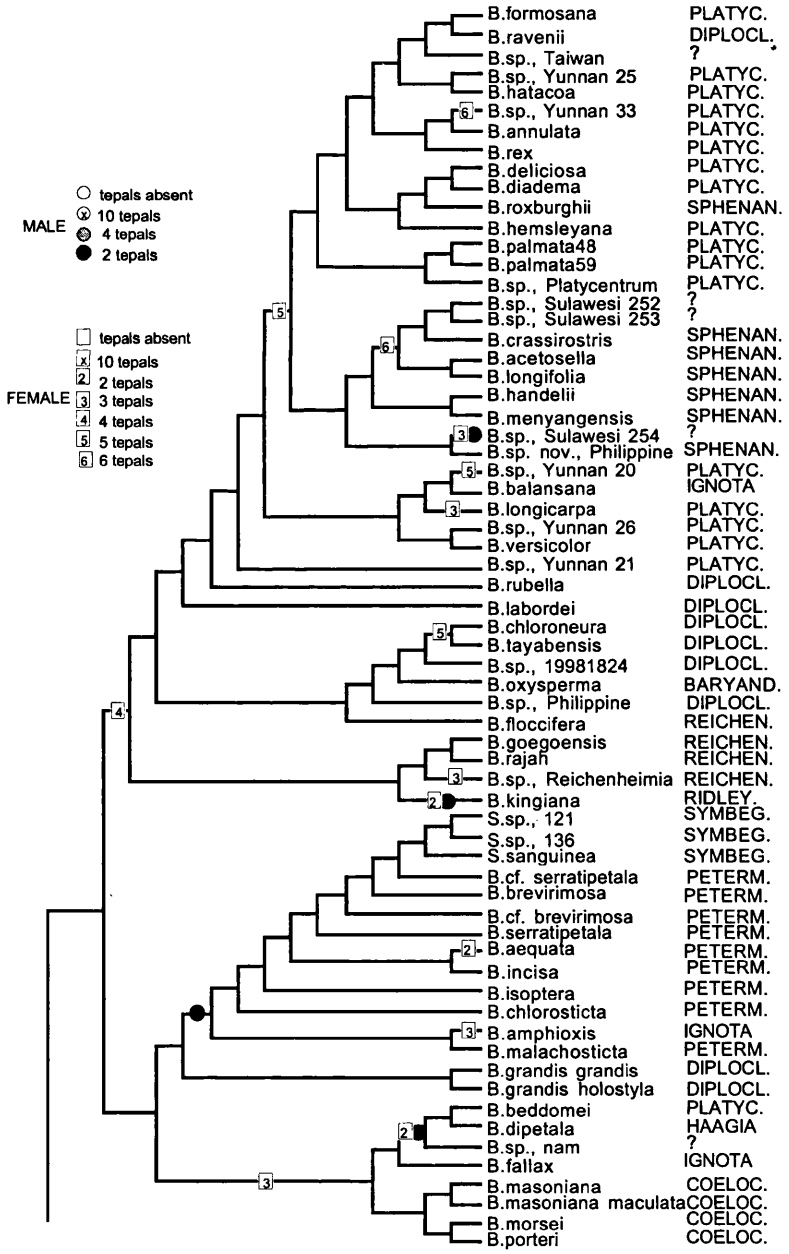


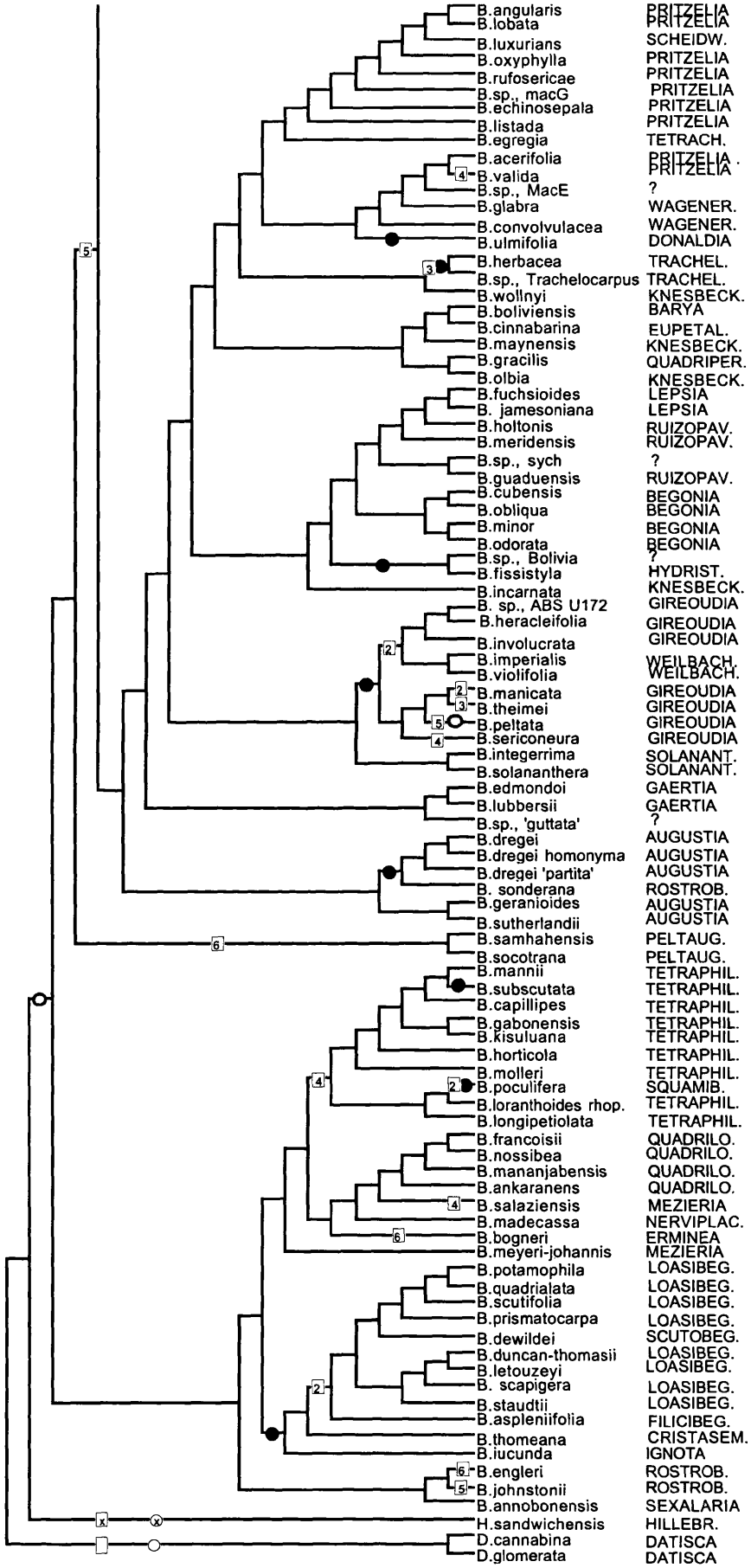

B.angularis PRITZELIA
B.lobata PRITZELIA
B.luxurians SCHEIDW.
B.oxyphylla PRITZELIA
B.rufosericae PRITZELIA
B.sp., macG PRITZELIA
B.echinosepala PRITZELIA
B.listada PRITZELIA
B.egregia TETRACH.
B.acerifolia PRITZELIA
B.valida PRITZELIA
B.sp., MacE ?
B.glabra WAGENER.
B.convolvulacea WAGENER.
B.ulmifolia DONALDIA
B.herbacea TRACHEL.
B.sp., Trachelocarpus TRACHEL.
B.wollnyi KNESBECK.
B.boliviensis BARYA
B.cinnabarina EUPETAL.
B.maynensis KNESBECK.
B.gracilis QUADRIPER.
B.olbia KNESBECK.
B.fuchsioides LEPSIA
B. jamesoniana LEPSIA
B.holtonis RUIZOPAV.
B.meridensis RUIZOPAV.
B.sp., sych ?
B.guaduensis RUIZOPAV.
B.cubensis BEGONIA
B.obliqua BEGONIA
B.minor BEGONIA
B.odorata BEGONIA
B.sp., Bolivia ?
B.fissistyla HYDRIST.
B.incarnata KNESBECK.
B. sp., ABS U172 GIREOUDIA
B.heracleifolia GIREOUDIA
B.involucrata GIREOUDIA
B.imperialis WEILBACH.
B.violifolia WEILBACH.
B.manicata GIREOUDIA
B.theimei GIREOUDIA
B.peltata GIREOUDIA
B.sericoneura GIREOUDIA
B.integerrima SOLANANT.
B.solananthera SOLANANT.
B.edmondoi GAERTIA
B.lubbersii GAERTIA
B.sp., 'guttata' ?
B.dregei AUGUSTIA
B.dregei homonyma AUGUSTIA
B.dregei 'partita' AUGUSTIA
B. sonderana ROSTROB.
B.geranioides AUGUSTIA
B.sutherlandii AUGUSTIA
B.samhahensis PELTAUG.
B.socotrana PELTAUG.
B.mannii TETRAPHIL.
B.subscutata TETRAPHIL.
B.capillipes TETRAPHIL.
B.gabonensis TETRAPHIL.
B.kisuluana TETRAPHIL.
B.horticola TETRAPHIL.
B.molleri TETRAPHIL.
B.poculifera SQUAMIB.
B.loranthoides rhop. TETRAPHIL.
B.longipetiolata TETRAPHIL.
B.francoisii QUADRILO.
B.nossibea QUADRILO.
B.mananjabensis QUADRILO.
B.ankaranens QUADRILO.
B.salaziensis MEZIERIA
B.madecassa NERVIPLAC.
B.bogneri ERMINEA
B.meyeri-johannis MEZIERIA
S
B.potamophila LOASIBEG.
B.quadrialata LOASIBEG.
B.scutifolia LOASIBEG.
B.prismatocarpa LOASIBEG.
B.dewildei SCUTOBEG.
B.duncan-thomasii LOASIBEG.
B.letouzeyi LOASIBEG.
B. scapigera LOASIBEG.
B.staudtii LOASIBEG.
B.aspleniifolia FILICIBEG.
B.thomeana CRISTASEM.
B.iucunda IGNOTA
B.engleri ROSTROB.
B.johnstonii ROSTROB.
B.annobonensis SEXALARIA
H.sandwichensis HILLEBR.
D.cannabina DATISCA
D.glomerata DATISCA

B. Tepal characters: Figure 10.15 shows characters 34 (male tepal number) and 47 (female tepal number).

Male tepal number takes 16 steps out of a possible 45 (ci 0.19; ri 0.69). Although there is homoplasy in this character, it tends to be reliable in grouping clades, e.g. the '*Petermannia*' clade (*B. malachosticta* - *Symbegonia*) has two tepals (i.e. lacks petals). Female tepal number takes 28 out of a total of 60 possible steps (ci 0.21; ri 0.59). Again, tepal number tends to be reliable between groups.

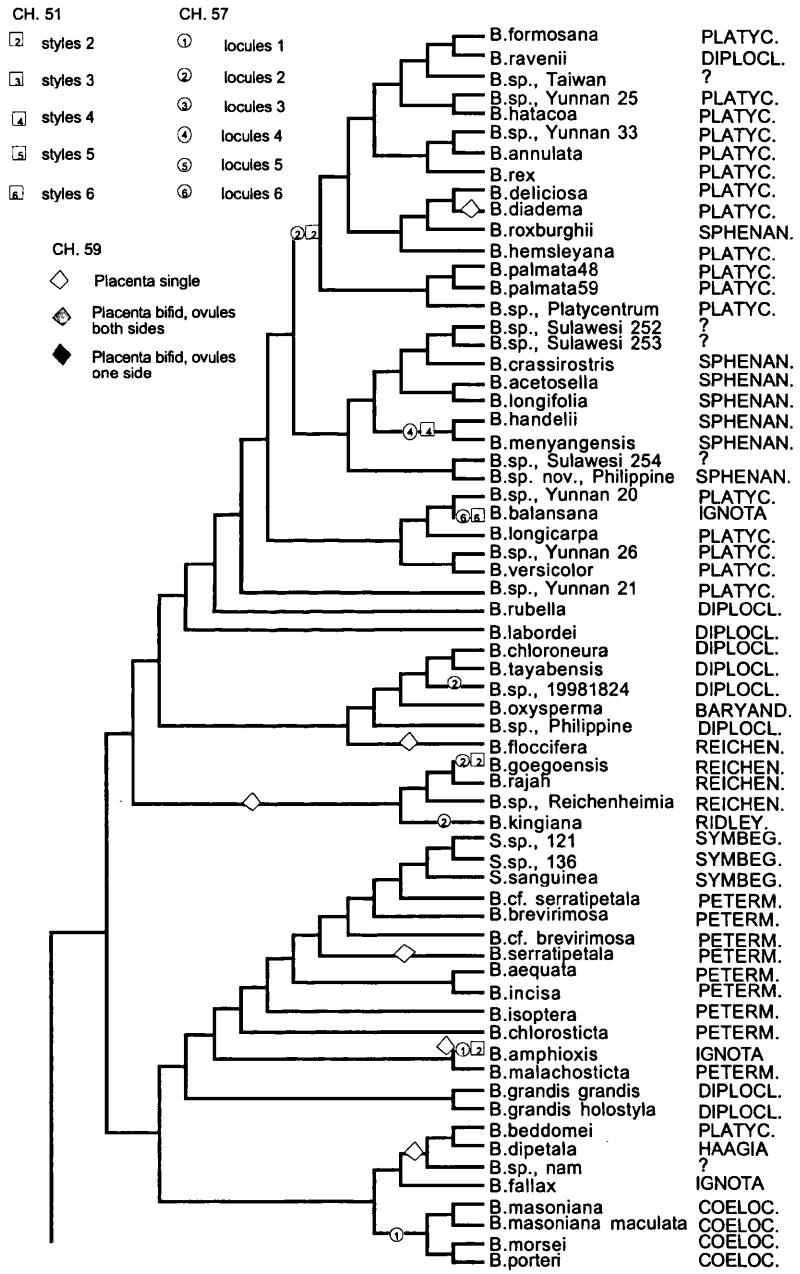
Figure 10.15: Male and female tepal number, ACCTRAN optimisation

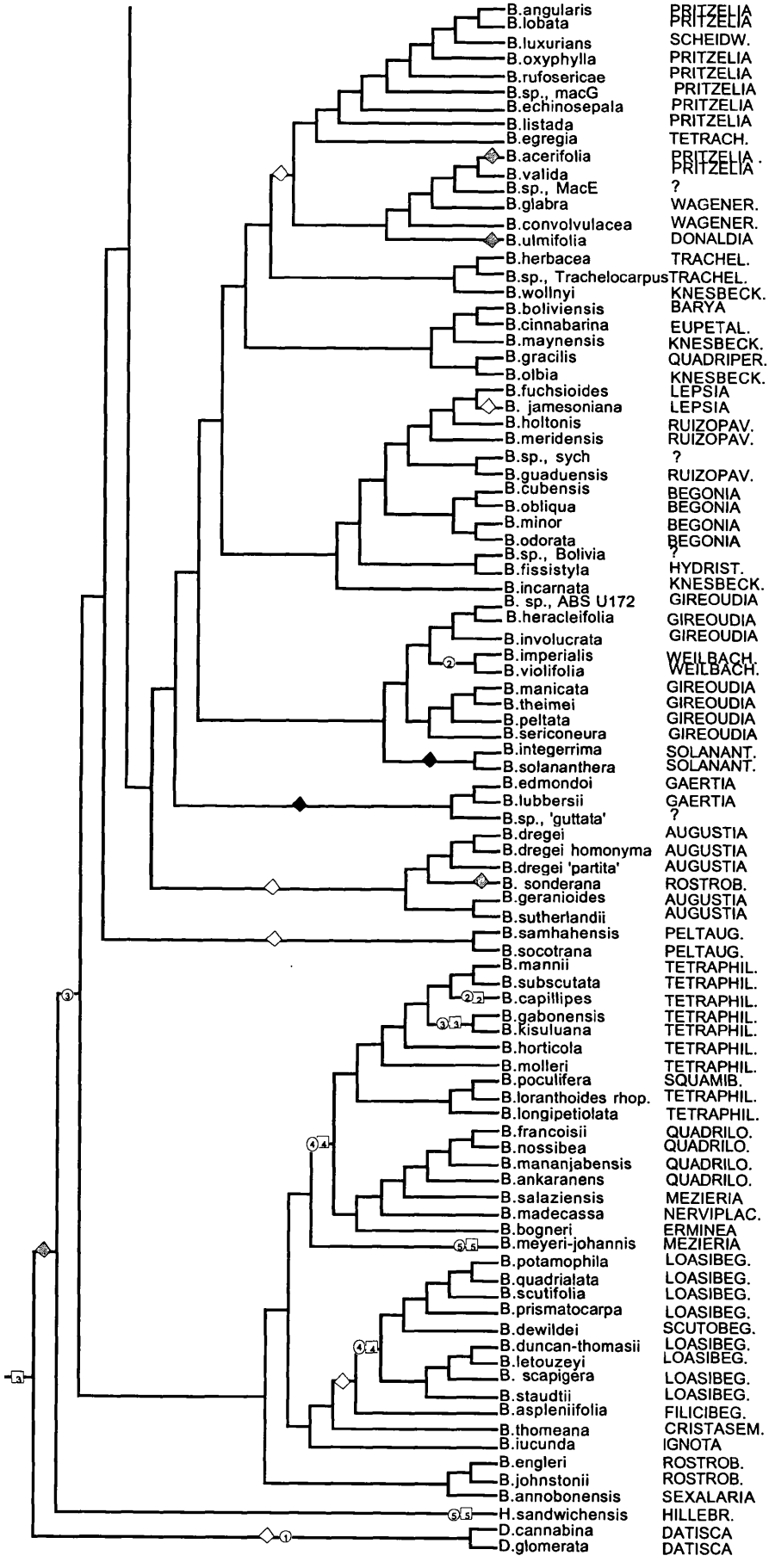



B.angularis PRITZELIA
B.lobata PRITZELIA
B.luxurians SCHEIDW.
B.oxyphylla PRITZELIA
B.rufosericae PRITZELIA
B.sp., macG PRITZELIA
B.echinosepala PRITZELIA
B.listada PRITZELIA
B.egregia TETRACH.
B.acerifolia PRITZELIA
B.valida PRITZELIA
B.sp., MacE ?
B.glabra WAGENER.
B.convolvulacea WAGENER.
B.ulmifolia DONALDIA
B.herbacea TRACHEL.
B.sp., Trachelocarpus TRACHEL.
B.wollnyi KNESBECK.
B.boliviensis BARYA
B.cinnabarina EUPETAL.
B.maynensis KNESBECK.
B.gracilis QUADRIPER.
B.olbia KNESBECK.
B.fuchsioides LEPSIA
B. jamesoniana LEPSIA
B.holtonis RUIZOPAV.
B.meridensis RUIZOPAV.
B.sp., sych ?
B.guaduensis RUIZOPAV.
B.cubensis BEGONIA
B.obliqua BEGONIA
B.minor BEGONIA
B.odorata BEGONIA
B.sp., Bolivia ?
B.fissistyla HYDRIST.
B.incarnata KNESBECK.
B. sp., ABS U172 GIREOUDIA
B.heracleifolia GIREOUDIA
B.involucrata GIREOUDIA
B.imperialis WEILBACH.
B.violifolia WEILBACH.
B.manicata GIREOUDIA
B.theimei GIREOUDIA
B.peltata GIREOUDIA
B.sericoneura GIREOUDIA
B.integerrima SOLANANT.
B.solananthera SOLANANT.
B.edmondoi GAERTIA
B.lubbersii GAERTIA
B.sp., 'guttata' ?
B.dregei AUGUSTIA
B.dregei homonyma AUGUSTIA
B.dregei 'partita' AUGUSTIA
B. sonderana ROSTROB.
B.geranioides AUGUSTIA
B.sutherlandii AUGUSTIA
B.samhahensis PELTAUG.
B.socotrana PELTAUG.
B.mannii TETRAPHIL.
B.subscutata TETRAPHIL.
B.capillipes TETRAPHIL.
B.gabonensis TETRAPHIL.
B.kisuluana TETRAPHIL.
B.horticola TETRAPHIL.
B.molleri TETRAPHIL.
B.poculifera SQUAMIB.
B.loranthoides rhop. TETRAPHIL.
B.longipetiolata TETRAPHIL.
B.francoisii QUADRILO.
B.nossibea QUADRILO.
B.mananjabensis QUADRILO.
B.ankaranens QUADRILO.
B.salaziensis MEZIERIA
B.madecassa NERVIPLAC.
B.bogneri ERMINEA
B.meyeri-johannis MEZIERIA
B.potamophila LOASIBEG.
B.quadrialata LOASIBEG.
B.scutifolia LOASIBEG.
B.prismatocarpa LOASIBEG.
B.dewildei SCUTOBEG.
B.duncan-thomasii LOASIBEG.
B.letouzeyi LOASIBEG.
B. scapigera LOASIBEG.
B.staudtii LOASIBEG.
B.aspleniifolia FILICIBEG.
B.thomeana CRISTASEM.
B.iucunda IGNOTA
B.engleri ROSTROB.
B.johnstonii ROSTROB.
B.annobonensis SEXALARIA
H.sandwichensis HILLEBR.
D.cannabina DATISCA
D.glomerata DATISCA

C. Ovary characters: Figure 10.16 shows characters 51 (style number), 57 (locule number) and 59 (placentation number).

Figure 10.16: Ovary characters, ACCTRAN optimisation




B.angularis PRITZELIA
B.lobata PRITZELIA
B.luxurians SCHEIDW.
B.oxyphylla PRITZELIA
B.rufosericae PRITZELIA
B.sp., macG PRITZELIA
B.echinosepala PRITZELIA
B.listada PRITZELIA
B.egregia TETRACH.
B.acerifolia PRITZELIA
B.valida PRITZELIA
B.sp., MacE ?
B.glabra WAGENER.
B.convolvulacea WAGENER.
B.ulmifolia DONALDIA
B.herbacea TRACHEL.
B.sp., Trachelocarpus TRACHEL.
B.wollnyi KNESBECK.
B.boliviensis BARYA
B.cinnabarina EUPETAL.
B.maynensis KNESBECK.
B.gracilis QUADRIPER.
B.olbia KNESBECK.
B.fuchsioides LEPSIA
B. jamesoniana LEPSIA
B.holtonis RUIZOPAV.
B.meridensis RUIZOPAV.
B.sp., sych ?
B.guaduensis RUIZOPAV.
B.cubensis BEGONIA
B.obliqua BEGONIA
B.minor BEGONIA
B.odorata BEGONIA
B.sp., Bolivia ?
B.fissistyla HYDRIST.
B.incarnata KNESBECK.
B. sp., ABS U172 GIREOUDIA
B.heracleifolia GIREOUDIA
B.involucrata GIREOUDIA
B.imperialis WEILBACH.
B.violifolia WEILBACH.
B.manicata GIREOUDIA
B.theimei GIREOUDIA
B.peltata GIREOUDIA
B.sericoneura GIREOUDIA
B.integerrima SOLANANT.
B.solananthera SOLANANT.
B.edmondoi GAERTIA
B.lubbersii GAERTIA
B.sp., 'guttata' ?
B.dregei AUGUSTIA
B.dregei homonyma AUGUSTIA
B.dregei 'partita' AUGUSTIA
B. sonderana ROSTROB.
B.geranioides AUGUSTIA
B.sutherlandii AUGUSTIA
B.samhahensis PELTAUG.
B.socotrana PELTAUG.
B.mannii TETRAPHIL.
B.subscutata TETRAPHIL.
B.capillipes TETRAPHIL.
B.gabonensis TETRAPHIL.
B.kisuluana TETRAPHIL.
B.horticola TETRAPHIL.
B.molleri TETRAPHIL.
B.poculifera SQUAMIB.
B.loranthoides rhop. TETRAPHIL.
B.longipetiolata TETRAPHIL.
B.francoisii QUADRILO.
B.nossibea QUADRILO.
B.mananjabensis QUADRILO.
B.ankaranens QUADRILO.
B.salaziensis MEZIERIA
B.madecassa NERVIPLAC.
B.bogneri ERMINEA
B.meyeri-johannis MEZIERIA
B.potamophila LOASIBEG.
B.quadrialata LOASIBEG.
B.scutifolia LOASIBEG.
B.prismatocarpa LOASIBEG.
B.dewildei SCUTOBEG.
B.duncan-thomasii LOASIBEG.
B.letouzeyi LOASIBEG.
B. scapigera LOASIBEG.
B.staudtii LOASIBEG.
B.aspleniifolia FILICIBEG.
B.thomeana CRISTASEM.
B.iucunda IGNOTA
B.engleri ROSTROB.
B.johnstonii ROSTROB.
B.annobonensis SEXALARIA
H.sandwichensis HILLEBR.
D.cannabina DATISCA
D.glomerata DATISCA

Style number takes 11 out of a possible 29 steps (ci 0.36; ri 0.72) and locule number takes 14 out of 36 possible steps (ci 0.29; ri 0.69). These two characters are quite strongly correlated; most taxa have the same numbers of locules as styles. However, there are some discrepancies, e.g. species from section *Coelocentrum* (*B. morsei*, *B. porteri* and *B. masoniana*) have one locule and three styles; those from section *Weilbachia* (*B. imperialis*, *B. violifolia*) have two locules and three styles.

The number of placentae takes 16 out of 37 possible steps (ci 0.13; ri 0.60). The change from two placental branches to one appears to have occurred independently in several lineages including *Filicibegonia/Loasibegonia/ Scutobegonia* (*B. aspleniifolia* - *B. potamophila*), *Peltaugustia* (*B. socotrana* and *B. samhahensis*), *Augustia* (*B. sutherlandii* - *B. dregei*) and *Reichenheimia* (*B.* sp., *Reichenheimia* - *B. goegoensis*).

10.4 Micromorphology - congruence with other data sets

10.4.1 Introduction

One way to test novel phylogenetic hypotheses is to compare them to other information, which has not been used in their generation. A commonly used example is the mapping of chromosome counts across a cladogram to see whether particular numbers support any of the clades (see Chapter 11).

Botanical literature offers a wealth of morphological and micromorphological characters; although these are not always initially discussed in a phylogenetic context, once we have a phylogeny it is possible to map these characters across it, to see whether they are useful in delimiting groupings or not (and whether some of the clades reconstructed by the cladogram need reconsidered in the light of the new information).

There are several papers concerned with micromorphological characters of *Begonia*, including detailed trichome structure (Bona & Alquini, 1995; Shui, Li & Huang, 1999), stem anatomy (Carlquist, 1985; Lee, 1974), anther endothecial cells (Tebbitt & MacIver, 1999), stigmatic papillae (Panda & de Wilde, 1995), seed (Bouman & de Lange, 1982, 1983; Keraudren-Aymonin, 1983; de Lange, 1988; de Lange & Bouman, 1986, 1992, 1999; Seitner, 1972) and pollen (van

den Berg, 1983, 1985). The characters concerned have not been coded and included in the non-DNA analysis, largely due to time factors, but also because the authors of these studies have been working with different taxa; this would violate the 'all data from the same individual' rule followed in this study.

In the light of the ITS cladograms produced here, however, previous conclusions about some of these micromorphological characters can be reexamined.

10.4.2 The Data Sets

10.4.2.1 Anther endothecial cells: The endothecium is a subepidermal layer of the anther, and usually possesses lignified or cellulose thickenings (Tebbitt & MacIver, 1999). Tebbitt and MacIver (1999) found six classes of endothecial patterns within the 173 species of Begoniaceae they looked at; those relevant to this study are U-shaped, perforate base plate, tympanate base plate, and endothecial thickening absent. They could find no correlation between the endothecial pattern and anther morphology or dehiscence (Tebbitt & MacIver, 1999).

10.4.2.2 Stigmatic papillae: Panda and de Wilde (1995) looked in detail at the stigmatic papillae of 65 species of *Begonia*, and at whether the styles are dry or wet. They found that all species of Begoniaceae (from the genera *Begonia* and *Symbegonia*; *Hillebrandia* was not available) have unicellular stigmatic papillae, and suggest that this reinforces the monophyly of the group.

10.4.2.3 Seed: The seed in Begoniaceae is characterised by the presence of specialised testa cells known as collar cells. These are longitudinally stretched cells which form a transverse ring around one end of the seed, and have not been found in any other angiosperm family. On germination, the walls between the seed lid and the collar cells split along the middle lamellae, and the walls between the collar cells split (Bouman & de Lange, 1983).

Hillebrandia seed have a rather irregular border between the collar cells and the seed lid. *Datisca* seeds are broadly similar to those of Begoniaceae, and also germinate by a seed lid which tears off along the middle of the lamellae, but they have no collar cells (Bouman & de Lange, 1983). *Begonia* seeds are almost always ellipsoid, and normally straight, although curved seeds are found in the American sections *Solananthera* and *Begonia* (Bouman & de

Lange, 1983). They usually measure between 300 μm and 600 μm long (de Lange & Bouman, 1999).

Most of the species in the genus *Begonia* are anemochorous. De Lange and Bouman (1999) suggest three adaptations to wind dispersal:

1. increased surface area to volume ratio
2. decreased mass, e.g. with air filled testa cells
3. promoted laminar airflow (microturbulance) due to surface roughness.

Species with special adaptations to wind dispersal are mostly climbers and/or epiphytes, e.g. sections *Wageneria* and *Solananthera*. Rain is also utilised as a disperser. The African sections *Filicibegonia*, *Loasibegonia* and *Scutobegonia* have indehiscent fruits which rot, and the seeds are thought to be carried in rain-wash. The Asian section *Platycentrum* has fruits which are adapted to rain ballistics; the two shorter wings of the recurved fruits form a cup to catch raindrops. Within the fleshy-fruited African sections *Baccabegonia*, *Mezieria*, *Squamibegonia* and *Tetraphila*, the seed are relatively large, and may possess arils (de Lange, 1988). These differences in fruit and seed structure appear to relate directly to habitat and dispersal.

African *Begonia* show the "greatest diversity in type of seed dispersal, especially in view of the relatively small number of species" (approximately 140 out of the global total of c. 1400 species) (de Lange & Bouman, 1992). They include the largest and smallest *Begonia* seeds (*B. ebolowensis* Engler, mean length 2240 μm; *B. iucunda*, mean length 220 μm) (de Lange & Bouman, 1999).

10.4.2.4 Pollen: Van den Berg (1983) conducted a study on the pollen types of the three genera of the Begoniaceae. He recognised three types, '*Hillebrandia*-type' (resembling some *Begonia*-type pollen), '*Symbegonia*-type' (in an "isolated position compared to the other types within the family", p. 59, although in an "extremely derived position", p. 64) and '*Begonia*-type' (based on *B. oxyloba*, *B. johnstonii*, *B. quadrialata* and *B. ampla*, although he felt that there may be more types within the genus). Van den Berg went on to look at the pollen of African *Begonia* (1985). He split the genus into several pollen types.

10.4.2 Results

10.4.3.1 Anther endothecial cells: From the list below (from Tebbitt & MacIver, 1999), of endothecial thickening-type for the species included in the ITS analysis, it can be seen that most species have U-shaped thickenings; the exceptions are the American section *Solananthera*, the Asian section *Petermannia* (which clearly should include *B. amphioxis*) and the related genus, *Symbegonia*, and the isolated Socotran section *Peltaugustia*. The other two species which have perforate-tympanate base plates (marked with *) also have U-shaped plates.

Absent:	*Symbegonia*: *S. sanguinea*, *S.* sp.
Perforate-tympanate:	*Peltaugustia*: *B. socotrana*; *B. loranthoides* ssp. *rhopalocarpa**, *B. prismatocarpa**
Perforate:	*Petermannia*: *B. brevirimosa*, *B. chlorosticta*, *B. incisa*, *B. isoptera*, *B. malachosticta*, *B. serratipetala*; Ignota: *B. amphioxis*; *Solananthera*: *B. intergerrima*, *B. solananthera*
U-shaped	*H. sandwichensis*; Africa: *B. prismatocarpa**, *B. staudtii*, *B. meyeri-johannis*, *B. salaziensis*, *B. dregei*, *B. geranioides*, *B. sutherlandii*, *B. johnstonii*, *B. sonderana*, *B. annobonensis*, *B. mannii*, *B. molleri*, *B. squamulosa*, *B. loranthoides* ssp *rhopalocarpa**; America: *B. obliqua*, *B. ulmifolia*, *B. involucrata*, *B. manicata*, *B. theimei*, *B. fissistyla*, *B. foliosa*, *B. angularis*, *B. lobata*, *B. rufosericae*, *B. gracilis*, *B. holtonis*, *B. luxurians*, *B. egregia*, *B. herbacea*, *B. convolvulacea*, *B. glabra*; Asia: *B. masoniana*, *B. grandis*, *B. tayabensis*, *B. dipetala*, *B. annulata*, *B. deliciosa*, *B. diadema*, *B. hatacoa*, *B. hemsleyana*, *B. versicolor*, B. *floccifera*, *B. goegoensis*, *B. kingiana*, *B. acetosella*, *B. handelii*, *B. longifolia*, *B. roxburghii*.

10.4.3.2 Stigmatic papillae: Within the genus *Begonia* there appears to be no phylogenetic pattern to the distribution of the five categories of stigmatic papillae type Panda and de Wilde (1995) recognised; they do not match traditional taxonomy, geographic distribution, or the 26S, ITS and *trnC - trnD* phylogenies presented here, with the exceptions of types IV (clavate, only in section *Weilbachia*) and V (lageniform, only in section *Solananthera*).

10.4.3.4 Seed: De Lange and Bouman (1992) subdivided the species they examined into categories based on seed micromorphology. They found three major groups, which included ten smaller classes, each of which included several types. The species included in this thesis fall into these categories:

1. '*Augustia*' type	a.	*B. dregei, B. homonyma*
	b.	*B. geranioides*
	c.	*B. annobonensis, B. johnstonii*
	d.	*B. sonderana*
	e.	*B. sutherlandii*
	f.	*B. engleri*
2. '*Peltaugustia*' type		*B. socotrana*
3. '*Cristasemen*' type		*B. thomeana*
4. '*Filicibegonia*' type	a.	*B. aspleniifolia*
	b.	*B. iucunda*
5. '*Scutobegonia/Loasibegonia*'	a.	*B. potamophila, B. scutifolia*
	d.	*B. hirsutula*
	e.	*B. quadrialata, B. prismatocarpa, B. scapigera, B. staudtii*
6. '*Mezieria*' type	a.	*B. salaziensis*
	d.	*B. meyeri-johannis*
8. '*Squamibegonia*' type		*B. poculifera*
9. '*Tetraphila*' type	a.	*B. mannii, B. horticola, B. subscutata, B. molleri*
	b.	*B. loranthoides* ssp *rhopalocarpa*
	c.	*B. capillipes*
	d.	*B. gabonensis*
	e.	*B. squamulosa, B. kisuluana, B. longipetiolata*
10. Madagascar	a.	*B. bogneri*
	e.	*B. nossibea*
	g.	*B. ankaranensis, B. francoisii, B. madecassa, B. mananjebensis*

The major sectional groupings in Africa, according to de Lange and Bouman (1992), are:

A. *Mezieria*, *Baccabegonia*, *Squamibegonia*, *Tetraphila*
B. *Augustia*, *Sexalaria*, *Rostrobegonia*
C. *Filicibegonia*, *Scutobegonia*, *Loasibegonia*

Within America (de Lange & Bouman, 1999), where there are c. 600 recognised species of *Begonia*, there is rather less quantifiable diversity in seed structure. A few sections show "a special seed structure characteristic at the sectional level" (p. 24). All these sections have restricted geographical distributions. They are:

Brazilian: *Trachelocarpus* (*B. herbacea*)
Solananthera (*B. solananthera, B. integerrima*)
Scheidweileria (*B. luxurians*)
Wageneria (*B. glabra, B. convolvulacea*)
Trendelbergia

Andean:	*Casparya*
	Gobenia
	Hydristyles
	Rossmannia
	Warburgina
Central Am., Mexico, Caribbean:	*Urniforma*

Species names where given are those which were examined by de Lange & Bouman, 1992, which are included in the ITS analysis.

10.4.3.4 Pollen: Pollen types (van den Berg, 1985) of relevance to the species examined in this thesis are:

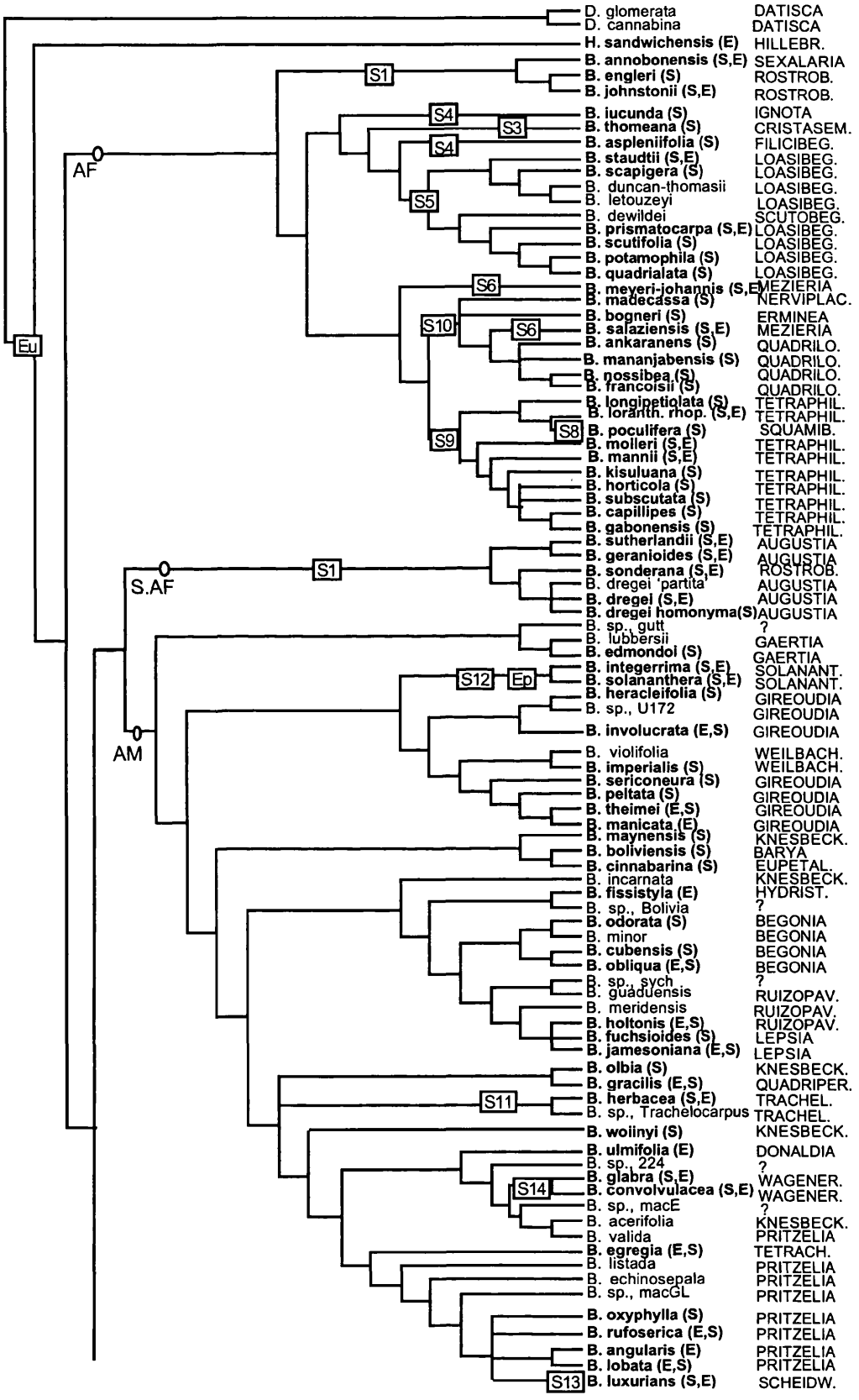
1.	*comorensis*-type	*B. meyeri-johannis*
3.	*thomeana*-type	*B. thomeana*
5.	*eminii*-type	*B. horticola, B. loranthoides, B. mannii, B. molleri*
6.	*komorensis*-type	*B. subscutata, B. kisuluana*
7.	*cavallyensis*-type	*B. capillipes*
8.	*squamulosa*-type	*B. squamulosa, B. crassipes*
10.	*poculifera*-type	*B. poculifera*
12.	*annobonensis*-type	*B. annobonensis*
13.	*dregei*-type	*B. dregei, B. engleri, B. geranioides, B. homonyma, B. johnstonii, B. partita, B. socotrana, B. sonderana, B. sutherlandii*
14.	*filicifolia*-type	*B. aspleniifolia, B. iucunda*
15.	*quadrialata*-type	*B. hirsutula, B. potamophila, B. prismatocarpa, B. quadrialata, B. scapigera, B. scutifolia, B. staudtii*

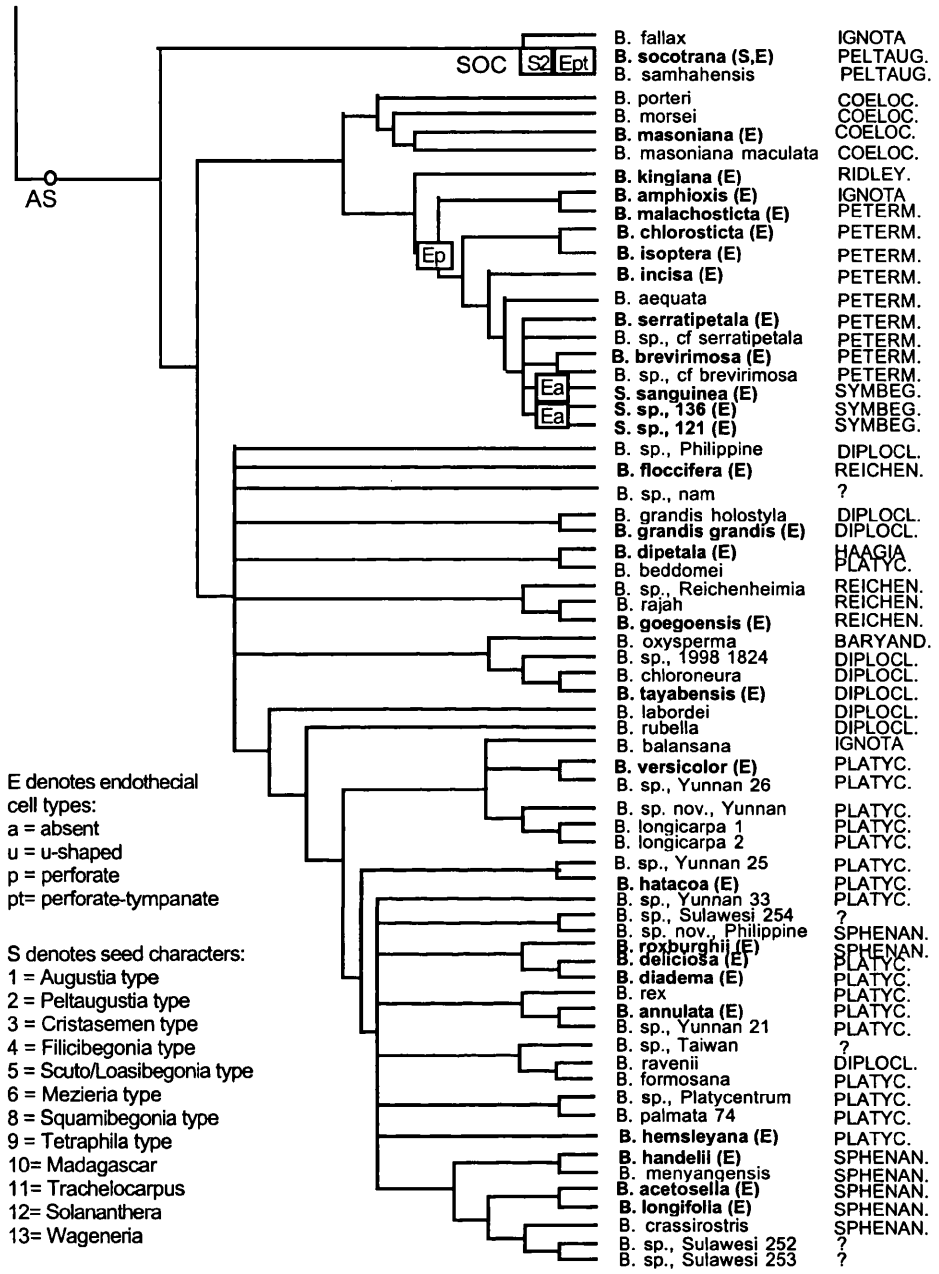
Within Madagascar, van den Berg found no distinct groups of pollen types.

10.4.3.5 Mapping the characters

In order to trace how well the characters described above fit the ITS phylogeny of Begoniaceae, the topology collated in Chapter 7 as the 'Jigsaw' tree (Figure 7.22) was taken and endothecial cell types and seed types were mapped across it (see Figure 10.17). Pollen type was not mapped because most of the characters are uninformative; also, van den Berg has only looked at African species and so his groups are only relevant to a subsection of the ITS tree. Stigmatic papillae were not mapped because it was patently clear that the classes Panda and de Wilde described were homoplastic.

Figure 10.17: ITS phylogeny, with endothecial cell types and seed types





Species in this phylogeny which were mentioned in the micromorphology papers are highlighted in bold. Those in Tebbitt & McIver (1999) have an 'E' after the species name; those mentioned in de Lange & Bouman (1992, 1999) have an 'S'.

10.4.4 Discussion:

10.4.4.1 Anther endothecial cells: It appears that U-shaped plates are phylogenetically basal in the Begoniaceae, being shared by *Hillebrandia* and by all African taxa examined. There have been changes in endothecial cell types in three independent lineages - that leading to section *Petermannial Symbegonia*, that leading to section *Solananthera*, and that leading to section *Peltaugustia*. Perforate cell plates have evolved more than once, so this is not a character without homoplasy, but it certainly reinforces the (already very

clearly defined) lineages of *Petermannia*, *Solananthera* and *Peltaugustia.*

10.4.4.2 Seed: The seed micromorphology of Asian species has not yet been comprehensively surveyed, although *Symbegonia* seed were examined by Bouman and de Lange (1983); they are very small, and mostly consist of collar cells, with a general morphology which agrees "fully with the general seed characters of the genus *Begonia*" (p. 78). Seitner (1972) also included some Asian species in his observations, and De Lange and Bouman (1999) have looked at the seed of some *Diploclinium* and some *Sphenanthera* species; all appear to belong to the "ordinary *Begonia* seed type" (de Lange & Bouman, 1999, p. 26).

From the data here, most of the seed types fit the cladogram (and conventional *Begonia* groupings) well. In America, where most seed of most species is apparently fairly uniform, the sections *Solananthera* and *Trachelocarpus* are picked out. These sections are highly distinct and already well characterised by many characters. The far less distinct sections *Scheidweileria* and *Wageneria*, both of which appear within section *Pritzelia*, are also picked out. Species in section *Wageneria* are among the most widely distributed American *Begonia*; the section is found throughout Central and South America (although absent on the Guianas) (Doorenbos, Sosef & de Wilde, 1998). Perhaps their seed are particularly well adapted to dispersal.

The homoplastic seed types are '*Filicibegonia*-type' and '*Mezieria*-type', although if '*Cristasemen*-type' and '*Scutol/Loasibegonia*-type' were modifications of '*Filicibegonia*-type', then there would be no inconsistency. '*Mezieria*-type' is found in *B. meyeri-johannis* and *B. salaziensis.* This suggests that it may be premature to split this section on the results of ITS analysis, as there is data which suggests that it may be monophyletic after all (or alternatively de Lange and Bouman may have, in the absence of major differences, assigned a common seed type based on the existing taxonomies).

10.4.4.3 Pollen: Van den Berg (1985) put his pollen types into an evolutionary context; he found that sections *Mezieria*, *Baccabegonia*, *Cristasemen* and *Filicibegonia* have the most primitive size, shape and endoaperture; Madagascan species also have a "relatively low evolutionary level" (p. 82). From these basic types, "the developments have taken place in a number of directions, sometimes diverging, sometimes converging" (p. 67).

10.5 Discussion

Within the limited time-frame of this project, molecular analyses give an evolutionary scenario which appears more congruent with gross morphology (and with previous sectional treatments) than a straight cladistic analysis based on non-DNA characters alone. The problems with the characters which hold together some of the clades in the non-DNA analysis are obvious, and they have also proved problematic in previous studies.

Tebbitt (1997) produced morphological cladograms which split the section he was studying, section *Sphenanthera*, into at least three clades. Several of the taxa resolved as sister to the African sections *Tetraphila* and *Mezieria* (others resolved within the Asian sections *Petermannia* and *Platycentrum*). Species from sections *Tetraphila*, *Mezieria* and *Sphenanthera* all have fleshy, indehiscent, frequently wingless fruits, and although Tebbitt is not explicit about which characters change on each clade, it is likely that fleshiness and related characters (fruit shape and dehiscence) may be responsible for this pattern. Badcock (1998) obtained a similar fleshy-fruited clade from her morphological analyses, which included species from Asia (*B. roxburghii*), America (*B. oacacana*) and Africa (*B. salaziensis*, *B. poculifera*, *B. meyeri-johannis* and *B. mannii*). Whether the fruit is dry or fleshy is strongly correlated with the mode of seed dispersal, as fleshy fruits are associated with zoochory and dry fruits, with wind dispersal. In a genus the size of *Begonia*, it does not seem improbable that unrelated species have evolved similar adaptations to seed dispersal, and therefore that a clade based on this character may well be the result of convergence or parallel adaptation. Similar groupings occur in the non-DNA cladogram which I have produced, with a clade of fleshy-fruited African and Asian species. Clearly, reassessing the homology of this character may reveal different types of fleshiness, although such anatomical work was not conducted as part of this analysis.

Morphological character evolution makes more sense when reconstructed over a combined analysis of non-DNA and ITS sequence data (Figures 10.11, 10.12). Although there is homoplasy in almost all characters, they are still useful for grouping clades within the genus. This locally informative nature of morphological characters has been documented in other plant groups, e.g. Pennington, 1995, in *Andira* Juss. (Fabaceae).

The presence or absence of a trichome ring at the top of the petiole (see Figure 10.14) has been used to distinguish between sections *Augustia* (absence: *B. sutherlandii* - *B. dregei*) and *Rostrobegonia* (presence: *B. engleri*, *B. johnstonii* and *B. sonderana* are included here) (Irmscher, 1961; Doorenbos, Sosef & de Wilde, 1998). One exception to this group is *B. sonderana*, which although traditionally placed in section *Rostrobegonia* on the basis of unbranched placenta, is resolved in molecular and combined analyses in *Augustia*. In this respect it is noteworthy that *B. sonderana* is one of only two *Rostrobegonia* species quoted by Doorenbos, Sosef and de Wilde as being "without a tuft of hairs" (p. 176); the presence of a tuft of hairs is congruent with its phylogenetic placement in *Augustia*. The sharing of *Augustia* characters (the tuft of hairs) and *Rostrobegonia* characters (bifid placentae) lends some support to the comment from Doorenbos, Sosef and de Wilde that *Rostrobegonia* is "closely related to sect. *Augustia* and possibly identical with it" (p. 178). However, in the molecular and combined analyses these sections are well separated. (It is interesting that the character which misleads in this case is that of placental branch number, which has "always played an important role in the classification of *Begonia*" (Doorenbos, Sosef & de Wilde, 1998, p. 28).)

Within sections *Loasibegonia* and *Scutobegonia* the character state 'peltateness' is variable (see Figure 10.14), in that it occurs both with the point of insertion in the centre of the leaf in some taxa, and with a highly asymmetric point of insertion in others. There does not appear to be any clear trend to this, as the species with the least asymmetric insertion, *B. scapigera*, is well within the clade.

Given the anatomical division of *Begonia* tepals into petals and sepals, it is interesting to see that, for the male flower, almost every change on the tree is caused by the loss of two tepals (Figure 10.15). Only on one occasion (*B. peltata*) is there a subsequent reversal back to four. If this result holds up to further scrutiny, it would be desirable to look at the vascularisation of the tepals in *B. peltata*, in case they are all sepals or all petals rather than that one organ has been lost and then regained.

Female tepal numbers are far more variable, and may show increases and decreases within clades. This may correlate with a more complex relationship between petal and sepal number in the female, than the male, flower.

The number of locules in *Begonia* appear to correlate quite strongly, but not absolutely, with the number of styles. Loss of one locule and one style (from three to two) is a good defining character for *Platycentrum* (*B.* sp., *Platycentrum* - *B. formosana*) (on this topology, Figure 10.16).

Placental branch number, like locus number, has, as mentioned previously, been considered important in *Begonia* evolution. Most *Begonia* species have two branches, with ovules on both sides of them. However, there have been 11 independent cases of the loss of one of these branches over this topology (Figure 10.16) (as well as subsequent reversals back to two branches). Therefore, it is not a suitable character for *Begonia* classification if used in isolation, but rather, must be considered along with a suite of other characters.

Perhaps disappointingly, none of the characters considered here will, in isolation, split *Begonia* into distinct morphological chunks - and too high an emphasis on any one character will almost certainly lead to polyphyly (or paraphyly). Although it will be possible to reclassify the genus in such a way that monophyly is upheld, this is unlikely to be achieved using clear-cut morphological apomorphies, but instead, with suites of (sometimes overlapping) morphological characters.

10.6 Summary

The morphological characters found in *Begonia* species have been discussed, and a non-DNA data matrix constructed. This matrix has been analysed (MP) alone and in combination with an ITS sequence matrix. The ITS matrix was also analysed, in order that comparisons between tree statistics could be drawn, and areas of agreement with the non-DNA topology could be established. The combined ITS - non-DNA analysis was used to look at how well the individual morphological characters fit the tree, and a few leaf, tepal and ovary characters were reconstructed across the ITS - non-DNA analysis topology.

Cladistic analysis of molecular data have produced evolutionary scenarios which are more congruent with gross morphology (and prior sectional treatments) than analyses based on non-DNA characters.

Micromorphological data sets published by previous authors were examined to see how well the data fits the ITS phylogeny produced in a previous chapter. Seed and anther endothecial cell characters are generally congruent with the molecular phylogenies.

Figure 10.18 *Begonia* Leaves - Colour plate

a

B. aspleniifolia Hook.f. ex A.DC.
sect. *Filicibegonia*
GL 001 097 97

b

B. johnstonii Oliv. ex. Hook.f.
sect. *Rostrobegonia*
E 1999 0653

c

B. lyman-smithii Burt-Utley & Utley
sect. *Gireoudia*
GL 003 155 94

d

B. sericoneura Liebm.
sect. *Gireoudia*
GL 009 124 82

e

B. tayabensis Merr.
sect. *Reichenheimia*
GL 006 035 89

Figure 10.19: *Begonia* Inflorescences - Colour plate

10.19.1 American species

a

B. theimei C.DC ex J.D. Sm., sect. *Gireoudia*
GL 002 093 79

b

B. heracleifolia Cham. & Schlecht., sect. *Gireoudia*
GL 001 126 83

c

B. involucrata Liebm., sect. *Gireoudia*
GL 004 100 57

d

B. maynensis A.DC., sect. *Knesbeckia*
GL 001 107 92

e

B. herbacea Vell., sect. *Trachelocarpus*
E 1973 1857

10.19.2 Asian species

a

B. diadema Linden ex Rodigas
sect. *Platycentrum*
GL 001 117 97

b

B. sp., sect. *Platycentrum*;
GL 004 033 96

c

Symbegonia sanguinea Warb.
GL 003 127 93

d

B. roxburghii A.DC.,
sect. *Sphenenthera*
GL 001 068 98

10.19.3 African species

a

B. socotrana Hook.f.,
sect. *Peltaugustia*;
E 1999 0424

b

B. loranthoides Hook.f. ssp
rhopalocarpa (Warb.) J.J.deWilde
sect. *Tetraphila*; GL 030 079 97

c

B. mannii Hook., sect. *Tetraphila*
GL 008 067 80

Figure 10.20 Male *Begonia* flowers - colour plate

a

B. sp., Yunnan no. 25
sect. *Platycentrum*
E 1998 0061

b

B. handelii Irmsch.
sect. *Sphenanthera*
E 1998 0050

c

B. boisiana Gagenp.
sect. *Ignota*
GL 002 033 96

d

B. letouzeyi Sosef
sect. *Loasibegonia*
GL 027 079 97

e

B. loranthoides Hook.f.
sect. *Tetraphila*
GL 002 087 84

f

B. ampla Hook.f.
sect. *Squamibegonia*
E 1999 0258

g

B. ampla Hook.f.
sect. *Squamibegonia*
E 1999 0258

Figure 10.21 Female *Begonia* flowers - colour plate

a

S. sanguinea Warb.
GL 003 127 93

b

B. brevirimosa Irmsch.,
sect. *Petermannia*; E 1982 1108

c

B. sp., Yunnan no. 25
sect. *Platycentrum*
E 1998 0061

d

B. chlorosticta M.J.S.Sands
sect. *Petermannia*
E 001 167 94

e

B. boisiana Gagnep.
sect. *Ignota*
GL 002 033 96

f

B. molleri Warb., sect. *Tetraphila*
GL 036 079 97

g

B. herbacea Vell.
sect. *Trachelocarpus*
E 1973 1857

h

B. convolvulacea (Klotzsch) A.
DC. sect. *Wageneria*
E 1979 1884

Figure 10.22 *Begonia* Fruits - colour plate

a

B. johnstonii Oliv. ex Hook.f.
sect. *Rostrobegonia*
E 1999 0653

b

B. oxyloba Welw. ex Hook.f.
sect. *Mezieria*
E 1998 2761

c

B. hatacoa Buch.-Ham. ex
D.Don sect. *Platycentrum*
GL 001 029 97

d

B. sp., ABS U205
GL

11. Cytology and Phylogeny in *Begonia*

11.1 Introduction

Cytogenetics can offer phylogenetic information at several different levels, from genome size, through morphology (shape, size, number and behaviour of chromosomes during meiosis and mitosis) to the structural organisation of genetic information along the lengths of individual chromosomes (Sessions, 1996).

11.1.1 Karyotype: There have been very few attempts to karyotype *Begonia* chromosomes. According to Arends (1985) the chromosomes of African species are generally (sub)metacentric, with trends towards increasing asymmetry in some sections (and because Arends associates primitiveness with symmetry, he regards the less symmetric sections, particularly section *Squamibegonia*, as more advanced). Arends (1970) found that the somatic chromosomes of 'Elatior' *Begonia* hybrids (*B. socotrana* x tuberous *Begonia* hybrids (probably from crosses between Bolivian and Peruvian species)) could be separated into longer and shorter chromosomes. The longer chromosomes have been inherited from the tuberous *Begonia* hybrids used in the original crosses, and the shorter ones came from the male parent, *B. socotrana.* In general, though, the small size of *Begonia* chromosomes has led to most authors being concerned simply with counts. For instance, Legros and Doorenbos (1971), who had produced chromosome counts from 190 species of *Begonia* at this time, found that the only species where they could recognise individual chromosomes is *B. nepalensis* (A.DC.) Warb (*B. gigantea* Wall.), from section *Monopteron* (A.DC.) Warb.

11.1.2 Numbers: Several studies have examined chromosome counts for *Begonia*. This is in part because of the huge amateur interest in *Begonia* cultivars; the chromosome numbers are of interest to growers concerned with the crossability of different species and cultivars. "Crosses between species [of *Begonia*] of similar chromosome numbers are usually successful while those between groups of dissimilar numbers if successful are usually sterile" (McGregor, 1969, p. 230).

Legro and Doorenbos have been the most prolific counters of *Begonia* chromosomes (1969, 1971, 1973). They give counts for over 220 species; they found 22 different chromosome numbers, which range from 16 (in *B. nepalensis*) to 156 (*B. acutifolia* Jacq., section *Begonia*). Apparently "[t]his complicated situation is considerably clarified if the species are arranged into sections. Most sections were found to be characterised by one basic chromosome number, from which the other numbers within the section (if any) have been derived by polyploidy" (p. 167, 1973).

However, none of these studies place *Begonia* chromosome numbers (and *Begonia* sections) into a formal phylogenetic context. Although Doorenbos, Sosef and de Wilde (1998) briefly discuss chromosome numbers under each sectional heading in their revision of *Begonia* sections, this is of limited value given that at the time no phylogeny of *Begonia* was available and thus the sections may not represent monophyletic units; furthermore, no picture of the direction of evolutionary change can be drawn.

11.2 Methods

My literature review for chromosome counts in the Begoniaceae revealed 604 published counts. These represent 255 species, in 47 sections. There are also a number of counts for hybrids and/or cultivars of horticultural interest. The taxonomy of the species has been revised according to Doorenbos, Sosef and de Wilde (1998), and all the counts are presented on the accompanying CD-ROM.

The ITS 'jigsaw' tree (Chapter 7, Figure 7.22) has been used as a framework for consideration of chromosome numbers; existing counts for species in the tree have been annotated onto the figure (Figure 11.1). For tree and node support indices, refer back to Figure 7.3.

Where there are several differing counts for one taxon, the more recent counts (Legro and Doorenbos, 1969, 1971, 1973) have been preferred, because these were felt to be more reliable. Some of the earliest counts

predate the squash technique (as described in Jong, 1997), and were made by sectioning the cells. This could easily lead to errors. For instance, the work of Heitz (1927) was disregarded by Legro and Doorenbos (1969), due to "the high incidence of incorrect results" (p. 193). However, all *Begonia* chromosome counts (including suspect ones) are included in a table (14.10) presented on the CD-ROM in the interests of completeness.

Legro and Doorenbos (1969, 1971, 1973) use the symbol '+' to indicate the presence of "stainable fragments.....about a third [the size] of the smallest chromosomes" (1969, p. 193); this has been followed in the reports of their counts on the accompanying CD-ROM.

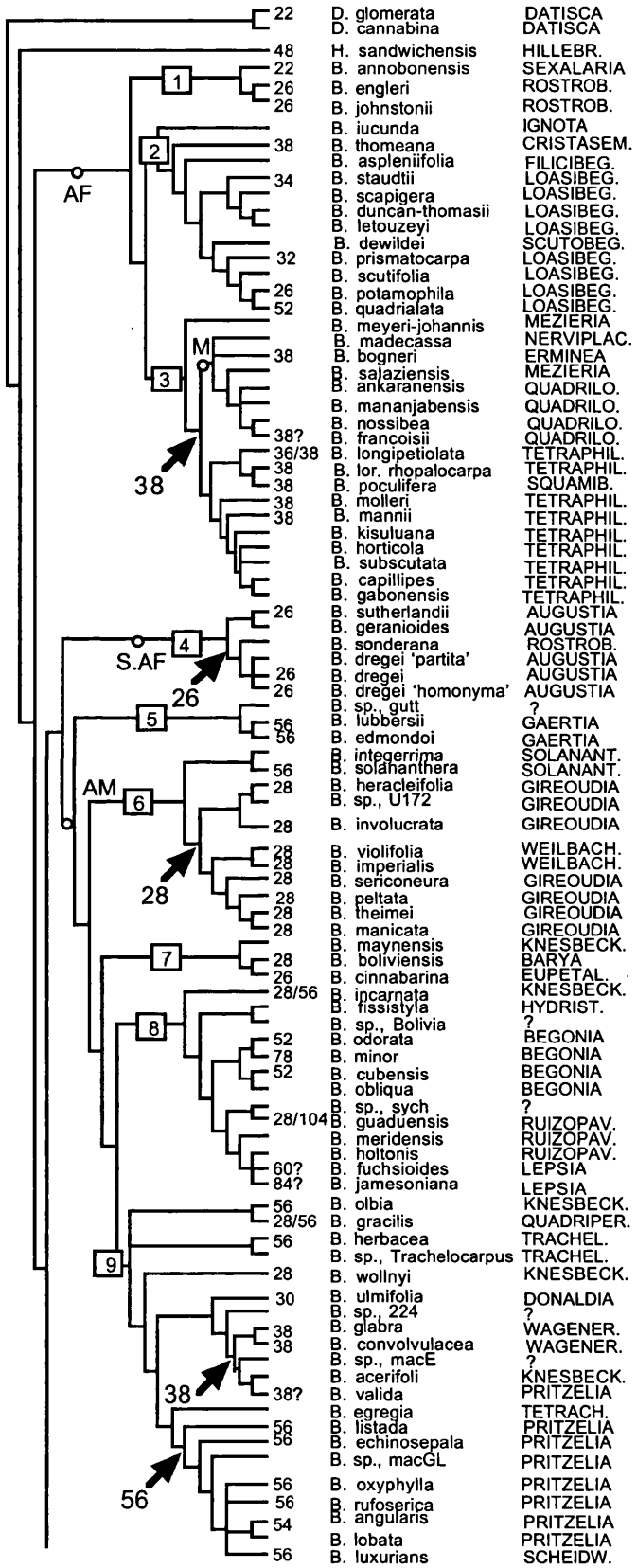
A further potential source of error is identification of material, both in the molecular and cytological studies. In the current molecular studies, voucher specimens have been deposited at E. In the course of this study it was not possible to check voucher specimens for chromosome counts (indeed, many are not supported by vouchers) and this represents a weakness in any evaluation of chromosome number evolution in *Begonia*.

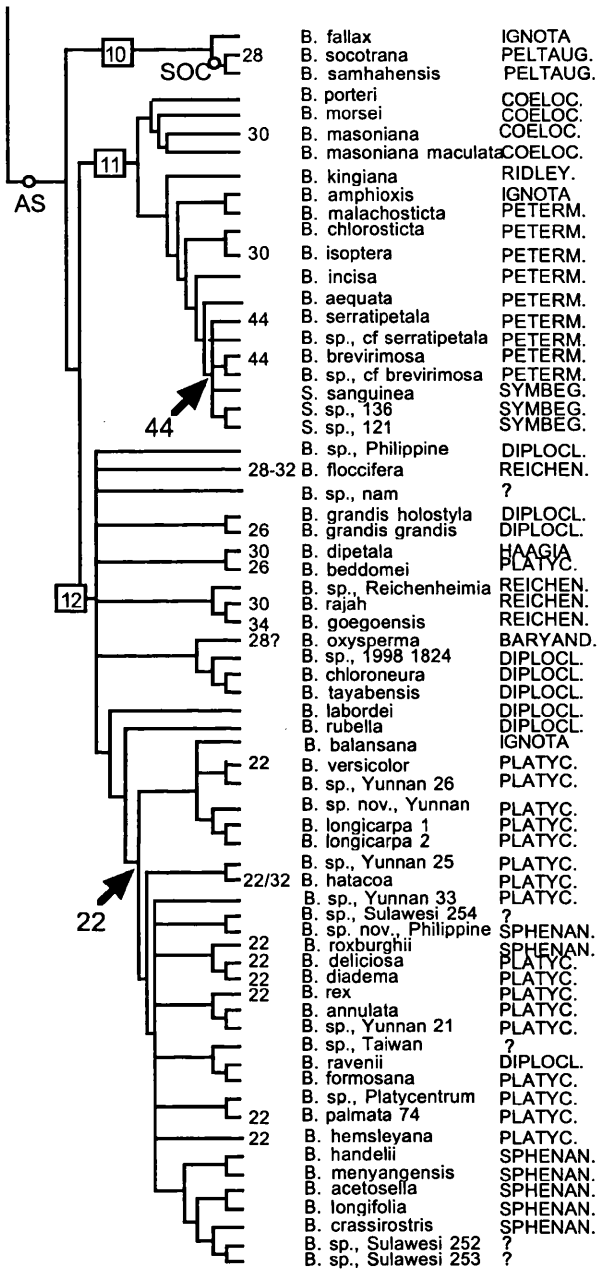
11.3 Results

See Figure 11.1, which represents the 'jigsaw' topology for ITS sequence data analysis from Chapter 7 (Figure 7.22), with chromosome numbers from the table on the CD-ROM annotated on. There are also seven arrows, which mark nodes which appear to be characterised by a particular chromosome number.

12 clades have been annotated. Of these, one to four are African, five to nine are American and ten to twelve are Asian. These will form the basis for discussion about trends in chromosome numbers.

Figure 11.1: Chromosome counts across an ITS phylogeny





☐ denotes a clade marker

11.4 Discussion

Plotting preexisting chromosome numbers onto a tree like this is a highly frustrating exercise. Patterns begin to emerge but either something does not quite fit, or a count for a crucial taxon is missing. Given that almost all the taxa used in this analysis are in cultivation in Scotland, reaffirming counts and filling in gaps is a possibility, and, given also that patterns **are** emerging, this is something which should be looked at in the future.

11.4.1 Africa

African species on this phylogeny have chromosome counts of 22 (one species), 26 (six species), 32 (one species), 34 (one species), 38 (eight species) and 52 (one species).

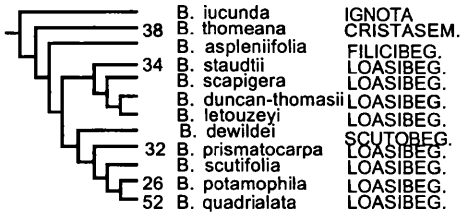
Clade 1: The counts in this clade are 22 and 26 (the lowest numbers from this continent) (see Figure 11.2). Section *Rostrobegonia*, to which the species with the counts of 26 belong, also includes species with counts of 28 (which have not been included in this ITS phylogeny); however, the section is possibly polyphyletic (one species from it resolves in clade 4) and so the placement of the species with 2n = 28 cannot be estimated.

Figure 11.2: Clade 1 (Africa)

Clade 2: The numbers in the clade are varied (see Figure 11.3); three of the counts fit the polyploid series described in Legro and Doorenbos (1969) of 2x = 26, 3x = 38 (loss of one from 39[11]), 4x = 52; however, counts of 32 and 34 do not fit this series. Obviously if there are clear patterns in this clade, more sampling is needed to reveal them.

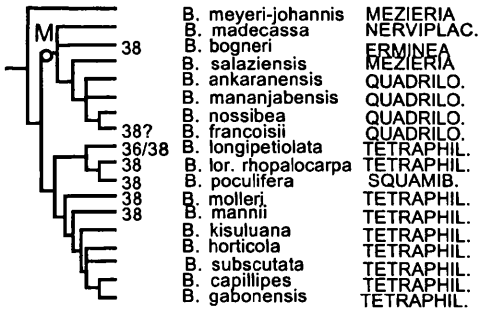
[11] The loss a chromosome usually leads to inviability in diploids; however, higher polyploids have a buffering effect, and aneuploids can survive (Heslop-Harrison, 1953).

Figure 11.3: Clade 2 (Africa)



Clade 3: This clade is apparently characterised by 2n = 38 (see Figure 11.4); however, there are tetraploids reported in some species from section *Tetraphila* (which were not sampled here) (Arends, 1992). Counts are needed for *B. meyeri-johannis* and *B. salaziensis*; the only species from section *Mezieria* which has been counted is *B. seychellensis*, which has 2n = 26 (Legro & Doorenbos, 1973); it may be that the situation in the clade containing the Madagascan and *Mezieria* species is more complex than is suggested here.

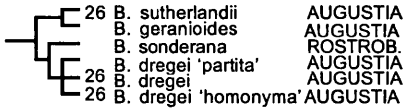
Figure 11.4: Clade 3 (Africa)



Clade 4: This clade consists of southern African species, which have resolved as basal to all the America species. This clade is apparently characterised by 2n = 26 (se Figure 11.5). Previous authors have considered the 2n = 26 species in Clade 1 to be inseparable from the species in Clade 4 - Doorenbos, Sosef & de Wilde, 1998. These taxa are well separated in this ITS phylogeny.

N.B. There is a mistake in Doorenbos, Sosef and de Wilde (1998, p. 68): "2n = 56 (*B. dregei*, *B. homonyma*, *B. princeae*)" SHOULD read 2n = 26 (pers. comm., Doorenbos, 1999).

Figure 11.5: Clade 4 (southern African)



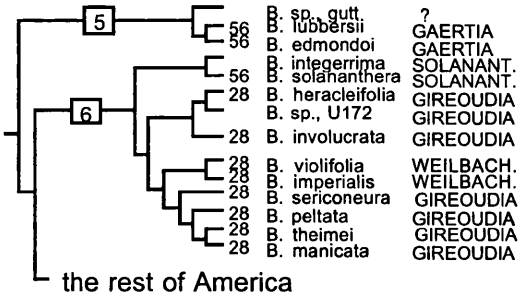
11.4.2 America:

American species on this phylogeny have chromosome counts of 26 (one species), 28 (13 species), 30 (one species), 52 (two species), 54 (one species), 56 (12 species), 60? (one species), 78 (one species), 84? (one species) and 104 (one species).

The chromosome number 2n = 28 occurs in several clades representing a broad geographical area. It is the lowest widespread number, and may be basal in America. The tetraploid 2n = 56 appears to have arisen several times; occasionally it characterises clades; at other times it appears along with the diploid number within species (and the morphology is not apparently distinguishable).

Clades 5 & 6: There are two numbers recorded from these clades, 2n = 56 (in clade 5 and part of clade 6) and 2n = 28 (see Figure 11.6). Six species from section *Gaerdtia* (clade 5) have been counted with 2n = 56 (of which, two are included in this phylogeny), while the one other species in section *Solananthera* (*B. radicans* Vell.) also has 2n = 56. The rest of this clade represents sections *Gireoudia* and *Weilbachia* (36 species from these sections have been counted, and all have 2n = 28; eight of these species are represented here). The 2n = 56 taxa probably represent tetraploids based on 2n = 28.

Figure 11.6: Clades 5 and 6 (America)

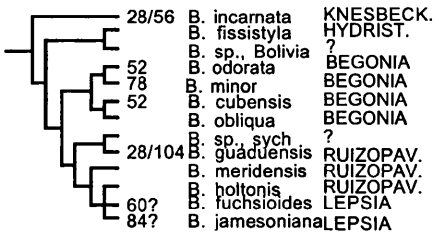


Clade 7: Counts exist for two of the taxa in this clade (see Figure 11.7). Although only *B. cinnabarina* from section *Eupetalum* is represented here, chromosome counts exist for seven species from the section; two of then have 2n = 26, the other five, like *B. boliviensis* (section *Barya*), have 2n = 28.

Figure 11.7: Clade 7 (America)

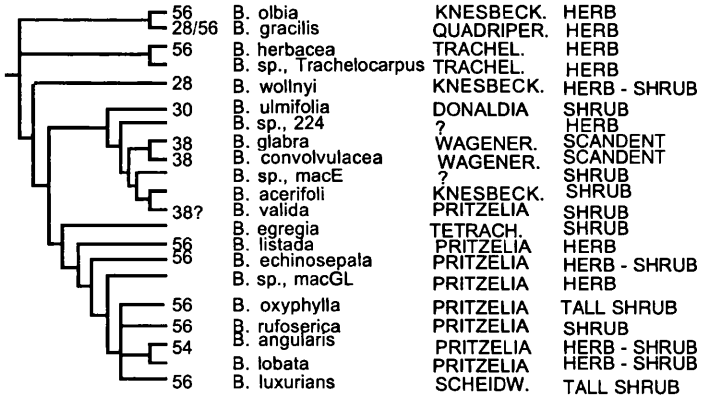
Clade 8: This clade is highly heterogeneous, with a mixture of numbers between 2n = 28 and 2n = 104 (both of which have been recorded within a single species, *B. guaduensis*) (see Figure 11.8). Section *Begonia* is included in this clade, with counts of 2n = 52 and 2n = 78. Legro and Doorenbos (1969) speculated that high and varied chromosome counts within this section may reflect the long amount of time many of the species within it have been in cultivation, and the possibility of hybridisation since cultivation. Counts from recently wild-collected material would enable this to be tested. Alternatively, this may represent a predisposition towards polyploidy within this clade.

Figure 11.8: Clade 8 (America)



Clade 9: This large clade is partially unresolved; most species have 2n = 56 (see Figure 11.9). Legros and Doorenbos commented in 1969 that the scandent/ trailing species in section *Pritzelia* are always 2n = 38, while the shrubby species are always 2n = 56. There is certainly some sort of pattern within this clade, but the clear-cut morphological correlation has been lost. The shrubby *B. ulmifolia* (section *Donaldia*), with 2n = 30, is sister to a clade which includes scandent species from section *Wageneria* (*B. glabra* and *B. convolvulacea*) and shrubby species from section *Pritzelia* (*B. valida*), which have 2n = 38. Sister to *B. ulmifolia* and the 2n = 38 clade is a clade which appears characterised by 2n = 56 (although one taxon shows 2n = 54). Species in this clade are largely shrubs; *B. luxurians* (section *Scheidweileria*) and *B. oxyphylla* (section *Pritzelia*) can grow several metres tall. However, *B. listada* is a delicate rhizomatous species (Karegeannes, 1981). Chromosome and ITS evidence do certainly suggest that some form of division of this clade is needed. The origin of the number of 2n = 38 is a bit of a mystery; Legro and Doorenbos (1969) speculate that this is part of a series 2x = 26, 3x = 38, 4x = 52; there is little evidence for this in this phylogeny with no other counts from this series recorded; instead the clade is predominantly based around the 28 - 56 series.

Figure 11.9: Clade 9 (America)



11.4.3 Asia:

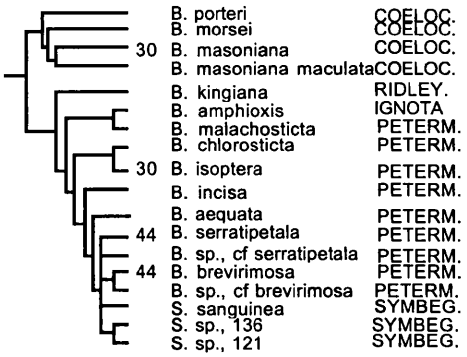
Asian species on this phylogeny have chromosome counts of 22 (eight species), 26 (two species), 28 (three species, including one from Socotra), 30 (four species), 32 (two species), 34 (one species) and 44 (two species).

Clade 10: The clade is shown in Figure 11.10. This is not exactly an 'Asian' clade *per se*; this clade, sister to all other Asian species, includes one Indian and two Socotran species (the geographic affinity of Socotra could be argued to be either Asian or African). 2n = 28, the count for *B. socotrana*, is the most frequent count on this phylogeny and therefore does not allow us any speculation about one of the more isolated and unusual taxa (section *Peltaugustia*) in *Begonia*.

Figure 11.10: Clade 10 (Asia / Socotra)

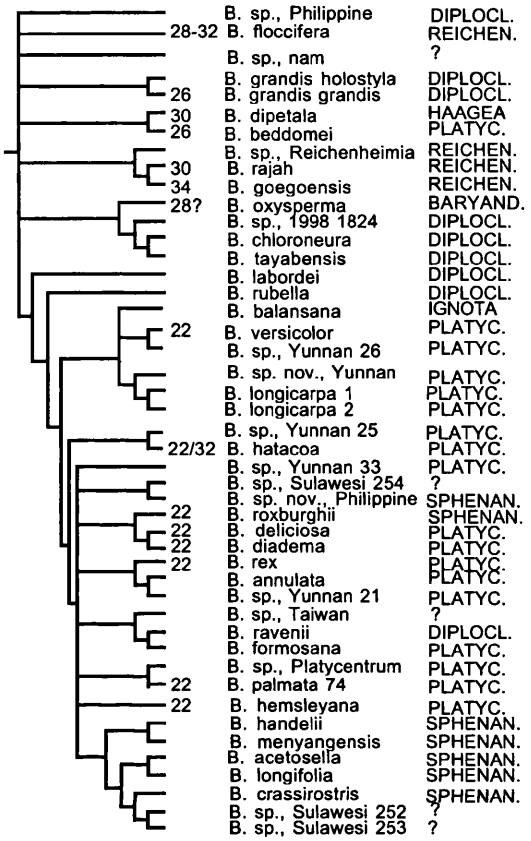
Clade 11: It is hard to say anything meaningful based on so few counts (see Figure 11.11). The pattern appears clear, but whether it would hold up to the addition of any more data is impossible to say. Legro and Doorenbos (1971) speculate that the count of 2n = 44 in section *Petermannia* represents a triploid of 2n = 30 (45, with one chromosome lost), not a tetraploid of species with 2n = 22. This phylogeny tends to support that view, particularly as all the Asian species with 2n = 22 are confined to Clade 12. There may be a geographic correlation in section *Petermannia*, with 2n = 30 found in western species and 2n = 44 only in New Guinea (Legro & Doorenbos, 1969); however, only eight taxa out of c. 200 have been counted for this section (including two from New Guinea with 2n = 30); it is too soon to generalise. It would however be very interesting to have counts for the *Symbegonia* species.

Figure 11.11: Clade 11 (Asia)



Clade 12: At the base of clade 12 there is a large polytomy (see Figure 11.12); there is a variety of chromosome numbers within this, and few indications of trends. The resolved part of this clade appears to be characterised by 2n = 22; there is some premium on obtaining counts for *B. rubella* and *B. labordei*, both from section *Diploclinium*. *Diploclinium* is a large and polyphyletic section which is "a show-case of the difficulties one meets when trying to delimit sections" (Doorenbos, Sosef & de Wilde, 1998, p. 93); species assigned to it have chromosome numbers 2n = 22, 26, 28, 32, 36, 38 and 44. *B. picta*, which is morphologically similar to *B. rubella* and *B. labordei*, has been counted and, tantalisingly, has 2n = 22. Section *Platycentrum*, to which most of the species in the resolved clade belong, has had 2n = 22 counted for 15 species (seven of which are represented in this ITS phylogeny) (and also a few counts of 2n = 44); Section *Sphenanthera*, which is included within section *Platycentrum*, has had two counts, one (*B. roxburghii*, sampled here) of 2n = 22; the other, for *B. robusta*, is 2n = 88.

Figure 11.12: Clade 12 (Asia)



11.4.4 Summary of cytological patterns

Table 11.1 Chromosome trends: Summary of CD-ROM Table.

	No. species counted (out of total)	No. individual counts	No. different Chrom. no.s	Range of numbers
AFRICA	36 spp (out of c. 140)	98 counts	27 numbers	20 - 76 (i.e. 56)
AMERICA	160 spp (out of c. 600)	260 counts	28 numbers	24 - 156 (i.e. 132)
ASIA	59 spp (out of c. 645)	116 counts	22 numbers	16 - 88 (i.e. 72)

Looking at Table 11.1, it seems that the most cytological diversity in *Begonia* is in America, at least in terms of range of numbers (although it is possible that more sampling in Asia, where less than 10% of species have been counted as opposed to c. 27% in America, may reveal a wider range

of numbers there). Africa, where c. 26% of species have been counted, has the smallest range of chromosome numbers, and Asia has the lowest number of different counts.

This is interesting, because Africa has the greatest amount of ITS sequence divergence, while ITS sequences from America and Asia are far less divergent, and molecular evidence (the ITS phylogeny) suggests that the African lineages are older than the Asian and American lineages.

Some general trends in chromosome number are apparent - for example, the prevalence of 2n = 22 in the Asian *Platycentrum* clade (within clade 12). It also appears that polyploidy has occurred several times. The polyploid 2n = 56, for example, characterises part of the American *Pritzelia* clade (clade 9), but has also arisen on other occasions, for example, within clade 7 (in *B. incarnata*) and in clades 5 and 6. Transitions up an euploid series are far easier than transitions down a series, thus in cases like clade 6, which have 2n = 28 and 2n = 56, 28 is probably the basal number. Thus numbers of chromosomes are frequently homoplastic in *Begonia*.

The occurence of what have been suggested to be triploids (e.g. 2n = 3x = 44 in clade 11; 2n = 3x = 38 in clades 2 and 3) does not correlate with sterility. For example, *B. brevirimosa* (2n = 44) and *B. ampla* (2n = 38; not sampled for ITS, but consectional with *B. poculifera* in clade 3) certainly set plentiful seed in cultivation at E. If the polyploid series suggested by Legro and Doorenbos (1969) are right, it could be that these are higher polyploids and therefore the basic number for *Begonia* is lower than any of the numbers which have been found. More evidence that the basic number is low and that nearly all species are polyploid lies in the aneupolid series which have been reported in *Begonia* (e.g. counts of 20, 21, 22, 23, 24, 25, 26 in Africa). The species in clade 1, with 2n = 22 and 2n = 26, set copious seed.

11.4.5 Hybridisation in *Begonia*

There are clearly many different chromosome numbers within *Begonia*. Although the source of this variety is not clear based on this study, there are a few probable causes: polyploidy, aneuploidy and hybridisation.

Peng and Chen (1991) investigated the endemic Taiwanese species, *B. buimontana* Yamam. (2n = 30). The male flowers of this species always drop before anthesis[12] , and pollen is nearly completely aborted. Meiosis in the species is also abnormal, with "some sticky, often disoriented bivalents and at least 11 univalents; multivalents are often present" (p. 997). Peng and Chen suspected that this plant is of hybrid origin, from a cross between *B. palmata* (2n = 22; section *Platycentrum*) and *B. taiwaniana* (2n = 38; section *Diploclinium*). They made artificial crosses between the putative parents, and obtained F1s which resemble *B. buimontana*, drop their male flowers prematurely, and have a somatic chromosome number 2n = 30 (i.e. a novel number).

Peng and Chen think that *B. buimontana* is represented only by F1 hybrids in the wild, because the populations are very uniform morphologically, and because the experimentally derived hybrids are very similar to wild plants. A few wild origin plants have been found which have set seed, but although some of the seed ("probably derived from back crosses with the putative parental species" - p. 998) was germinated in a greenhouse, it died at the cotyledon stage. They put the maintenance of these hybrids, once established, down to the perennial habit, and suggest that expansion of the distribution of this hybrid can be achieved by recurrent natural hybridisations.

[12] The character of male flower opening or dropping (as rang alarm bells for Peng & Chen, 1991) must be investigated in the plant's natural habitat. *B. listada* in cultivation at the Royal Botanic Garden, Edinburgh and Glasgow Botanic Garden (pers. obs.), and at the Royal Botanic Garden, Kew (pers. comm., Sands, 2000) drops its male flowers before they open. However, descriptions of the species (Smith & Wasshausen, 1981; Karegeannes, 1981) are with open male flowers. Thus the problem may lie in its growing conditions under glass. Legro and Doorenbos (1971) counted *B. listada* as 2n = (76), although Doorenbos (pers. comm., 1999) gives a new count of 2n = 56, which is the most common chromosome number in section *Pritzelia* (it is found in 31 species out of 41 which have been counted).

Although this hybrid does appear to be largely sterile, the fact that it was first described in 1933 and could still be found in the wild over 50 years later, distributed through three counties in Taiwan, suggests that either the original plants have a relatively long lifespan and have spread clonally, or that the *B. palmata* - *B. taiwaniana* cross has occurred repeatedly (or rare sexual events may be sufficient to maintain populations). Either way, the longer the (largely) sterile plant can survive in the wild, the greater the chance of polyploidy conferring it a degree of fertility, or of introgression with one or other of the parents. Had this plant been fully fertile, an hybrid origin would not necessarily have been investigated; thus how many of what act as good *Begonia* species are the results of reticulations rather than divergent speciation it is not possible to say.

There is also a record of a natural *Begonia* hybrid in Malaysia, between *B. decora* Stapf and *B. venusta* Ridl., both from section *Platycentrum* (Teo & Kiew, 1999). Six hybrid populations have been identified using morphological characters; within them, some individuals are morphologically more similar to one parent, and some more similar to the other. The pollen germination of the hybrid plant is c. 97%, and seed germination is 98%. Teo and Kiew conclude that the fertile hybrids back-cross with the parents to produce hybrid swarms; there appear to be no genetic barriers between the morphologically distinct species *B. decora* and *B. venusta*.

Assuming these hypotheses of hybridisation are correct (some molecular studies may be informative), caution must be used in the interpretation of cladograms based on species of *Begonia*. Where data have a strong geographic structure, as is the case in the *Begonia* ITS analyses, this may either reflect the real evolutionary history of the genus, or a series of reticulations between plants which grow together. Comparisons between chloroplast and nuclear phylogenies can be instrumental in untangling these questions (Rieseberg & Soltis, 1991).

That one of the recorded wild hybrids (Peng & Chen, 1991) is a cross between species from two sections somewhat negates any argument that

reticulations are most likely between closely related species so will not affect overall topology. Further, the cross is between species with different chromosome numbers. Hybridisation may thus confound phylogenetic inference, especially if just a single region is used.

How badly reticulation events affect other species within *Begonia* depends somewhat on how common or rare such events are in nature. Different species are commonly found growing in similar habitats; *Begonia* flowers do not appear to be adapted to specific pollinators (except a few probable shifts from the common condition of insect pollination, to bird pollination, in the Andes and in New Guinea), and different species often have very similar flower structure; many *Begonia* species are cross-fertile; so the main barrier to hybridisation may be some form of temporal separation. Most *Begonia* do not produce nectar, so pollen is thought to be the only reward. Therefore sophisticated temporal separations based on the timing of nectar-release would not be a consideration. Of course, there may be many other ways plants maintain their identities, and it remains that there are little empirical data unambiguously documenting hybridisation in natural populations of *Begonia*.

11.5 Summary

604 chromosome counts, representing 239 species, have been gathered and are presented on the accompanying CD-ROM. Counts for species which are represented in the ITS phylogeny have been mapped onto the phylogeny; this phylogeny has then been used to look for trends in chromosome number in *Begonia* lineages.

Most of the cytological diversity in *Begonia* appears to be in America, which has the largest range of chromosome numbers. There are some trends within clades, for example the prevelance of 2n = 22 in the *Platycentrum* clade; there is also some homoplasy, with the same number generated several times, for example, 2n = 56 in clades 5 and 6, 7 and 9.

12. Evolution, Biogeography and the Begoniaceae

12.1 Introduction

The production of cladograms for a group is only the start of an interpretative process; converting a cladogram to a phylogeny may involve little more than accepting it as a picture of the evolutionary pattern within a group; alternatively, some less parsimonious or optimal solution may be accepted (hopefully with some sort of justification), as is the case in Sosef (1994). In previous chapters I have considered cladograms produced from nuclear and chloroplast sequence data (26S, ITS and *trnC-trnD*) and from morphology. I have also considered the available cytological information for the family. From these data sets it should be possible to say something about evolution and biogeography within the Begoniaceae.

Before discussing the evolution and biogeography of *Begonia*, it is worth briefly reviewing some major events in Earth's history over the last c. 150 million years, to place into context the environment in which the genus has evolved.

12.2 Geology through time

For dates of the geological time periods discussed, see Table 12.1.

Table 12.1 Geological Time Scale

ERA	PERIOD (Hallam, 1994)	PERIOD (Bennett, 1997)	EPOCH	AGE (Ma)
Cenozoic	Neogene	Quaternary	Holocene	0.01
			Pleistocene	1.64
		Tertiary	Pliocene	5.2
			Miocene	23.3
	Palaeogene		Oligocene	35.4
			Eocene	56.5
			Palaeocene	65
Mesozoic	Cretaceous			145.6
	Jurassic			208.0

12.2.1 CRETACEOUS: The Gondwanan continent consisted largely of the land masses currently known as South America, India, Africa, Madagascar, Australia, New Zealand and Antarctica. On the eastern side of what is now Africa, Madagascar and then India were joined. Madagascar / India began to separate from mainland Africa during the early Cretaceous (Hallam, 1994). Madagascar reached the position it now occupies (relative to Africa) about 105 Ma. Sea also began to open up on the western side, between South Africa and Argentina, about 130 Ma (Scotese, Gahagen & Larsen, 1988); all the connections between Africa and South America were severed during the Late Cretaceous, about 95 to 80 Ma (Parrish, 1993).

12.2.2 PALEOGENE: There have been suggestions, based on freshwater frog and snake distributions (Hallam, 1994, p. 148-151), that some form of land bridge occurred between the Rio Grande Rise (South America) and the Walvis Ridge (southern Africa) in the south, and/or the Ceara (South America) and Sierra Leone Rise (Africa) to the north, during the late Cretaceous to early Palaeocene (c. 65 Ma). The Walvis Ridge is thought to have been completely submerged by the end of the Eocene (Parrish, 1993), while the Rio Grande Rise may have been submerged by the late Oligocene (Thiede, 1977); a Palaeogene sweepstakes route[13] is thought more plausible than a continuous land corridor, "not just because of geological considerations but also the strong endemism of African and South American mammals" (Hallam, 1994, p. 165).

During the Mesozoic, temperate forests extended as far as the polar regions, and there was a wide tropical-subtropical zone; the mid-Eocene flora in western Europe was predominantly tropical. Global cooling occurred across the Eocene - Oligocene boundary; this was probably associated with the development, in the early Oligocene, of a circum-Antarctic oceanic circulation system, and led to the development of glaciation and ice-sheet formation (Hallam, 1994).

[13] "Chance crossings or migrations across a water barrier or other major biogeographic obstacle by rafting or other means of transport" (Lincoln, Boxhall & Clark, 1982, p. 240).

India separated completely from Madagascar-Africa a little before 90 Ma (Veevers, Powell & Johnson, 1980); it carried on it many African plant species. The descendants of these species dispersed into SE Asia after the collision of the Indian plate with Asia in the mid Eocene. This pattern can be traced in several palm taxa (Morely, 1998). The collision of India and Asian caused the thickening of the Tibetan crust between the mid Eocene and early Miocene; the Tibetan plateau reached its present elevation c. 8 Ma (Windley, 1995).

After falling considerably at the end of the Cretaceous, sea levels rose in the Palaeocene and the Eocene. However, another large fall in sea level occurred in the mid Oligocene, when levels could have dropped by up to 100 m (Hallam, 1992).

Although most of the Mozambique Channel is over 2000 m deeep, there is evidence of a land bridge between Africa and Madagascar between the mid Eocene and lower Miocene (c. 45 - 26 Ma) (McCall, 1997).

The major global fall in sea level in the mid Oligocene exposed large areas of Sundaland and Sunda shelf; there was probably more dry land than at any subsequent time until the end of the Cenozoic. Around 25 Ma the north Australian margin came into contact with Sulawesi and the Halmahera arc, possibly creating a discontinuous land connection across the Philippines into Sulawesi (Hall, 1998).

12.2.3 NEOGENE: In the early Miocene Africa collided with Eurasia. It is thought that there was also a major increase in aridity 20 to 30 Ma, causing a reduction in the amount of surface water (and consequently, extinctions in aquatic vertebrates in the western United States - Hutchinson, 1982). By the mid to late Miocene, the cooling of the global climate caused southern Africa to undergo aridification; closed forest was fragmented, and replaced by woodland and savanna.

There is evidence that tropical Africa was cooler and drier in the Pleistocene than it is today (Bonnefille, Roeland & Guiot, 1990); this may have

influenced the area occupied by rain forest species during these times (Sosef, 1994). Refugia sites for species during periods of glaciation have been postulated for a number of regions around the globe (for detailed reviews on European phylogeography, see Hewitt, 1996, and Ferris, King & Hewitt, 1999; for the Amazon region, see the paper by Haffer, 1969, based on bird distributions); in tropical Africa, refuges have been proposed in several regions including Cape Three Points, Ghana; the coast of Sierra Leone / Liberia; Cameroon / Gabon; and eastern Zaire (Sosef, 1994).

North and South America were isolated until late in the Neogene (Hallam, 1994); the Panama Isthmus became emergent early in the Pliocene, allowing relatively free migration between the continents.

A large area of land may have been exposed between Australia and Sulawesi by the late Miocene / early Pliocene (Hall, 1998). Although many of the islands of eastern Indonesia are thought to be very young (e.g. Seram, Irian Jaya, eastern Sulawesi), the island chains of the Philippines and Halmahera probably had emergent land with tropical plant cover through most of the Tertiary (Hallam, 1994).

Intermittent dry periods are recorded during the Neogene in the Sunda region of Asia (these are reflected by maxima of Gramineae pollen) (Morley, 1998). Such significant climatic fluctuations, as polar ice-caps expanded and retreated, were a feature of the Holocene; during these periods sea levels rose and fell globally; land links were formed and lost, e.g. across the northwest European shelf, the Bering Strait, and the Sunda shelf (Indonesia/Malaysia) (Hallam, 1994).

South Sulawesi (to the east of Wallace's line) today shows geological affinity with the Sunda plate and floristic affinity to the Eocene floras of India, Java and SE Kalimantan. The New Guinea flora, to the west of Wallace's line, is speculated to be a product of the mingling of East Malesian, Sundanian and Australian floras in the Miocene (Morley, 1998).

The uplift of north-central Borneo caused a mass of sediment into deltas in north and east Borneo. From c. 20 Ma, there has probably always been land exposed in the region of Sulawesi. From 15 to 5 Ma more of Borneo emerged, and volcanic activity and land mass collisions led to intermittent emergence of many points of land (Hall, 1998). However, concurrently, deep basins also formed (e.g. Sulu Sea, Banda Sea) which would have formed new barriers at the same time as new pathways were also forming. At the present moment "there are more highland areas, and a greater area of land [in SE Asia] than at any time during the last 30 million years" (Hall, 1998, p. 122).

12.2.4: Summary of main points:

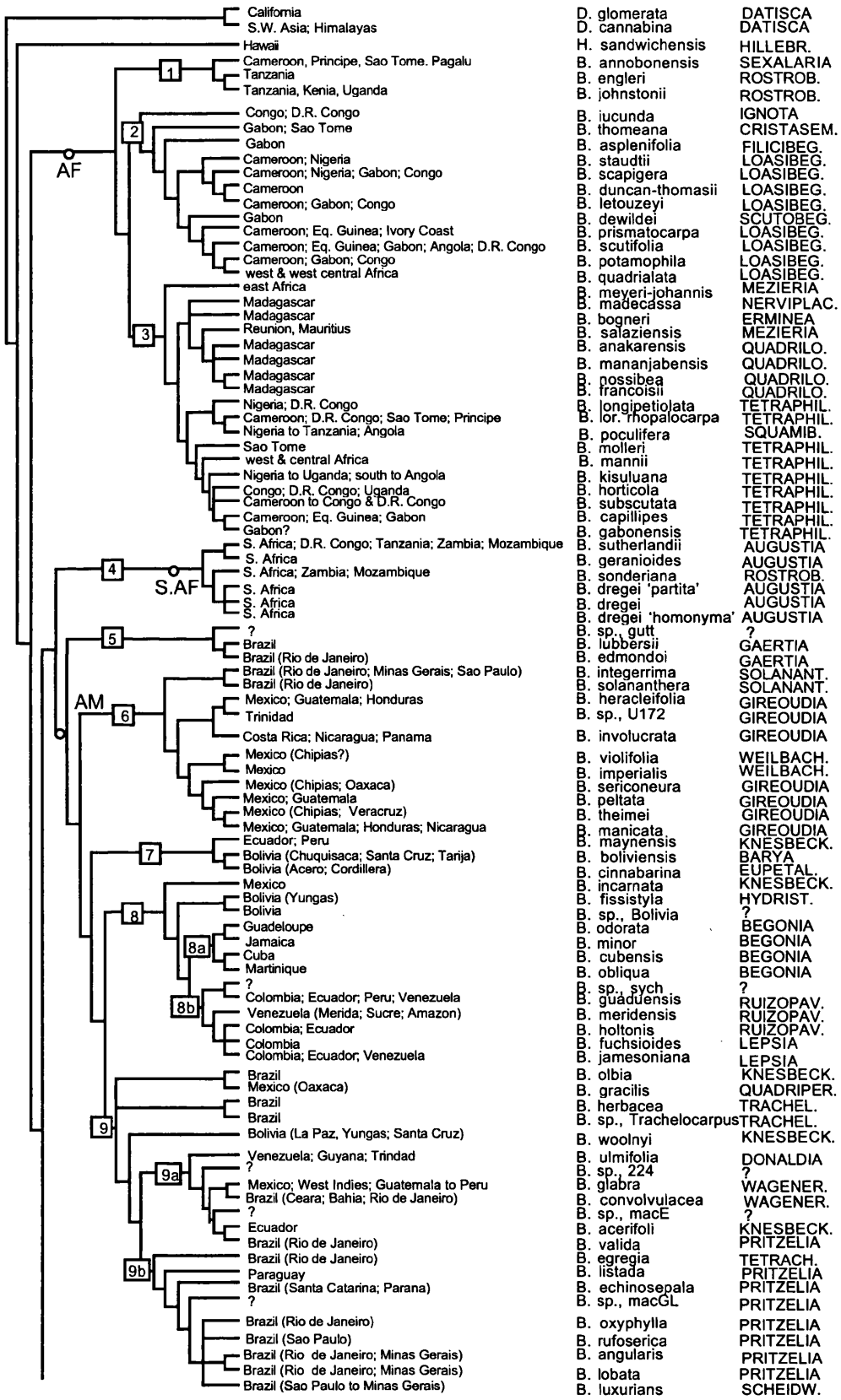
Cretaceous	Gondwanaland.
c. 130 Ma	Sea opens between S. Africa and Argentina.
early Cretaceous	Madagascar / India separate from Africa.
c. 105 Ma	Madagascar reached present position relative to Africa (Hallam, 1994, p. 139).
95-80 Ma	All land connections between Africa and South America lost.
c. 90 Ma	India separates from Madagascar/Africa (Hallam, 1994, p. 139).
c. 65 Ma	landbridge forms between Africa and South America (Rio Grande Rise - Walvis Ridge).
Eocene	Sea levels rise.
mid-Eocene	Indian plate collides with Asia. Tropical flora in western Europe.
end Eocene	Walvis Ridge (South Atlantic Ocean, southern Africa) submerged. Global cooling.
mid Oligocene	Global sea levels fall by up to 100 m. Wallace's line - barrier to plant dispersal.
late Oligocene	Rio Grande Rise (South Atlantic Ocean, South America) submerged.
early Miocene	Africa collides with Eurasia.
Miocene	East Malesian / Sundan / Australian floras mix.
45-26 Ma	Africa / Madagascar land bridge (McCall, 1997, p. 663).
30-20 Ma	major global rise in aridity.
c. 25 Ma	N. Australia in contact with Sulawesi/Halmahera arc; discontinuous land connection across Philippines into Sulawesi (Hall, 1998).
mid-late Miocene	aridification, southern Africa; forest fragmented.
early Pliocene	Panama Isthmus emerges.

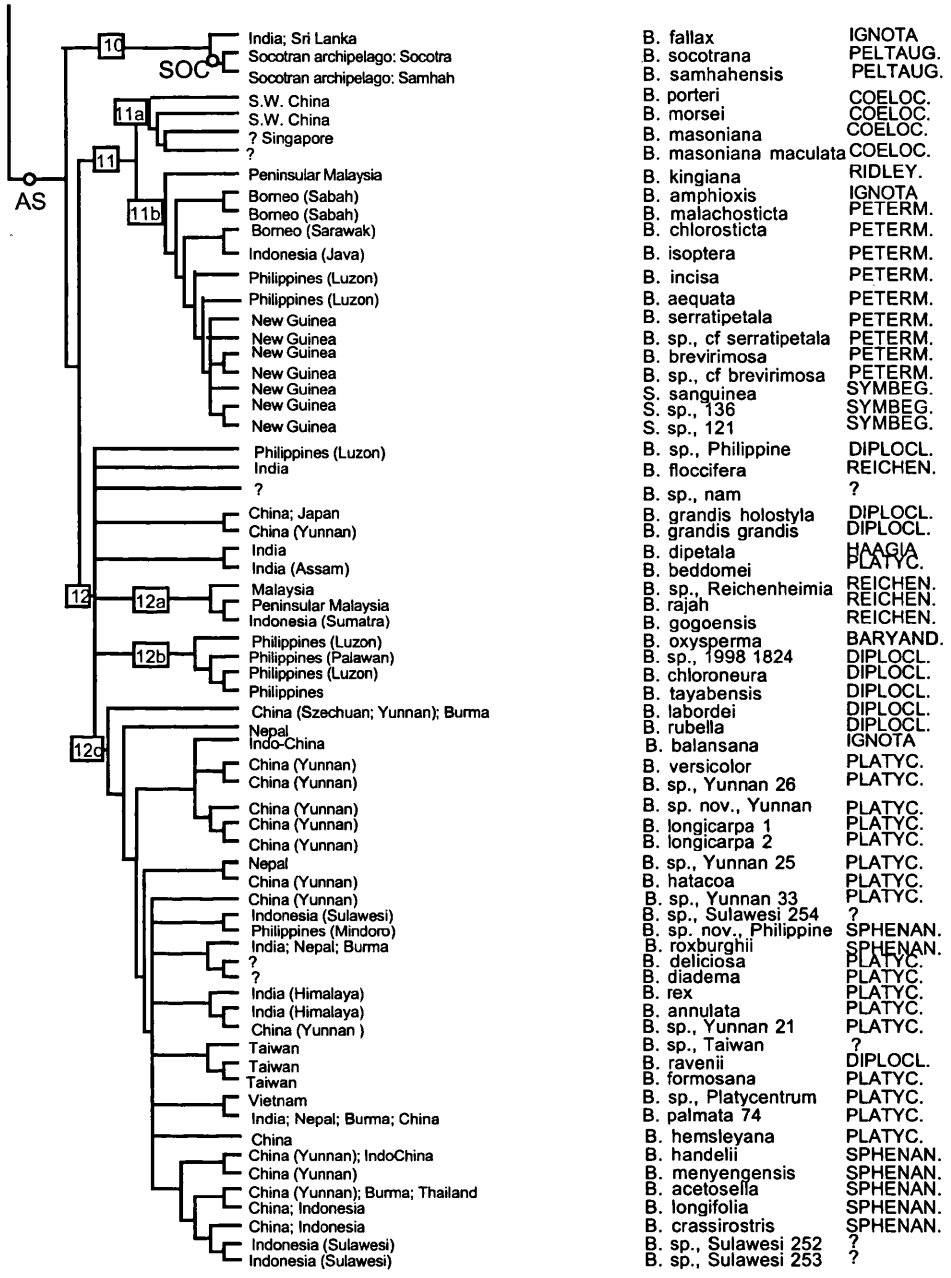
12.3 Geographic Origins

12.3.1 Introduction: Either *Begonia* is a very ancient group, and its modern day distribution is explicable in terms of plate tectonic events (the explanation preferred by Sharp (1996, 2000), who argues vociferously if not convincingly for a Gondwanan origin for the genus) or we need to look to more recent events (e.g. dispersal and land bridges formed during sea level fluctuations) to explain its biogeography. The age of the genus *Begonia* is pertinent to this question; there is, however, no fossil record to guide us. Guo and Ricklefs (2000) give the Cucurbitales a phylogenetic grade of six (compared to, for example, zero for the gymnosperms, one for the magnoliids, four for the Brassicales), suggesting that they consider them to be a relatively derived group among angiosperms. However, simplistic arguments (ancient group = vicariance; modern group = dispersal) will not necessarily reflect modern day distribution patterns; ancient events can be overlaid by more recent events, and vicariance can occur contemporaneously with, after, or prior to, dispersal.

Gondwanaland: *Nothofagus* is the classic example of a genus with a "typical austral distribution" (Humphries & Parenti, 1999, p. 129). Early *Nothofagus* pollen is recorded from the Cretaceous in Australia and Antarctica, South America and New Zealand; it has not been found in South Africa and India. Morley (1998) uses this as evidence that the latter were "well separated from Gondwanaland at the time of its [*Nothofagus*] initial radiation" (p. 215). *Nothofagus* subsequently dispersed from Australasia to Papua New Guinea and Irian Jaya, apparently correlating with the uplift of the New Guinea mountains; it appears in the Birds' Head of Irian Jaya in the late Miocene but did not disperse further west, presumably because it could not disperse across water (Morley, 1998). *Begonia*, on the other hand, is found in South Africa, India and South America. Its absence from Australia/New Zealand suggests that it was not dispersed across Gondwanaland when the continent broke up.

Figure 12.1: ITS phylogeny of *Begonia*, with geography marked on





(Boxed numbers on clades will be discussed later in this chapter)

Figure 12.1 is the ITS 'jigsaw' phylogeny of *Begonia* (Figure 7.22), with geographic distributions of species (taken from Doorenbos, Sosef & de Wilde, 1998) marked on. There are strong geographic correlations in this phylogeny (e.g. a Madagascan clade; a New Guinea clade). There are two possible reasons species may show such geographical correlations. Firstly, if species are poorly dispersed and sister taxa therefore tend to remain in geographic proximity; secondly, if hybridisation is common and the pattern relates to reticulation events between geographically (rather

then necessarily phylogenetically) proximal taxa. For the purposes of discussion, the former will be assumed to be responsible for the majority of the phylogenetic patterns observed. In order to test the latter possibility, a chloroplast phylogeny could be compared to the patterns reconstructed in this nuclear ribosomal treatment.

12.3.2 *Datisca*: The two species in *Datisca* have a disjunct distribution, occurring in S.W. Asia and California. Because the Californian species (*D. glomerata*) exhibits the unusual breeding system of androdioecy (Liston, Riesberg & Elias, 1990) while the Asian species (*D. cannabina*) is dioecious, the genus has been fairly widely studied (Liston, Riesberg & Elias, 1989, 1990; Liston, Riesberg & Hanson, 1992; Riesberg, Hanson & Philbrick, 1992; Swensen, Mullin & Chase, 1994).

Liston, Riesberg and Hanson (1992) dated the divergence of the two *Datisca* species at around 10 Ma (late Miocene), based on cp-DNA mutation rates. This is the last period when there were land connections between the temperate deciduous forest which spanned the northern hemisphere. Thus the two species may be the result of the past fragmentation of a formerly more continuous range. Eurasia and North America have seen considerable climate fluctuations within even the Quaternary. Much of the North American temperate flora is thought to be relictual of a flora formerly far more widely distributed through the northern hemisphere; many of the disjuncts between eastern Asia and North America (including the two *Datisca* species) may be Tertiary relicts formerly more widespread across higher latitudes during the Palaeogene (Guo & Ricklefs, 2000). A less reliable prior estimate for divergence times for the two *Datisca* species, extrapolated from Nei's mean genetic identity values for isozyme data, was 10 - 40 Ma (Liston, Riesberg & Elias, 1989, p. 538); this is the value quoted by Guo and Ricklefs (2000) in their analysis of eastern Asian - North American disjuncts.

Divergence times based on ITS are discussed in the next section.

12.3.3 *Hillebrandia*: The Hawaiian Islands are 3,900 km from the closest continent, with the highest known rate of endemism for any major archipelago. Emergent land has existed in the location of the islands for the past c. 70 Ma (Kim et al., 1998). The progenitors of all the c. 1000 species of native angiosperms are thought to have got to the Hawaiian islands by long-distance dispersal; Malaysia, North America, northern South America, Australia, New Zealand and South America have all been proposed to have floristic affinities with Hawaii (Kim et al., 1998). Affinities of the native plants, which may have been diverging from their mainland sister groups for up to 70 million years, can be difficult to determine.

Kim et al. (1998) investigated the phylogeny of the endemic Hawaiian genus *Hesperomannia* (Asteraceae); they found it to have affinities with African taxa (c. 15,000 km away); they dated the divergence between the African and Hawaiian taxa at 17-26 Ma. Seelanan, Schnabel and Wendel (1997) also found links between African and Hawaiian taxa in the family Malvaceae, with a sister group relationship between *Kokia* Lewton (Hawaii) and *Gossypioides* Skovsk. ex J.B.Hutch. (East Africa-Madagascar) dated c. 3 Ma (Pliocene), necessitating an hypothesis of long-distance trans-oceanic dispersal. These studies are relevant to *Hillebrandia* because our analyses suggest that *Hillebrandia* is sister to *Begonia*, and that its nearest relatives may be found in Africa. This cannot be explained in any way but as a long distance dispersal event, given that Hawaii is not hypothesised to have been connected to any mainland. Possibly *Hillebrandia* belonged to a more widespread lineage which has undergone extinctions in other geographic locations. Without finding either fossil evidence or extant relatives in some other location it is not possible to decide whether this is the case.

If Begoniaceae ITS DNA sequence divergences occurred following a regular molecular clock, a date could be put on the *Hillebrandia* / *Begonia* divergence. Given considerable difficulties in aligning sequences from the two genera so that many positions have been excluded from our phylogenetic analyses, comparing ITS rates with those in other plant groups is not reliable. Calibration by dated fossil remains is not possible,

as there are none, and calibration by geological events (e.g. the oldest known age of emergent land at Hawaii) would be extremely circular. Further, even within a more recent Asian clade, *Begonia* ITS does not appear to show clock-like behaviour (see later discussion) and so there is no reason to suppose that it may have in the past.

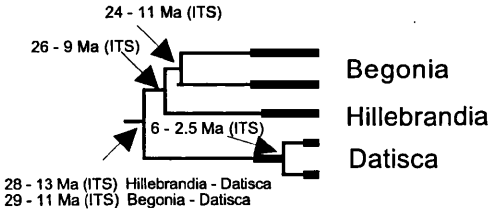
Such (rather compelling) provisos aside, using a molecular clock with estimates of 0.79% - 1.57% nucleotide substitutions per million years (Sang et al., 1994; Sang, Crawford & Stuessy, 1995) for ITS (uncorrected pairwise distances, with the 5.8 region excluded), very approximate values can be put on clades (see Table 12.2).

Table 12.2: Rough clade ages assuming an ITS molecular clock

TAXA	SEQUENCE DIVERGENCE	AGE RANGE
Datisca c. - Datisca g.	9%	6 - 2.5 Ma
Datisca c. - Hillebrandia	42%	27 - 13 Ma
Datisca g. - Hillebrandia	44%	28 - 14 Ma
Datisca - B. johnstonii	34%	22 - 11 Ma
Datisca - B. valida	45%	29 - 14 Ma
Hillebrandia - B. meyeri-johannis	31%	20 - 9 Ma
Hillebrandia - B. solananthera	41%	26 - 13 Ma
B. nossibea - B. dewildei	37%	24 - 11 Ma

However, with pairwise divergence values for ITS 1 and ITS 2 up to c. 45% (when all the variable regions, which were excluded from analyses, are included), it is apparent that the alignment of some regions may be inaccurate; even were the alignment accurate, there is a high possibility that multiple hits will lead to underestimates of divergence using uncorrected pairwise differences.

Figure 12.2: Molecular clock - based estimates of lineage age

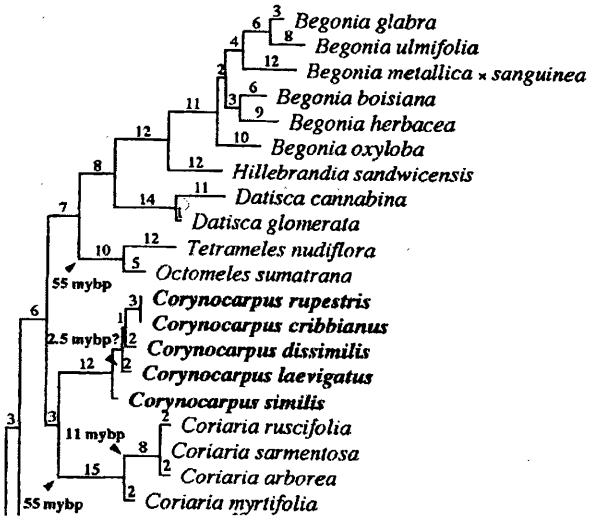


Drawn to scale of upper values of ITS value estimates, with the lower edges of the ranges marked in bold.

More data is needed to corroborate the basal divergences within Begoniaceae; estimates using maximum likelihood rather than uncorrected pairwise distances would also be more informative.

Wagstaff and Dawson (2000) suggest a date of 55 Ma for the Begoniaceae/ Datiscaceae lineage, based on early Eocene Tetramelaceae megafossils (see Figure 12.3). They also have an Oligocene leaf, stem and raceme fossil and lower Miocene pollen for *Coriaria*.

Figure 12.3: *rbcL* phylogeny of Coriariaceae, Corynocarpaceae, Tetramelaceae, Datiscaceae and Begoniaceae - from Figure 2, Wagstaff & Dawson, 2000 (p. 139)



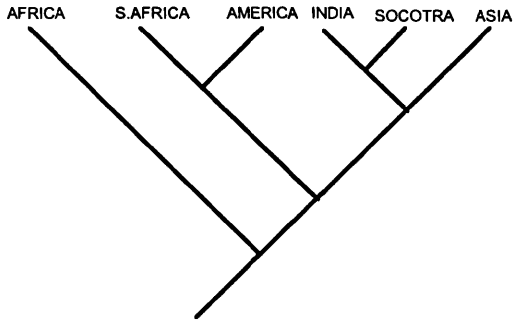
This suggests that the Begoniaceae / Datiscaceae lineage is **at least** 55 Ma (the fossil cannot give an upper limit for lineage age, and we cannot tell how long after the Begoniaceae / Datiscaceae lineage evolved the Begoniaceae evolved, so this date cannot be used as direct evidence for the age of Begoniaceae). *rbcL* branch lengths from the dated node to the terminals range from 23 (node to *Datisca glomerata*) to 51 (node to *B. ulmifolia*) (Wagstaff & Dawson, 2000). The relative rates test (Doyle & Gaut, 2000) gives a value of r = 0.451 (and so, does not support a time-calibrated clock).

It is possible, based on fossil evidence and on ITS and 26S molecular divergence values, that the Begoniaceae (*Begonia* and *Hillebrandia*) may be in the region of 60 to 20 million years old; more fossil evidence (and sequence from more genes) is needed to narrow this estimate.

12.3.4 *Begonia* - relationships from the cladograms

12.3.4.1 Continental relationships: Africa has been suggested as the area of origin for *Begonia* (e.g. van den Berg, 1995, p. 75); although species depauperate compared to the rest of the tropics (c. 140 species as compared to c. 1360 species), African taxa occur in morphologically isolated clades (e.g. *Tetraphila*; *Loasibegonia-Scutobegonia*), separated by suites of characters (e.g. flower colour, fruit fleshiness, seed and pollen micromorphology). The relationships suggested by this ITS analysis are summarised in Figure 12.4.

Figure 12.4: ITS-based geographic relationships of *Begonia* species

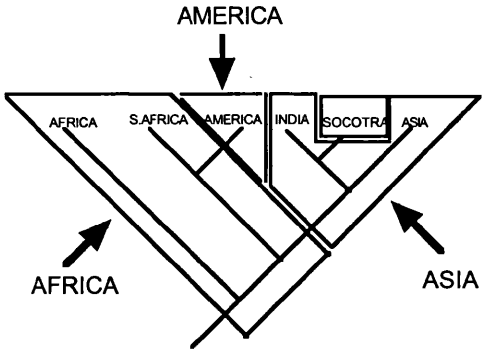


The paraphyly of African taxa can be interpreted in two ways: either with the African lineage older than lineages in Asia and America, or with there being two African lineages, one sister to the rest of *Begonia*, and the other, sister to all the American species of *Begonia*, derived from a more recent west to east dispersal event. These two options are shown in Figure 12.5. The explanation which has the most basal lineages in *Begonia* as African (i) is preferable to the other (ii) in terms of parsimony.

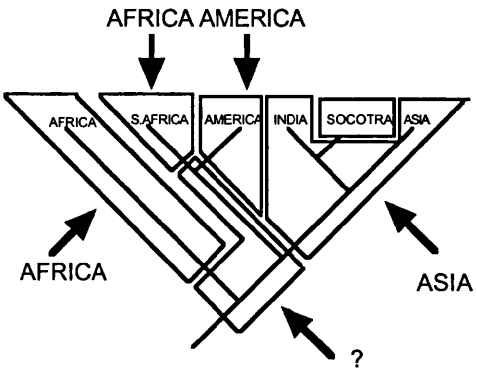
Figure 12.5: Geographic origins of *Begonia* lineages

12.5 a: CLADOGRAMS

i. Africa basal

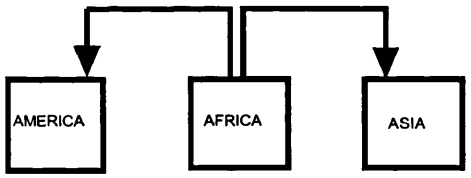


ii. Basal region unknown

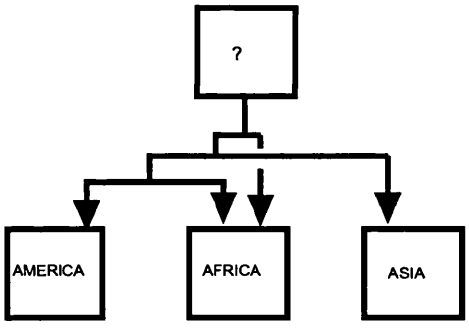


12.5 b. BLOCK DIAGRAMS

i. Africa basal



ii. Basal region unknown



One possible scenario (hypothesis one) is as follows:

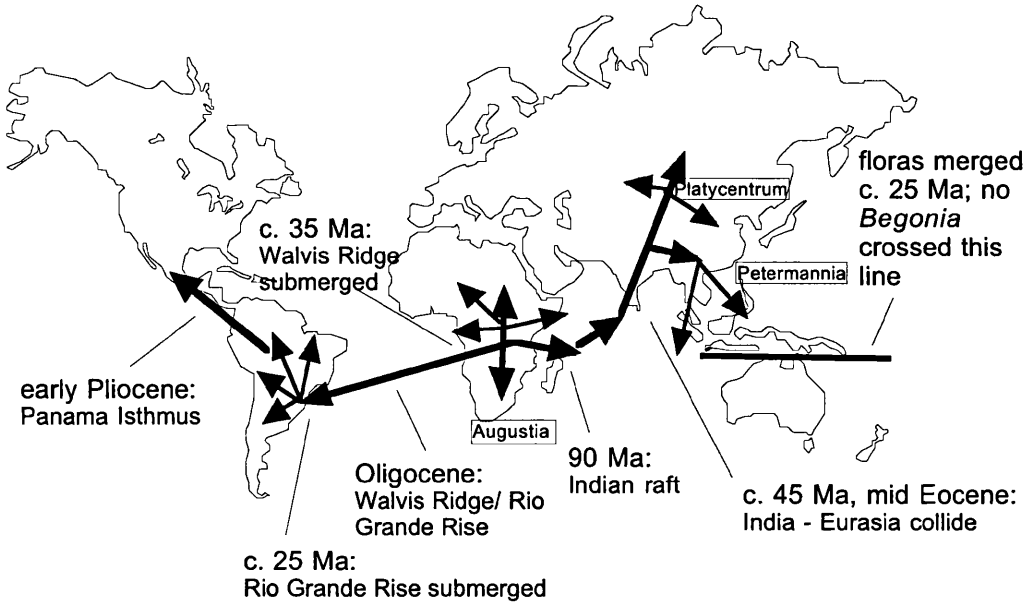
The ancestral *Begonia* evolved on Africa. The main African lineage contained all the taxa which were to become sections *Rostrobegonia*, *Sexalaria*, *Tetraphila*, *Squamibegonia*, *Cristasemen*, *Loasibegonia*, *Scutobegonia*, *Filicibegonia*, and all the Madagascan species. There was another lineage, possibly with an easterly distribution, across to the land mass which was to become India (see Figure 12.6 a). About 90 million years ago India separated from Africa/Madagascar and moved north, carrying on it some taxa from this lineage, the *Begonia* species which were to populate Asia. One lineage dispersed from India onto Socotra

(*Peltaugustia*); the rest of the species arrived in Asia in the mid Eocene; from there, they radiated out across Asia, with one clade (*Platycentrum*) diversifying particularly on the geologically active landscape of the Himalayas, and the other (*Petermannia*), particularly across the emerging and submerging islands of Malesia.

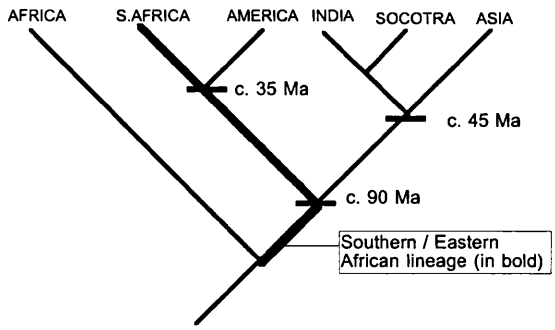
Other species from the eastern African lineage migrated towards the south of Africa and spread across to the west coast. From there, one lineage (*Augustia*) spread through southern Africa, while the other, the ancestor to all the American species, crossed (possibly) via the Walvis Ridge/Rio Grande Rise (probably before the onset of the Oligocene); the Rio Grande Rise was submerged by the late Oligocene, so *Begonia* would have arrived in America by this date. They then radiated, and crossed through Central America into Mexico, over the Panama Isthmus during the early Pliocene.

Figure 12.6: *Begonia* biogeography, hypothesis one

12.6 a: Fitting lineages across a modern-day world map



12.6 b: Fitting dates onto the cladogram



Although the order of these events fits the cladogram perfectly (see Figure 12.6 b), in order for India to carry *Begonia* to Asia, the date of origin for the genus would be over 90 million years ago (when India absolutely separated from Africa/Madagascar. Nearly all the diversity in Asia (c. 660 species) must have evolved during the last c. 45 million years, since India collided with Asia; most of the diversity in America (c. 600 species - de Lange & Bouman, 1999) would have occurred in the last 25 or so million years (although some diversification could have occurred on the Indian plate and/or South Atlantic land-bridge/islands). The Australian and south-east Asian floras came into contact about 25 million years ago; as *Begonia* has not been found on Australia or New Zealand, it seems likely that the genus had not reached islands like New Guinea at this point in time (rather than that it could have crossed to Australia and did not); Australia and New Zealand contain a large amount of habitat apparently suitable for *Begonia*.

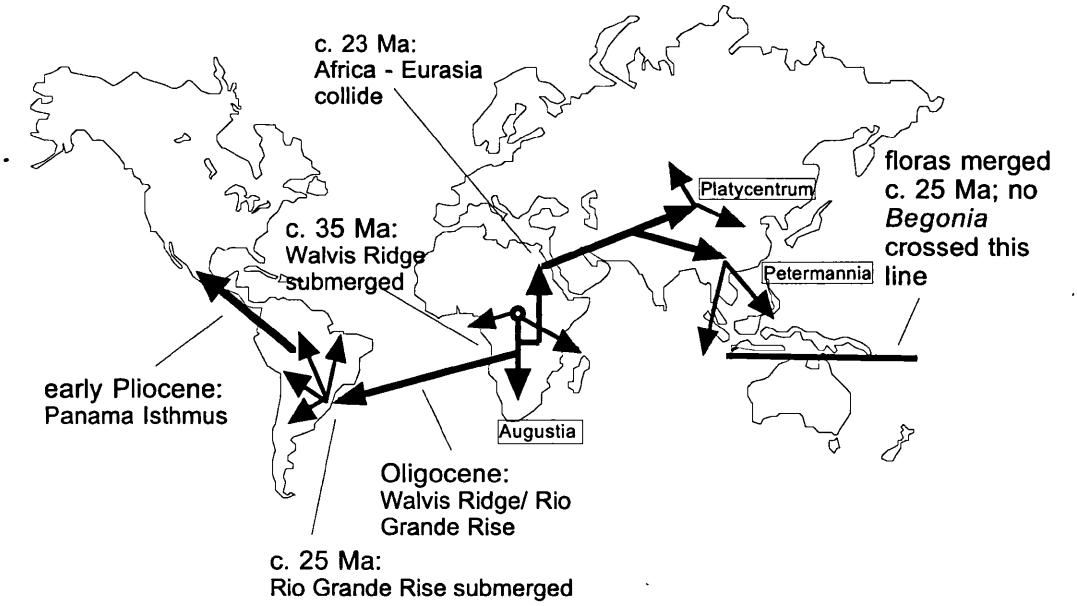
There are several examples of genera (e.g. Dipterocarpaceae Blume, *Gonystylus* Teijsm. & Binnend. (Gonystylaceae Gilg), *Ixonanthes* Jack (Linaceae DC ex Perleb.), *Eugeissona* Griffith. (Arecaceae) and *Durio* Adans. (Bombacaceae Kunth.)) which are today considered typically Malesian, having apparently rafted from Africa, radiated in Asia, and subsequently suffered range (and species number) reductions in Africa and India (Morley, 1998). Southern Africa, after all, suffered from aridification, forest fragmentation and savanna expansion about 10 million years ago; many *Begonia* species require a damp climate and forest cover to survive; it is possible that whole lineages were lost, and that the paraphyly of African *Begonia* (if extinct species were included) would be

more apparent.

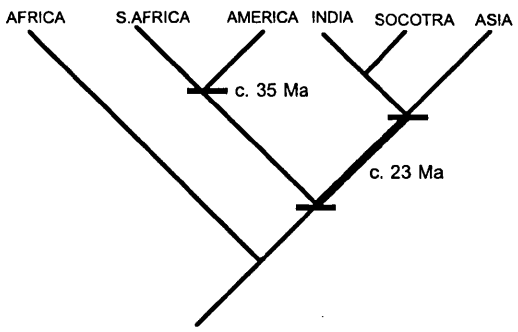
Still, so old a date sits a little uncomfortably on what is not generally considered a particularly basal angiosperm genus (and with the fossil date on the Begoniaceae/Datiscaceae clade of c. 55 Ma - Wagstaff & Dawson, 2000; Figure 12.3). Further, India, with massive geological activity and periods of aridification, would not have made a very hospitable raft for *Begonia*. An alternative hypothesis (hypothesis two) could be that the lineage of *Begonia* which went on to give rise to section *Augustia*/ American taxa/ Asian taxa separated into two clades while in Africa (see Figure 12.7 a); one migrated north; the other, south/south east (diversifying into *Augustia*/ all the America species, as described). The clade which moved north would have been able to cross into Eurasia about 23 million years ago, when Africa and Eurasia collided.

Figure 12.7: *Begonia* biogeography, hypothesis two

12.7 a: Fitting lineages across a modern-day map



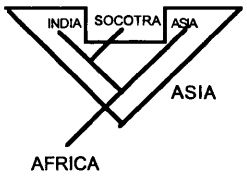
12.7 b: Fitting dates onto the cladogram



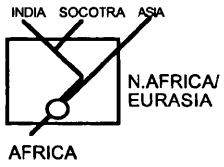
The chronological inconsistency (in the cladograms the Asian clade predates the American clade; see Figure 12.7 b) may be explained by extinctions in the north of Africa (along the bold line). Further, as the climate in Eurasia is now largely inhospitable to *Begonia*, taxa would have been lost from across Arabia (surviving only in Socotra). The sister relationship between Socotran and Indian taxa could be explained by lineage splitting in North Africa or Eurasia, closer geographically to the Socotran islands (Figure 12.8 b), with possible extinctions of other members of the lineages, prior to *Begonia* reaching, and radiating in, Asia, rather than by a dispersal event from India (Figure 12.8 a).

Figure 12.8: Asian and Socotran *Begonia* lineages

12.8a



12.8 b

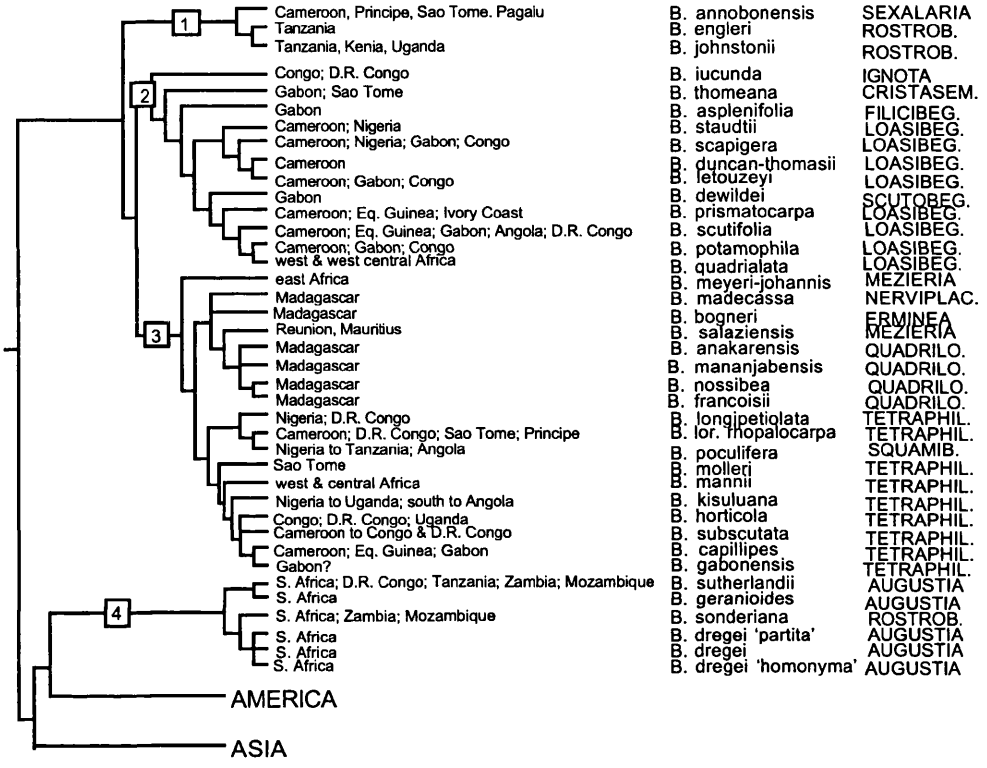


The following discussion is more concerned with the relationships within than between the continents; clade numbers are those marked onto the ITS phylogeny, Figure 12.1 (and were also used in Chapter 11 to discuss cytology - see Figure 11.1).

12.3.4.2 African clades

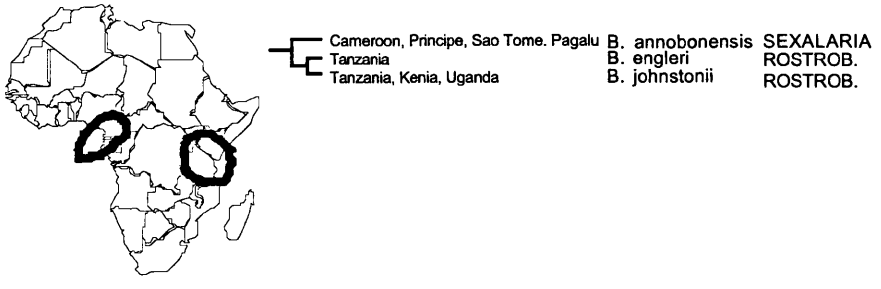
The African relationships from Figure 12.1 are redrawn in Figure 12.9 for convenience.

Figure 12.9: ITS-based relationships of African *Begonia* taxa



Clade 1: This clade appears as sister to all other east, west and central African species of *Begonia* (see Figures 12.9, 12.10).

Figure 12.10: Map of geographic distribution of species in Clade 1 and Clade 1 (Africa)

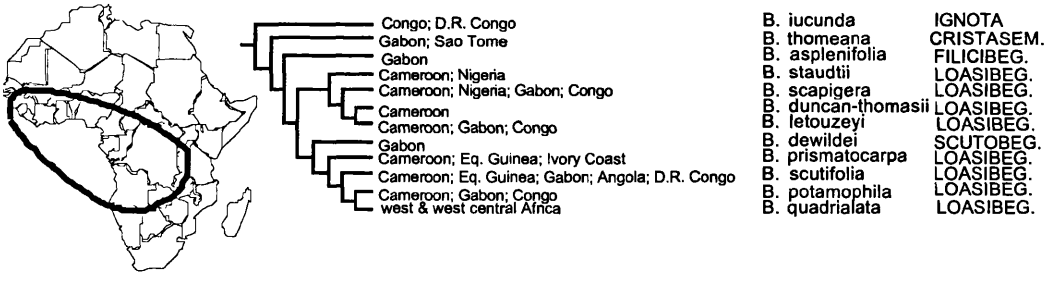


B. annobonensis (section *Sexalaria*) is a monocarpic species which can go

through several generations in a year. It is distributed from west Africa to the islands of Sao Tome and Pagalu. The other two taxa (*B. engleri* and *B. johnstonii*, section *Rostrobegonia*) are found on the east African mainland. Sampling more of the species from within section *Rostrobegonia* may give this clade a more continuous range. The section *Rostrobegonia* appears polyphyletic, so it is not possible to simply extrapolate its distribution from the distributions of species currently assigned to it.

Clade 2: This next African clade includes species from sections *Cristasemen*, *Filicibegonia*, *Loasibegonia* and *Scutobegonia* (see Figure 12.11).

Figure 12.11: Map of geographic distribution of species in Clade 2 and Clade 2 (Africa)



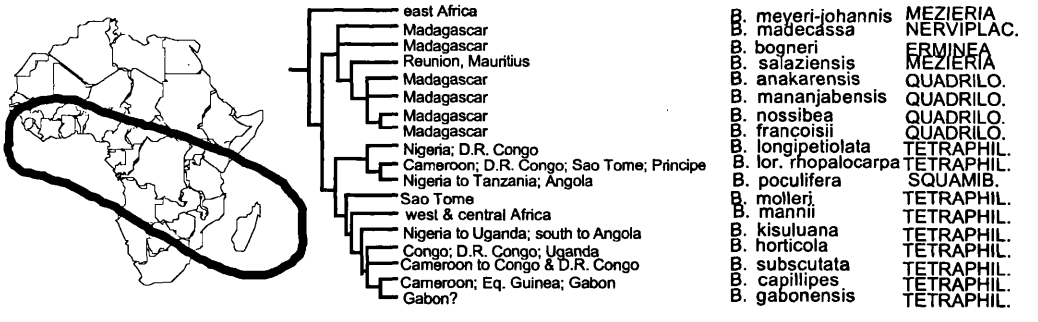
Many species within this group have relatively large, bright yellow, flowers, monochasial inflorescences, and, with the exception of the lianescent, ivy-like *B. thomeana* (section *Cristasemen*), they are all rhizomatous herbs. Species are distributed through west central Africa, from Guinea, east to the Democratic Republic of Congo, Rwanda and Burundi, and south to Angola, and west to the island of Sao Tome, Gulf of Guinea (*B. thomeana*). The idea that these species are related is not novel; De Lange and Bouman (1992) group sections *Filicibegonia*, *Loasibegonia* and *Scutobegonia* together according to seed characteristics.

Sosef (1994) looked at the biogeographic relationships of species within the sections *Loasibegonia* and *Scutobegonia* in an attempt to identify Pleistocene rain forest refugia. Although he managed to trace a few vicariance events, he suggests that the method he used was only capable

of tracing events of comparatively recent origin (the last glacial) and that vicariance events during previous glacial periods have been obscured "by renewed dispersal resulting in the display of floristic affinities rather than of vicariance in the data" (Sosef, 1994, p. 134). It may be, particularly on continental land masses, that biogeographic inferences should be restricted to broad (continental-scale) patterns rather than country-by-country comparisons. Also, if it is the case that there are several overlaid patterns in *Begonia*, for nowhere is this more likely to be true than for Africa, which appears to hold the oldest lineages in the genus.

Clade 3: This clade includes species from sections *Mezieria*, *Tetraphila*, *Squamibegonia* and all the Madagascan species of *Begonia* (see Figure 12.12).

Figure 12.12: Map of geographic distribution of species in Clade 3 and Clade 3 (Africa)



Despite great morphological variation within the island, all the sampled species from Madagascar are monophyletic. This has not previously been suspected, and demonstrates the potential for morphology to confuse. The Madagascan clade does include one non-Madagascan taxon, *B. salaziensis*, section *Mezieria*, from the Mascarine Islands (Reunion and Mauritius) to the east of Madagascar. Africa and Madagascar are currently c. 700 km apart, posing the question of how a lineage of *Begonia* got onto Madagascar.

Unlike *Begonia*, *Streptocarpus* Lindl. (Gesneriaceae) is thought to have colonised Madagascar three times from Africa (Möller & Cronk, in prep.,

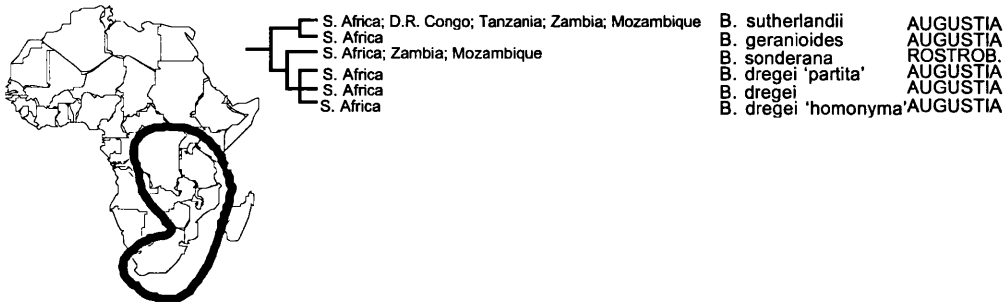
2001). Both genera are found in similar habitats (predominantly moist, shaded forest locations). Möller and Cronk (in prep., 2001), using a conservative estimate of 0.79 - 1.57% nucleotide substitutions per million years for ITS (Sang et al., 1994; Sang, Crawford & Stuessy, 1995), estimate maximum divergence time between African and Madagascan taxa to be 50 to 25 million years. Using the same substitution rates gives slightly younger divergence times for *Begonia* (between c. 21 and 10 million years, based on *B. duncan-thomasii* to *B. ankaranensis*, uncorrected pairwise distance 33%). Even doubling the maximal dates on these ranges does not put them into a suitable time-frame for Gondwanaland - based vicariance, as Madagascar is thought to have separated from Africa in the early Cretaceous.

Some form of land bridge between Africa and Madagascar has been suggested, from the mid Eocene to the early Miocene, 26 to 45 Ma (McCall, 1997). This is not far from the dates based on sequence divergence values (which will be underestimated for *Begonia*, given that uncorrected pariwise values were used), and also is more consistent with an age of c. 60 to 30 Ma for *Begonia*, as discussed before, than a Gondwanan disjunction would be.

Section *Tetraphila* is paraphyletic, also including species from section *Squamibegonia*. The species in this *Tetraphila/Squamibegonia* clade are widely distributed, not only on mainland Africa, but also to the west of Africa, in the Gulf of Guinea (on the islands of Sao Tome and Principe) and on the Mascarine Islands to the east of Africa. It is interesting that these widely distributed species (some species are recorded with disjunct mainland/island distributions, or from more than one island) include most of the fleshy-fruited *Begonia* species; fleshy-fruitedness is thought to correlate with bird dispersal, to which ocean is not necessarily a barrier.

Clade 4: These southern African species do not appear to be closely related to species from the rest of Africa, but appear as sister to an American clade (see Figures 12.9, 12.13).

Figure 12.13: Map of geographic distribution of species in Clade 4 and Clade 4 (Africa)

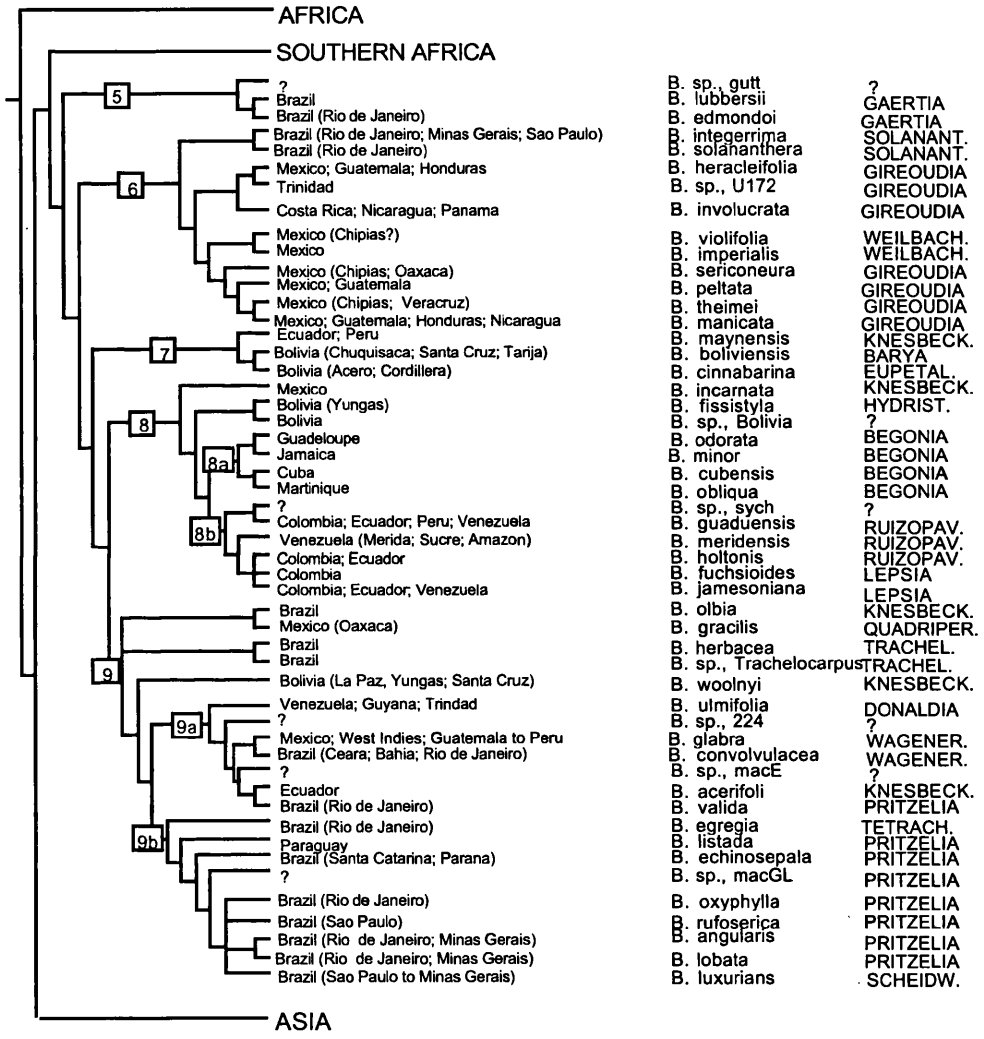


The southern African clade consists largely of species from section *Augustia*, although one species currently assigned to section *Rostrobegonia* (*B. sonderana*) is also included. Most of the species are from South Africa; they have underground tubers and either herbaceous or woody stems. The most widely distributed species is *B. sutherlandii*, which is unusual in that not only has it small tubers, but it can also produce tubercils in the leaf axes. *B. sutherlandii* can over-winter outside even when grown in the Scottish climate; its ability to withstand a range of temperatures and seasonality probably is responsible for its wide distribution. Many of the other species in this clade show some ability to withstand water shortages (and perhaps to regenerate after flash fires?), perenniating though a combination of tubers, a distinct caudex and/or thick woody trunks.

12.3.4.3 Americas

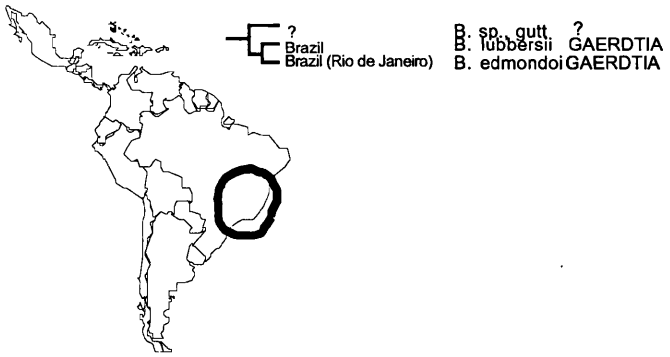
The American relationships from Figure 12.1 are redrawn in Figure 12.14 for convenience.

Figure 12.14: ITS-based relationships of American *Begonia* taxa



Clade 5: The most basal species in America, from section *Gaerdtia*, are found in eastern Brazil (see Figure 12.15).

Figure 12.15: Map of geographic distribution of species in Clade 5, and Clade 5 (America)



These species have somewhat woody stems (they belong to an horticultural class known as 'cane begonias') and generally bifid placentae which are unusual in that ovules are only on the outer surfaces (although *B. edmondoi* has undivided placentae, as do the Southern African species in the clade basal to this, clade 4, section *Augustia*). Species in section *Gaerdtia* are generally reasonably drought-tolerant; *B. lubbersii*, in cultivation, can survive leaf-drop.

Clade 6: Species in this clade are widely distributed, from Brazil to Mexico (see Figure 12.16).

Figure 12.16: Map of geographic distribution of species in Clade 6 and Clade 6 (America)

Because the species in clade 5 are also Brazilian, the biogeography in

clade 6 can be interpreted as a migration north, from Brazil to Mexico. Thus the Brazilian species *B. solananthera* and *B. integerrima* (section *Solananthera*) may represent an early lineage within clade 6 (at least, they form the less speciose half of the basal dichotomy). They are lianescent climbers and, like the species in the previous clade, are unusual in having ovules only on the outer surfaces of their bifid placentae; the fruits have three locules. Within the other half of the dichotomy, *B. violifolia* and *B. imperialis* (section *Weilbachia*) are small rhizomatous herbs; they have bifid placentation with ovules between the branches, but have only two locules in the fruits. All the other species have thick, relatively woody stems, bifid placentation with ovules between the branches, and three locular fruits. Several of the species in this clade possess asymmetric inflorescences.

Clade 7: Clade 7 is broadly Andean, from Ecuador to Bolivia (see Figure 12.17).

Figure 12.17: Map of geographic distribution of species in Clade 7 and Clade 7 (America)

B. maynensis, from Ecuador and Peru, has far smaller male than female flowers. This unusual character is also found in some species from the SE Asian section *Petermannia*. In this cladogram *B. maynensis* resolves as sister to two tuberous Bolivian species, *B. cinnabarina* and *B. boliviensis*. Both these species have orange to red flowers, and appear to show adaptations to bird pollination. The androecium of *B. boliviensis* in particular is very similar to that of the New Guinean *Symbegonia* species;

the styles also show some convergence. Both comparative flower size and the fused, swollen, coloured androecium are likely to be pollinator-specific adaptations. It is remarkable that within one South American clade and within one SE Asian clade the same two distinct morphological adaptations appear to have evolved independently.

B. maynensis has a thick woody stem, while the other species, from the Andean region, have perenniating tubers (they were involved in the crosses which gave rise to the modern range of tuberous 'Elatior' *Begonia* cultivars (Arends, 1970)). *B. bolivienesis* is currently ascribed to section *Barya*; the other two species also ascribed to this section are found in Peru. Although there are no obvious morphological similarities between *B. maynensis* and *B. boliviensis* and *B. cinnabarina*, and despite what appear to be very different pollination syndromes, clade 7 can therefore be circumscribed geographically.

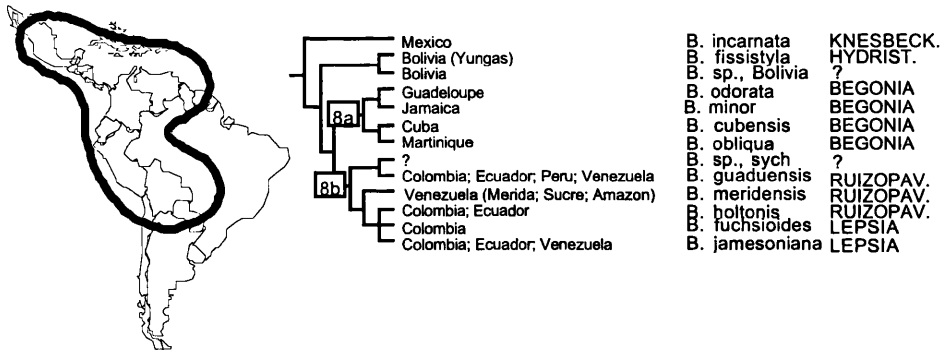
Section *Casparya* could not be sampled, as none of the 24 known species are in cultivation. The section is distributed in Central America and the Andean region, and characterised by fruits dehiscing through the backs of the locules, and being horned rather than winged (Doorenbos, Sosef and de Wilde, 1998). The relationship of this section to other *Begonia* sections is not known; however, when discussing possible pollination syndromes in other America *Begonia*, it may be relevant to point out that nectar production (not otherwise recorded in *Begonia*) has been observed in *B. ferruginea* L.f. (Vogel, 1998), and that morphologially, the flowers bear a striking resemblance to *B. boliviensis*.

Lower oxygen levels and overall temperatures in the higher Andean regions mean that insects are comparatively scarce; this may be what has driven the change from insect to bird pollination in these species. Pollen is no use to hummingbirds; *Begonia* flowers are thought (Vogel, 1998) to mimic other reward-bearing species, particularly *Fuchsia*. (This deception is also how *B. fuchsioides*, in the apparently unrelated section *Lepsia* (clade 8b), is thought to be pollinated). However, it is possible that *B. ferruginea* is not the only nectar-bearing species in *Begonia*; Vogel (1998) found no

specialised nectaries in its flowers; thus there would be no evidence of nectar secretion in herbarium material.

Clade 8: This clade has a wide distribution, through Central and South America (see Figure 12.18).

Figure 12.18: Map of geographic distribution of species in Clade 8 and Clade 8 (America)



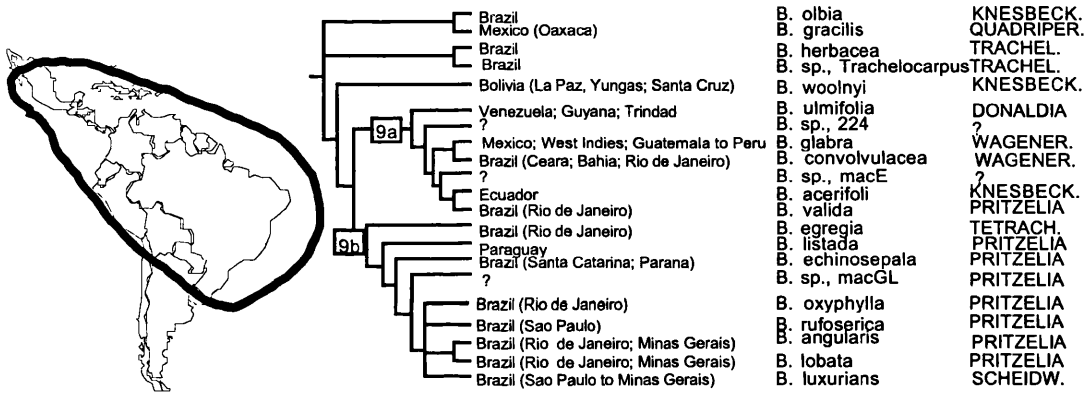
The basal species in this clade is from section *Knesbeckia* (*B. incarnata*). This species has a fleshy/woody stem, and is found in Mexico. Section *Knesbeckia* appears to be polyphyletic; as currently delimited, it includes 50 to 55 species distributed from Mexico to Bolivia. Two Bolivian species (probably) from section *Hydristyles* (the unidentified taxon is known only from a single herbarium sheet, and lacks female flowers) are sister to the rest of the taxa in clade 8. Without far more exhaustive sampling, it is not appropriate to comment on any biogeographic relationships between Mexico and Bolivia, as other species from section *Knesbeckia*, with other distributions, may resolve between these lineages.

Clade 8a (with the included species from section *Begonia*) occurs in the West Indies; its sister clade, **8b** (sections *Ruizopavonia* and *Lepsia*), includes species distributed through Colombia, Ecuador, Peru and Venezuela. The included species from sections *Begonia*, *Ruizopavonia*, *Hydristyles* and *Lepsia* are herbs with small, entire leaves, and are relatively intolerant to water shortages. The leaves tend to be held on a plane to either side of the stem, and in some species each node includes

one smaller and one larger leaf (*B. foliosa*; *B. fuchsioides*, section *Lepsia*).

Clade 9: There is a basal polytomy in this clade, with the positions of *B. olbia* and *B. gracilis* (sections *Knesbeckia* and *Quadriperigonia* respectively), the two species from the morphologically highly distinct section *Trachelocarpus*, and the final major American clade (*B. wollnyi* - *B. luxurians*) unresolved in relation to each other (see Figure 12.19).

Figure 12.19: Map of geographic distribution of species in Clade 9 and Clade 9 (America)



Section *Quadriperigonia* includes 17 to 19 species, mostly from Mexico, and is characterised by a terminal inflorescence and propagation by tubercles (Doorenbos, Sosef & de Wilde, 1998). More species from this section need to be sampled; the current limitation is lack of living material. *B. olbia* is a relatively woody, thick-stemmed species. The species in section *Trachelocarpus* are rhizomatous epiphytes, known only from eastern Brazil. They have distinctive beaked fruits, flowers with an unusual almondy scent, separate male and female inflorescences, distinctive seeds (de Lange & Bouman, 1999) and occur on very long branches in ITS analysis (see Figure 7.4). They also have an indumentum of unusual droplet-shaped glands (described as 'pearl-glands' by Doorenbos, Sosef & de Wilde, 1998). There is nothing in their morphology which gives any hint as to possible relationships; it is unfortunate that this phylogeny does not resolve their position. They have a chromosome number of 56 (counted for four species in the section); however, this is also found in species from

sections *Gaerdtia*, *Solananthera*, *Knesbeckia*, *Quadriperigonia* and *Pritzelia*/*Scheidweileria*, so offers no real clues.
Basal to the last resolved American clade is a fleshy-stemmed, woody, sometimes deciduous species (*B. wollnyi*); sister to *B. wollnyi* there are two clades. **Clade 9a** includes species from sections *Knesbeckia*, *Pritzeila*, *Donaldia* and *Wageneria*, and is distributed from Mexico, through the West Indies and Brazil, to Peru. Species in this clade tend to have a mass of small white flowers in a symmetrical inflorescence, and are shrubby in habit. An exception to the habit is the *B. convolvulacea*/*B. glabra* lineage; these species, from section *Wageneria*, are lianescent, with slightly woody stems. Section *Wageneria* has in the past been incorporated in section *Pritzelia* (e.g. Irmscher) but is separated out by Doorenbos, Sosef and de Wilde (1998) largely on the basis of its scandent habit.

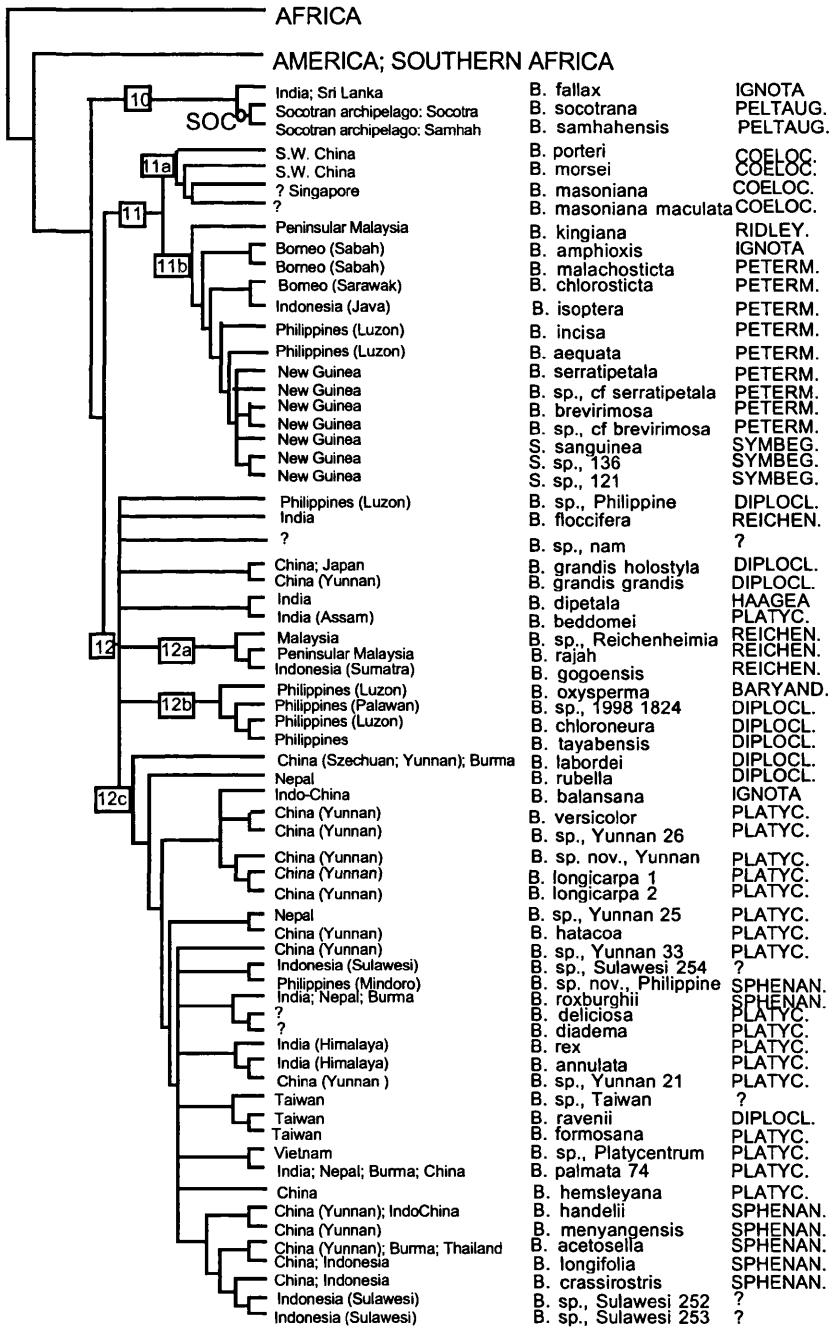
The other clade, **9b**, is almost exclusively Brazilian; with two exceptions (*B. egregia* (*Tetrachia*); *B. luxurians* (*Scheidweileria*)) all species are in section *Pritzelia*. The monotypic section *Tetrachia* has peltate leaves; section *Scheidweileria* is characterised by compound leaves. All other taxa in this clade have simple basifixed leaves. Again, this clade includes many taxa with huge inflorescences of small white flowers (although *B. listada* and *B. echinosepala* have far smaller inflorescences); further, the ovaries of the female flowers are densely hairy. With the exception of *B. listada*, a small, rhizomatous species, these taxa are shrubby; indeed, *B. oxyphylla* and *B. luxurians* can grow to several metres.

The nesting of *B. luxurians* (section *Scheidweileria*) within section *Pritzelia* is highly plausible, given that the key feature differentiating members of section *Scheidweileria* from section *Pritzelia* is leaf morphology (simple versus compound leaves).

12.3.4.4 Asia

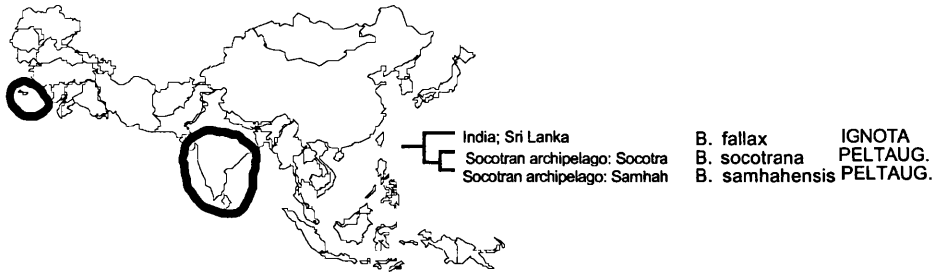
The Asian relationships from Figure 12.1 are redrawn in Figure 12.20 for convenience.

Figure 12.20: ITS-based relationships of Asian *Begonia* taxa



Clade 10: This clade is sister to all the Asian *Begonia* species (see Figure 12.20). For its geographic distribution, see Figure 12.21.

Figure 12.21: Map of geographic distribution of species in Clade 10 and Clade 10 (Asia / Socotra)



The sister group relationship of some Indian and Socotran species, which has already been discussed in relation to Figure 12.7, is unexpected in that most recent authors have considered the Socotran species to be related to southern African species from the section *Augustia* (e.g. van den Berg, 1983, using pollen characters). However, Hooker (1881) suggested that there may be a link between *B. socotrana* and some peltate fleshy-leaved Indian species from section *Reichenheimia*, like *B. floccifera*. Although this relationship was suggested by some analyses (see Chapter 5, Figures 5.5, 5.6, 5.7, 5.8, 5.9, 5.10) it is not resolved in the tree under discussion. *B. fallax*, the sister to the Socotran species, is a shruby plant which was placed in section *Diploclinium* by de Candolle (1864); Doorenbos, Sosef and de Wilde (1998) hint instead at an affinity with section *Haagea*.

Sequence divergence is high between the two Socotran endemics and other species in *Begonia*, and the Socotran species both show morphological adaptations (in the form of perenniating bulbils) to what is a very unusual environment for a *Begonia*, dry and seasonal. The period of reproductive isolation in an unusual environment may mean that the morphological characters suggesting affinity with section *Augustia* are misleading. The Bremer support values for the Socotra/*B. fallax* clade are low (four for the 177-taxon matrix, Figure 7.3; two for the morphological-analysis ITS matrix, Figure 10.9). In the combined ITS/morphology analysis, Figure 10.11, the Socotran species resolve as sister to everything

American, South African and Asian, while *B. fallax* resolves close to section *Haagea* (Bremer support value three). Thus it seems that it is too early to make the definitive statement about the position of this lineage, as adding more data may alter the ways its relationships are reconstructed yet again.

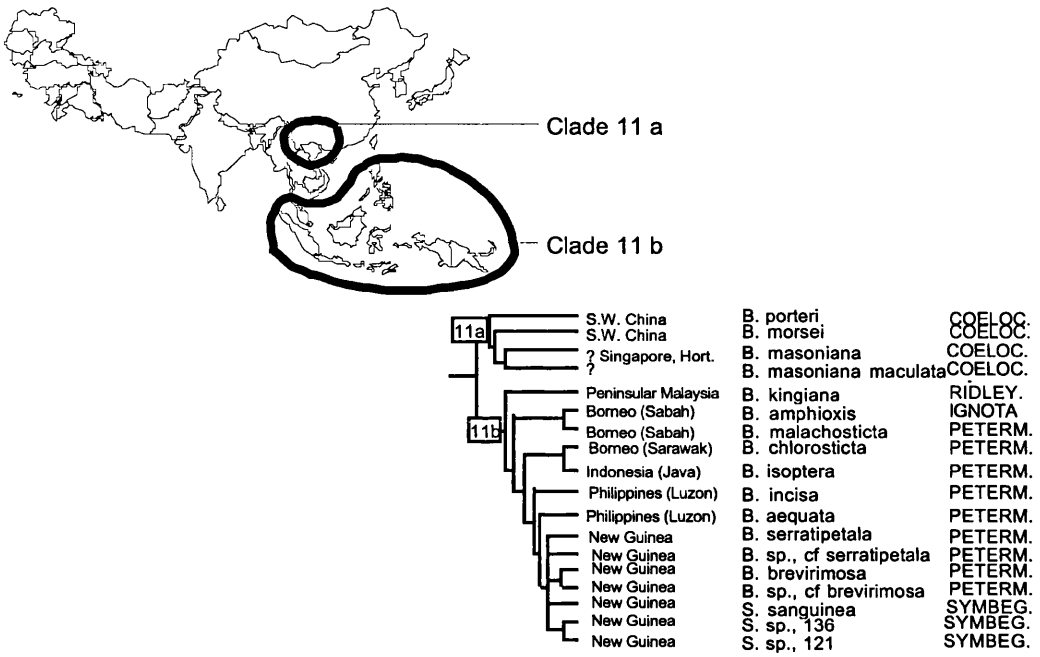
In this ITS phylogeny, the Socotran-Indian clade shares a common ancestor with all the other Asian species of *Begonia* (and *Symbegonia*). It may be that there are western Asian/Indian species which are basal in Asia; collections from across India are sorely needed. *B. samhahensis* was only discovered in 1995; perhaps more species of *Begonia* remain to be discovered in such atypical locations.

Clades 11 and 12: Some of the biogeographic patterns in SE Asia are "difficult to relate simply to geology" - such as why the distance between Borneo and Sulawesi (across Wallace's Line) appears difficult for plant groups to cross.

From Figure 12.20, it can be seen that all the sampled *Begonia* species from Borneo are in section *Petermannia*, and in the same lineage as taxa from Peninsular Malaysia (basal), Java, Luzon and New Guinea (clade 11b). However, the three sampled taxa from Sulawesi resolve within the predominantly Chinese *Platycentrum* clade (clade 12c). This is despite the morphological similarity of one of the Sulawesi taxa (no. 254) to section *Petermannia*. The analyses here are sectional rather than species level, so the true patterns can only be hinted at, and it possible that there are other Sulawesi species which have close relatives in Borneo.

Clade 11: Most of the species in this clade are Malesian (see Figure 12.22). The two halves of clade 11 are highly unbalanced, with c. 12 species in clade 11 a, and c. 200 in clade 11 b (assuming that the sections *Coelocentrum* and *Petermannia* are monophyletic, as the initial results presented here suggest).

Figure 12.22: Map of geographic distribution of species in Clade 11 and Clade 11 (Asia)

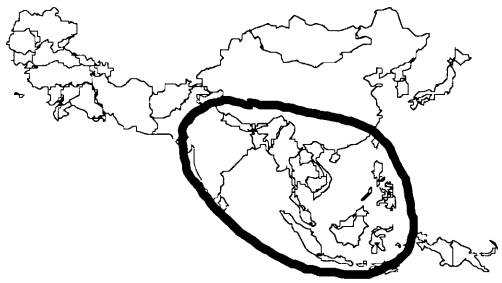


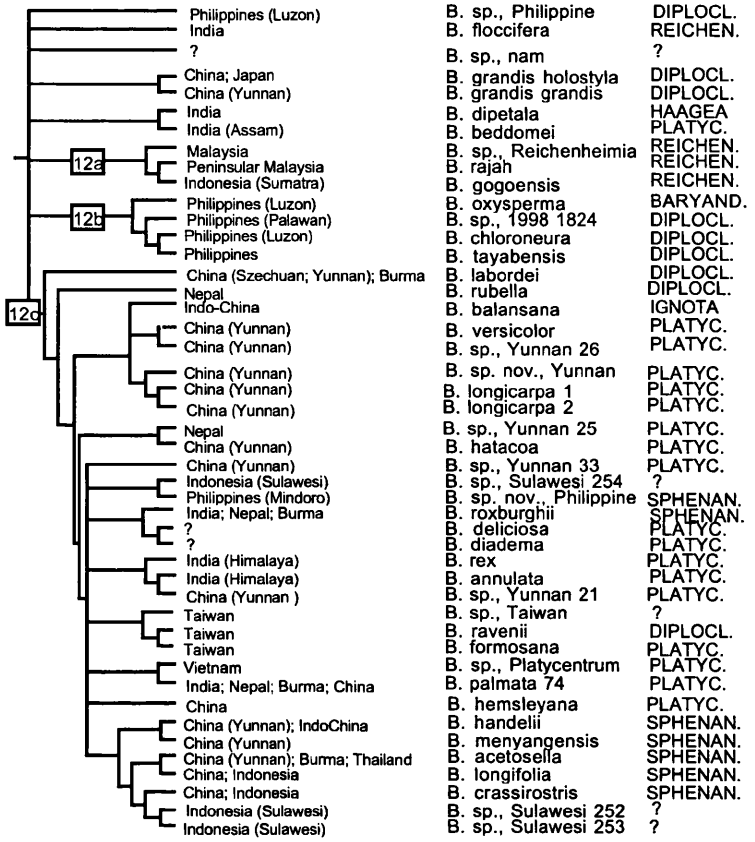
Clade 11a: This clade consists of species from section *Coelocentrum*; the section is distributed in south-western China and VietNam. *B. masoniana* was 'discovered' in a botanic garden in Singapore; there are no records of where it came from (Mason, 1957; Irmscher, 1959). Because every other known species in this section is from a relatively small area of northern VietNam / Yunnan / Guangxi it seems probable that *B. masoniana* is also from this region. The section is characterised by having unilocular ovaries with parietal placentation. This has been considered (e.g. Jin & Wang, 1994) as a primitive condition in *Begonia*, linking the species with some from section *Mezieria* in Africa, and therefore it has been suggested that species from section *Coelocentrum* are basal in Asia. There is no evidence for the primitivity of section *Coelocentrum* here (as Irmscher, 1939, suspected): unilocular ovaries appear to have evolved separately in Africa and Asia (where they have evolved at least twice, once in section *Coelocentrum* and once in the morphologically peculiar *Petermannia*-relative, *B. amphioxis*).

Clade 11b: The lineage of section *Petermannia* may have entered the Malesian islands via the Malaysian archipelago; its ancestor presumably migrated south from the continent. The two species on Sabah (including *B. amphioxis* as a member of the section) are monophyletic; the next clade consists of species from Sarawak and Java. Species from Luzon are paraphyletic, due to a monophyletic clade of seven taxa from New Guinea (including species of *Begonia* and of *Symbegonia*). What evidence there is suggests that oceans provide barriers to *Begonia* dispersal in this lineage at least, with a high degree of island endemism, and apparently no species crossing the Torres Strait (between New Guinea and Australia) or the Timor Sea to Australia. Of course, species may have crossed into Australia and subsequently suffered extinction; equally, there may be as-yet undiscovered *Begonia* on the continent. The former is thought unlikely as we are dealing with (comparatively) recent events; the latter, in part, because there is a vibrant amateur *Begonia* group in Queensland, Australia, which would surely have spotted them!

Clade 12: Lack of resolution at the base of this major Asian clade creates problems with interpretation. The relationships between clades from the Philippines, India, China, Malaysia / Sumatra and the major '*Platycentrum*' clade are unresolved (see Figure 12.23).

Figure 12.23: Map of geographic distribution of species in Clade 12 and Clade 12 (Asia)

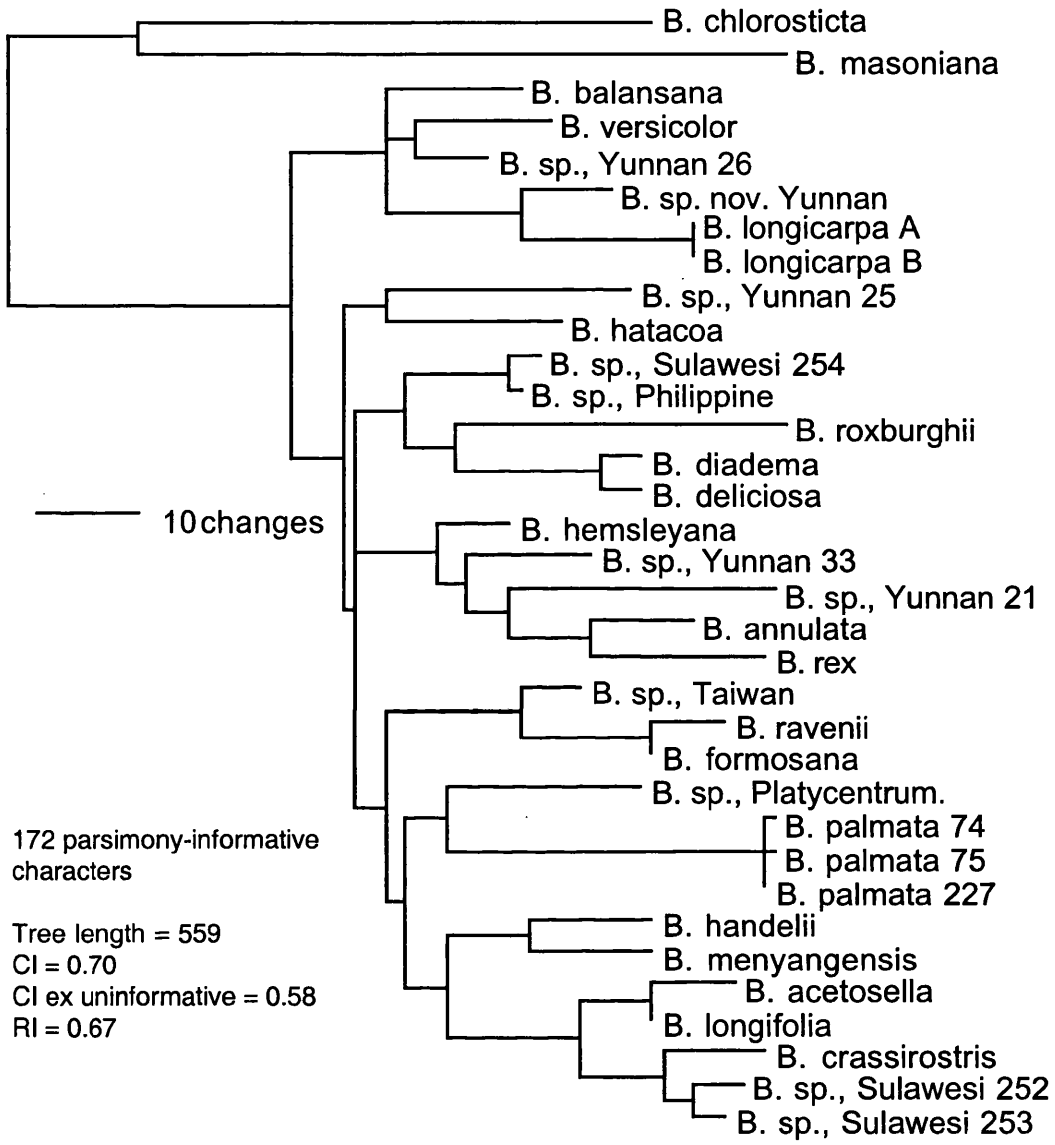




One clade (**12a**; section *Reichenheimia*) suggests migration from Malaysia to Sumatra; another (**12b**) shows the monophyly of four Philippine taxa. Within the largest resolved clade (**12c**), the basal taxa are from China, Indo-China, Burma and Nepal. Many of the species within this clade are unresolved; this resolution is not the result of conflict between many different most parsimonious trees in the consensus tree (the topology has been taken from the compartment analysis of these taxa, which resulted in only ten most parsimonious trees). Rather, the lack of resolution is due to low levels of sequence divergence, which may be due to a rapid radiation (faster than ITS can track).

Harking back to the phylogram presented for the compartment analysis, section *Platycentrum* (Figure 7.15; presented again here as Figure 12.24) and for the same taxa within the complete culled ITS analysis phylogram (Figure 7.4), there appears to be a rapid radiation (with little internal resolution) followed by a period of lineage differentiation.

Figure 12.24: Phylogram for *Platycentrum* clade compartment analysis (copied from Figure 7.14)

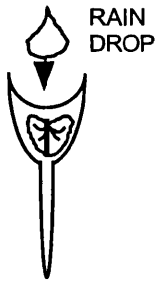


Many of the species in this clade are currently found around China - India - Himalaya - Burma - VietNam. The uplift of the Himalayas and Tibet continued long after the initial collision of India and Asia (estimates of a data for this collision vary, from the late Palaeocene, c. 60 Ma (Powell & Conaghan, 1973) to the Eocene (c. 40 Ma, Molnar & Tapponnier, 1975)) and had a significant impact on climate (Hallam, 1994): grassland spread where once was forest, and xerophytic scrub developed in rain-shadow areas. The Tibetan plateau reached its present elevation only 8 Ma; it is

thought that the "late Cenozoic climatic system was stongly influenced by the Tibetan plateau" (Windley, 1995), which had far-reaching effects including increases in the intensity of the Indian monsoon and changes in the vegetation patterns in Pakistan.

It is possible that the radiation in this group correlates with this Himalayan uplift and its associated effect on climate, which may have led to the fragmentation of ancestral ranges and subsequent divergence in isolation. However, the '*Platycentrum*' clade also has a number of morphological character changes associated with it, which could be interpreted as 'key innovations'. The majority of the species in this clade have two-locular ovaries, while the majority of *Begonia* species overall have three locules. This morphological change appears to correlate with the mode of seed dispersal. Seeds from the three-locular fruits are wind dispersed, while, in these two-locular species, the fruit recurves and the two smaller wings on the fruit form a cup which catches raindrops, shaking the seeds loose (de Lange & Bouman, 1999) (see Figure 12.25).

Figure 12.25: T.S., two-locular fruit, section *Platycentrum*



Assuming that rain-splash dispersal is advantageous over wind dispersal in some situations (wind dispersal tends to rely on dry fruit and may be problematic in a monsoon climate, for example) the 'innovation' of these two-locular fruits could have allowed the lineage which possessed it to radiate.

Further evidence for evolution of seed dispersal in this clade is a probable transition from rain-splash to zoochory, in the fleshy fruited *Sphenanthera* species.

Coming back to the question of Himalayan uplift versus morphological key innovations as a driving force for *Platycentrum* radiation, it is worth asking whether molecular clock estimates for this clade correlate with the timing of the Himalayan uplift.

The ITS phylogram for the *Platycentrum* compartment (Figure 7.15) has branch lengths ranging from 15 (polytomy to *B. hemsleyana*) to 40 (polytomy to *B. palmata*; polytomy to *B. roxburghii*). The relative rate ratio 'r' between these species is 0.375; only if r = 1 is there evidence for a time-calibrated molecular clock (Doyle & Gaut, 2000). N.B. the alignment of sequences from this region was not ambiguous and no positions were excluded due to uncertainty. Another DNA region may behave in a more clock-like manner, and thus be more suitable for molecular clock based hypotheses.

The uncorrected pairwise divergence within this *Platycentrum* / *Sphenanthera* clade range from 1.4% (between closely related species) to 11.5% (across the unresolved part of the tree). Using a conservative estimate of 0.79 - 1.57% nucleotide substitutions per million years (Sang et al., 1994; Sang, Crawford & Stuessy, 1995) provides dates in the order of 0.89 - 0.45 Ma between *B. longifolia* and *B. acetosella,* i.e. Pleistocene; 8 - 3.5 Ma between *B. roxburghii* and *B. palmata,* i.e. late Miocene. If the radiation of the *Platycentrum* clade did occur in the Miocene, the initial collision of India and Asia would have happened over 20 million years previously, and Tibet would have more or less reached its present elevation. However, given that ITS does not appear to evolve in a clock-like fashion in *Begonia*, there is little basis for accepting these dates as evidence.

Within the '*Platycentrum*' clade, species have radiated across Sulawesi / Philippines, China, the Himalayan region, Taiwan, Vietnam, Burma, Thailand and Indonesia. Certainly within the species attributed to section *Sphenanthera* (*B. handelii* - *B. crassirostris*) a migration appears to have occurred from China to Indonesia; a separate southern migration into Malesia is required to explain the 'Sulawesi no. 254' / 'Philippine sp. nov.' clade.

Dioecy is widespread in angiosperms, being present in 37 of Engler and Prantl's 51 orders (Bawa, 1980) (and remains highly polyphyletic in the more phylogenetic classifications produced by the A.P.G. e.g. Soltis, Soltis & Chase, 1999). The section *Sphenanthera* contains several species which are dioecious[15], e.g. *B. roxburghii*, *B. menyangensis* and *B. handelii*, which are included in this ITS phylogeny. The section does not appear to be monophyletic, as the dioecious species resolve in two separate clades (suggesting two separate gains of dioecy). These two clades are also supported by gross morphology, as *B. roxburghii* is a 'cane' *Begonia*, with tall upright stems to c. one metre, while *B. menyangensis* and *B. handelii* are more or less acaulescent, producing masses of large, strongly scented pale flowers near the ground surface. In the light of Bawa's (1980) suggestion that dioecy correlates with fleshy-fruitedness, it is interesting that dioecy is found only among the fleshy-fruited members of the *Platycentrum* / *Sphenanthera* clade. The majority of dioecious species are reported to be insect pollinated and animal dispersed, particularly in the tropics (Bawa, 1980). It is not known what disperses the seeds of the fleshy-fruited species in section *Sphenanthera*.

It would be interesting to compare population genetic structure of nucleotide and organelle markers in order to estimate pollen to seed flow ratios in sympatric *Begonia* species with different fruit types (e.g. zoochorous, rain splash dispersed, wind dispersed).

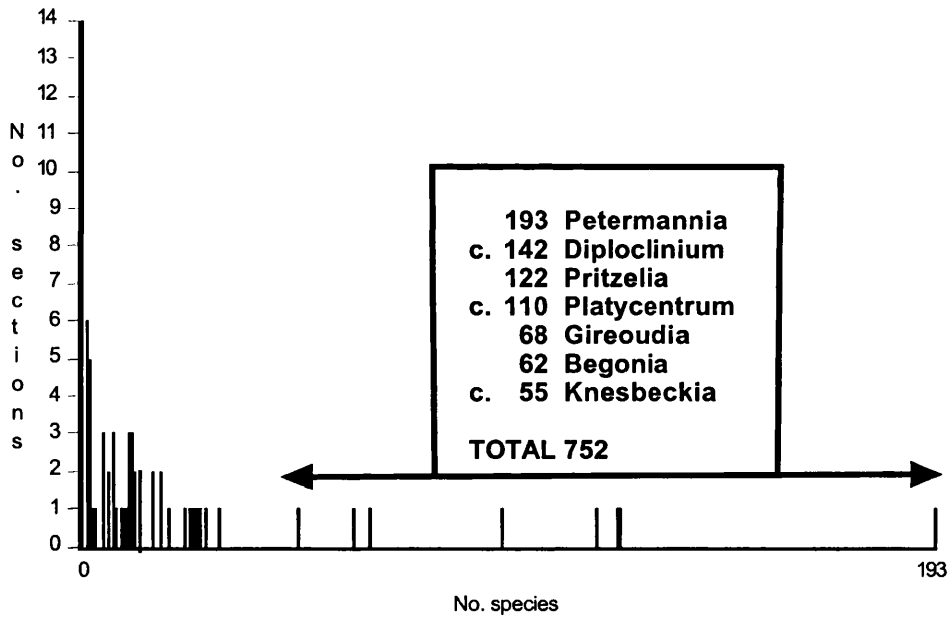
Without including taxa from the full distribution of *Begonia* in this region (e.g. species from Fiji, Haimahera) detailed island biogeographic conclusions cannot be made. However, *Begonia*, a genus with a range of narrow endemics (and very few widely dispersed species), appears eminently suited to biogeographic considerations; the presence of at least two unrelated clades of taxa on the south east Asian islands offers the possibility of using cladistic biogeography to compare and contrast independent distribution patterns.

[15] It can be difficult to determine dioecy based on herbarium collections, as plants may produce separate male and female inflorescences, which can be separated temporally. However, these reports are based on observations over many years, of plants in cultivation.

12.4 Why is *Begonia* a large genus?

The traditional sections in the genus *Begonia* conform to the 'hollow curve' distribution described in the first chapter of this thesis. Simplistically, this suggests that a lot of things look very similar (therefore are included in a few big sections) and a few things look very different (therefore are included in small to monotypic sections). Returning to the graph shown in Chapter 4 (Figure 4.1) it can be seen that over half the known species of *Begonia* are contained in only seven sections; the remainder are contained in 55 sections (Figure 12.26).

Figure 12.26: The number of species per section for *Begonia* (from Figure 4.1)



However, given that many of these sections (particularly sections *Knesbeckia* and *Diploclinium*) are not good monophyletic groups, it is clearly preferable to consider phylogeny over traditional classification.

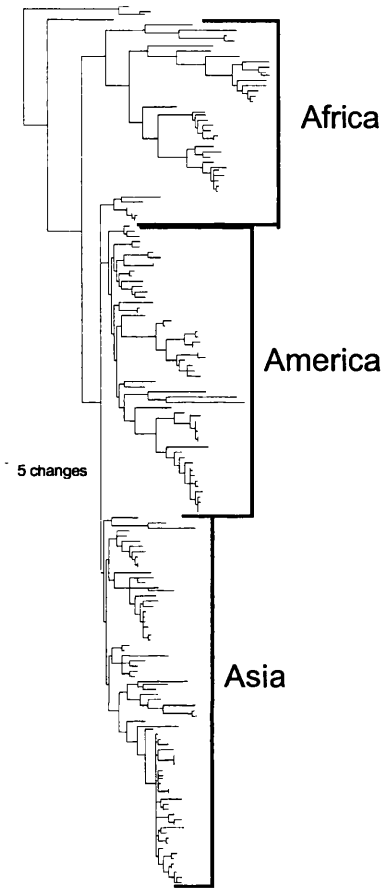
The first question is whether *Begonia* is a big genus because it is old (Willis, 1922, Age and Area hypothesis) or because it is young (Cronk, 1989, Relict hypothesis). The evidence strongly favours Cronk's hypothesis: the less basal clades (e.g. in America and Asia) are more species rich, more widely distributed and less differentiated morphologically; the older lineages (in Africa) are less speciose and better

differentiated morphologically[16]. Following Cronk (1989), the younger lineages in *Begonia* could be described as being in 'bloom' phase, while the African species may have been 'depleted by extinction', leaving them phenetically distinctive.

Looking back to the ITS phylogram (Figure 7.4) allows consideration of tree 'stemminess' (see Figures 1.2 to 1.5 for description). Figure 12.27 is reproduced from part of Figure 7.24, and shows branch lengths for the culled manual alignment ITS analysis. Most of the species included in the African clades are clustered relatively close to the ends of long branches, while the American and Asian clades, on the other hand, generally have shorter internal branch lengths. Although this is a broad generalisation, it supports the view that the African lineages are more ancient and divergent than the (younger) lineages in Asia and America. From a simplistic perspective, a casual glance at the ITS matrix (Appendix, 14.7) shows far greater difficulty in aligning African taxa with each other than aligning all the Asian and American species together.

[16] although, as Africa is also the best studied region taxonomically, it could be argued that these African sections are better delimited.

Figure 12.27: Tree shape, from a phylogram produced by analysis of the manually aligned, culled ITS data set (reproduced from Figure 7.24)



Accepting that African lineages show greater divergence than Asian and American lineages, it is worth asking whether the extant species in the different continents show a similar trend. One way of assessing this (assuming that there is some sort of regularity to sequence nucleotide divergences, i.e. some form of molecular clock) is comparing pairwise divergences between species in different lineages. Clearly there are risks of erroneous inference in such an exercise, due to differential taxon sampling within clades, and the delimitation of the clades selected for comparison. The clades which have been used correspond to monophyletic groups, separated by an obvious morphological disjunction. Comparing traditional sections was not possible due to the paraphyletic nature of the larger sections in the ITS phylogeny. Relative sampling densities are discussed below.

For section *Loasibegonia* ITS sequences were obtained for nine of its 19 species. Unfortunately, there is a possibility that the section is paraphyletic, with *Scutobegonia* nested within it. Only one of the 21 recognised species in section *Scutobegonia* could be included (i.e. ten species have been sampled from a clade of perhaps 40 species). The uncorrected pairwise divergences in this clade range from 13% to 1.3% [17]. Another African clade, *Tetraphila / Squamibegonia*, has been sampled for ten out of c. 30 species. Divergences range from 10.3% to 1.1% within the clade.

These can be compared to values obtained for lineages in Asia and America. Only a very small proportion of the species in these lineages have been sampled (12 - 13 species out of over 200, *Petermannia / Symbegonia* clade; 28 out of over 130, *Platycentrum / Sphenanthera* clade; 15 out of over 160, *Pritzelia / Weilbachia / Donaldia / Scheidweileria* clade). The uncorrected pairwise divergence values are 0.6% to 8.1% (*Petermannia / Symbegonia* clade); 0.1% to 9.5% (*Platycentrum / Sphenanthera* clade) and 0% to 20% (*Pritzelia / Weilbachia / Donaldia / Scheidweileria* clade).

Divergence ranges are generally higher in the two African lineages, but there is considerable overlap - enough to say that there must be equivalently recent species in all clades. The lower sampling levels in Asia and America mean that the extremes of ranges are less likely to have been included; therefore the true overlap could be higher. Although there are many weaknesses in this informal analysis, a preliminary generalisation is that *Begonia* species on Africa are unlikely to be orders of magnitude older than species in Asia and America (and in fact the species are probably of roughly comparable ages despite the ages of the lineages). This is being found increasingly in plant phylogenetics - even ancient lineages are currently composed of modern species (e.g. *Selaginella*, Bateman pers. comm., 2000; *Araucaria*, Setoguchi et al., 1998 - see section 1.4.4 B).

The other major factor in 'tree-shape', which was mentioned in the introduction and which tells us something about diversification, is balance.

[17] Because the alignment of all 177 taxa in the global ITS analysis is extremely gappy and, in places, ambiguous, uncorrected pairwise divergence values from the Compartment analyses (section 7.3.3) are cited.

However, because there are problems with the monophyly of some of the traditional sections, it can be difficult to know how many species are truly in each clade on the tree.

Figure 12.28 gives very approximate figures for species number per clade, taken from Doorenbos, Sosef and de Wilde (1998) and reliant on the assumption that no members of the sections included in each clade truly belong in another clade. This is, however, false for the African section *Rostrobegonia* (clade 1) which has one member (*B. sonderana*) which resolves in the *Augustia* clade (clade 4). Another highly problematic section is American, *Knesbeckia*, which resolves in several clades (clades 7, 8 and 9). Due to the uncertaintly surrounding their placement, its 50 - 55 species have not been added onto the totals for any clade (which partly explains why the total number of species in the South African / American clade (c. 612) is not the total of the numbers of species in individual clades (477); the rest of these 'missing' species belong to sections which have not been included in the analysis). Unsampled sections also explain the discrepancy between the total Asian species (c. 645) and the sum of the clades in Asia / Socotra (463). Although section *Diploclinium* is also polyphyletic, all its species fall within clade 12 in analyses so far.

Figure 12.28: ITS phylogeny of Begoniaceae, with approximate species no.

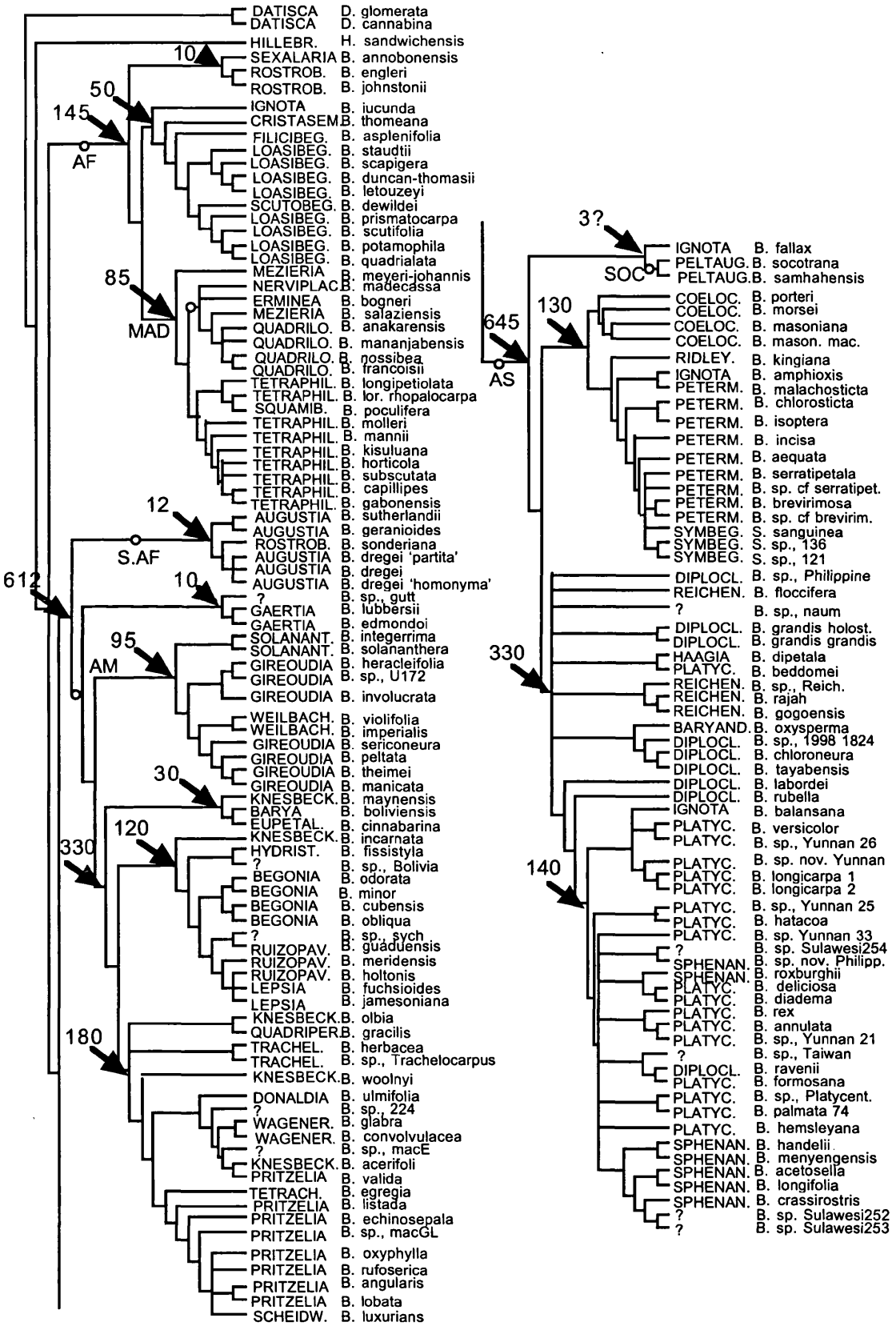
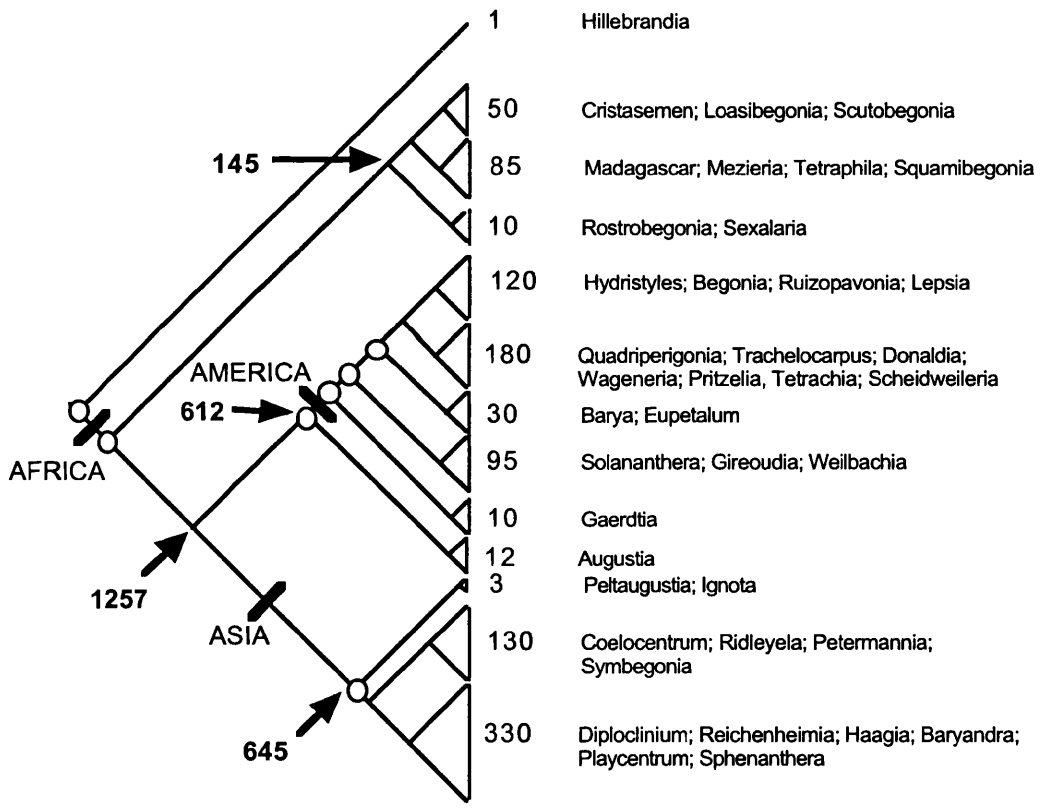


Figure 12.29 provides a summary of the numbers of species per clade (with provisos as to accuracy, as mentioned previously).

Following Guyer and Slowinski's (1993) definition of 'unbalanced' (having 90% or more of the total number of taxa along the more diverse branch of the dichotomy), unbalanced nodes are marked onto Figure 12.29 with open circles.

Figure 12.29: Summary diagram of species number per clade, for the 12 clades described previously



(clades are not drawn to scale)

The basal nodes in Begoniaceae, and also in both the American and the Asian clades, are unbalanced. This means that there are differences in the relative diversification and / or extinction rates between sister groups.

It is not possible to prove that extinction has occurred purely on the basis of a phylogeny; however it seems likely in the case of a genus which is almost

exclusively confined to moist tropical / subtropical forest, and which has been in Africa over a period when forests are thought to have contracted and aridification occurred. If the genus did migrate north from Africa into Eurasia, extinctions are also likely to have occurred there, as it is another region which has suffered aridification, and temperature cooling. Thus extinction could explain the unbalanced nature of many of the nodes within this ITS phylogeny (e.g. the African / rest of *Begonia* dichotomy unbalanced due to increased extinction in the African lineage; the unbalance in the Asian lineage due to increased extinction in the North of Africa and Eurasia).

12.5 Overview: The Evolution of *Begonia*

There is no hard evidence about the age of the genus *Begonia*: there are no fossils within the family Begoniaceae (and only the 55 Ma Tetramelaceae fossil cited in Wagstaff & Dawson, 2000, from a closely related family). Rough estimates based on sequence divergence are also difficult: ITS is very divergent between African lineages (which leads to alignment problems, which means that pairwise differences are highly unreliable between unrelated clades). It seems likely, however, that the alignment difficulties between African clades, and between Africa and (Asia/ America) confirm that African lineages are older than the other lineages in *Begonia*. (It is possible that Broulliet's 1995 pers. comm., that ITS is slow in *Begonia*, was due to not including species from Africa in his pilot study).

The similar numbers of species in America and Asia are interesting. Either *Begonia* has equivalent speciation rates in both continents and has been in each for similar amounts of time, or speciation and / or extinction rates differ on each continent and the balance is purely down to chance. There are certainly many undescribed species in Asia; I am less familiar with America. It may be that, if these could be taken into account, Asia would have more species.

Key innovations are difficult to prove; the presence of them should correlate with increased species richness (Dodd, Silvertown & Chase, 1999).

However, unbalanced clades appear to correlate to geographic rather than morphological changes; thus it appears that the key factor generating *Begonia* species is radiation into new habitats. Whether this is adaptive, into new niches, is debatable. *Begonia* do fill several obvious ecological niches (e.g. geophyte, epiphyte) but the majority of species grow happily under very similar conditions in cultivation, and the evidence for radical pollinator or disperser specialisation is slim (e.g. bird pollination in *Symbegonia*, *Barya* and *Casparya*; ant dispersal in species of *Tetraphila* (Bouman & de Lange, 1982); bird dispersal in species of *Tetraphila*, *Squamibegonia* and *Mezieria*; rain-splash dispersal in *Platycentrum*).

One fairly simple measure, which does not rely absolutely on a molecluar clock, is lineage diversification rate (ln N / t, where t = the age of a clade, and N = the number of species in it) (Wojchiechowski, Sanderson & Hu, 1999) (as discussed in Chapter 1, section 1.5.1). The genus *Begonia* contains c. 1400 species; I have been reluctant to put an absolute age on it; based on the datings of land bridges I have hypothesised *Begonia* might have used and Tetramelaceae fossils, a rough (and very broad) estimate would be 60 to 35 million years ago (probably Eocene). These values give diversification rates of between 0.12 and 0.21 (rates for the whole of *Begonia* will be biased downwards, as I expect extinctions to have occurred in Africa and Eurasia). Erikkson and Bremer's (1992) median value for continental plant families is 0.12 spp/Ma; clearly without a narrower date for the origin of *Begonia* it is not possible to say whether the genus shows above-average diversification.

The estimated dates for the Asian and American clades are purely speculative, based on biogeographic hypotheses with little corroborative evidence. Assuming *Begonia* did leave Africa along the Walvis ridge, the lineage must have separated by 35 Ma. The estimated diversification rate for the c. 600 American species is then 0.18 spp/Ma. A date of c. 25 Ma for the c. 645 Asian species would give a diversification rate of 0.26 spp/Ma. None of these values approach the estimation (Wojciechowski, Sanderson & Hu, 1999) of 0.71 spp/Ma for *Astragalus*. From these calculations it would appear that lineages in *Begonia* have far lower diversification rates than the

extremely large genus *Astragalus*. To obtain a value as high as c. 0.71 for 1400 species, the genus would have had to have originated in the mid Miocene, c. 10.2 Ma).

What are lacking are studies on speciation within *Begonia*. Hybridisation and polyploidy have been discussed in Chapter 11 and may contribute towards isolation and diversification. In addition, the geographical constraints to clade distribution indicate that allopatric speciation may be important in *Begonia*. Supporting evidence, of limited dispersal providing the potential for reproductive isolation by distance, comes from population level studies by Matolweni, Balkwill and McLellan (2000). They showed differential allelic fixation and high Fst values (the proportion of variation partitioned between populations) over even small geographic distances. Further studies are required in order to assess the scale over which populations become reproductively isolated, and the potential for localised adaptation and differentiation, particularly in the face of the homogenising effect of gene flow from sympatric relatives.

12.6 Taxonomic Changes Recommended

The aim of this thesis was to produce a phylogeny for the genus *Begonia*, and consequently the taxon sampling was directed at covering the morphological and geographical range of the genus. Although there are some sections (e.g. *Loasibegonia*) which are well represented, the levels of sampling per section are not adequate to produce a robust species level revision. Also, further analyses with increased sampling may reveal hidden homoplasy (Sanderson, 1990) and alter the overall topology. Furthermore, many of the taxa are represented by only one molecular data set (ITS); it is possible that alternate data sets (perhaps from different genomes) will affect the clades resolved. Thus any taxonomic comments are to be taken as preliminary.

However, in gross and micro- morphology there are some convincing examples of convergence in *Begonia* - for example, the similarity of flowers (particularly anther and style types) in *Symbegonia* in Asia and section *Barya* in South America, associated with bird pollination; also the homoplasy in endothecial cell types between section *Petermannia* in Asia and section *Solananthera* in South America. These cases, between widely phylogenetically separate taxa, should act as a warning: morphological convergences between closely related taxa have far more potential to mislead and may contribute to lack of monophyly in traditional sections.

12.6.1 Genera: There is a notable exception to the 'preliminary' nature of most of the comments in this chapter. *Symbegonia* has not always received generic status (e.g. Brummitt, 1992), and on the basis of results here and in other studies, should clearly be regarded, at most, as a section of *Begonia*. Although the ITS phylogeny does not resolve *Symbegonia* as monophyletic, and includes it within the section *Petermannia*, the morphological synapomorphies of the erstwhile genus (its extremely distinctive androecium and unique anther endothecial cells) suggest that adding more sequence data may recover its monophyly (i.e. that the problem is the lack of ITS characters in this part of the phylogeny). The inclusion of *Symbegonia* in *Begonia* does not just rest on ITS and 26S data alone - *trnC-trnD* (Badcock, 1998), *rbcL* and 18S (Swensen, Luthi &

Rieseberg, 1998) sequence data and morphological data all place *Symbegonia* deeply within *Begonia*.

Some may argue against the sinking of this genus into *Begonia*, given that it is easily recognisable in the field as a separate entity (Sands, pers. comm., 2000). However, not only are there several other sections in the genus for which such an argument could be made (*Tetraphila* (Africa), *Peltaugustia* (Socotran archipelago) and *Trachelocarpus* (America) for example), but treating *Symbegonia* as a separate genus implies that it is comparable to other genera - anyone unfamiliar with the true relationships may wonder how one genus in the Begoniaceae is so species rich (with around 1400 species) while another contains only around 14 species. Obviously, in the light of phylogeny, such comparisons are meaningless.

One important consideration is the role sections should have: are they practical subdivisions or should they reflect common evolutionary history? All the available evidence suggests that the taxon *Symbegonia* has evolved from within section *Petermannia*, rendering *Petermannia* paraphyletic if *Symbegonia* is even given sectional status. As yet, *Begonia* classifications do not go below the sectional level; with about 200 species in section *Petermannia* (and the additional c. 14 from *Symbegonia*) there is clearly a need for some subsectional division. Until this is in place, 'losing' *Symbegonia* amongst this huge morass of species would be foolish and its sectional status should therefore be upheld on this purely practical criterion.

The Begoniaceae appear to be a good natural group; within the family, the characters which separate *Hillebrandia* and *Begonia* are clear-cut, relating to tepal number, to the mode of dehiscence of the fruits, and to the position of the ovary. No such suite can be cited to distinguish sections of *Begonia*; although it may be possible to hive off some of the distinct African sections as separate genera, the obvious morphological divisions would leave *Begonia* para- (or even poly-) phyletic. Further, even given the (perhaps laudable) aim of restricting *Begonia* to a more manageable size, excising Africa would not accomplish it - although most of the morphological and

molecular divergence of the genus is African, most of the species are Asian and American. There are c. 140 species in Africa, and c. 1260 in Asia and America. The type, *B. obliqua* L., is an American species, section *Begonia*; therefore the name would remain with the exuberance.

12.6.2 Madagascan species: The monophyly of the Madagascan species is certainly strongly suggested by ITS and 26S data. However, incorporating all Madagascan species within a single section may not be the best way of dealing with them. Although this would be phylogenetically informative, this, although desirable, is not the sole purpose of sectional classifications. There are at least 48 species of *Begonia* on Madagascar (de Lange & Bouman, 1992) (judging by a recent RBG Kew expedition to Ambatoraky reserve on the northeast of the island (Baker, pers. comm., 2000) there are many more to be found), and one purpose of subgeneric classification must be to facilitate identification of these in the field. Due to sampling limitations, it is not possible to say whether the sections which are currently recognised are monophyletic, as only one individual was sampled from sections *Erminea* and *Nerviplacentaria*. Further, there are some doubts as to the usefulness of the current sections; the most recent flora of Madagascar (Keraudren-Aymonin, 1983) abandoned sectional treatment as untenable. However, there is certainly a place for some form of subdivision (whether sections or subsections) within the Madagascan species; a thorough cladistic analysis of the species on the island is required before this can be attempted.

12.6.3 African species: Section *Mezieria* resolves as polyphyletic in this study, despite only two species from the section being included. Suggestions that there are problems with this section can be found in the taxonomic revision published by Klazenga, de Wilde and Quene (1994). They produce a morphological cladogram based on eight characters, which gives four equally parsimonious trees. Of these trees, they accept the only one which is fully resolved (which gives a monophyletic *Mezieria*). However, the two characters which appear as synapomorphies for the section show reversals within the section. Furthermore, in two of the other equally parsimonious trees *Mezieria* is not monophyletic. Their study

includes six species within *Mezieria*, and two outgroups (section *Baccabegonia* and section *Squamibegonia*). Clearly, in the light of the ITS phylogeny, this is an inadequate test of monophyly; such a study should also contain the Madagascan species, as well as species from section *Tetraphila*. Such a study is currently being undertaken by Vanessa Plana, RBGE.

Without including more species from *Mezieria* in analyses, it is not possible to draw many conclusions, and perhaps greater sampling will lead to topological changes. However, on the basis of the ITS analyses presented here, it appears that *B. meyeri-johannis* should not be included in section *Mezieria* (the type of which is *B. salaziensis*). The inclusion of *B. meyeri-johannis* within section *Mezieria* has previously been questioned, because it differs from the other species by "its 5-locular ovaries with 5 styles, unisexual inflorescences and lianoid habit" (Klazenga, de Wilde & Quene, 1994, p. 310) (although they drew the opposite conclusion, that there is "neither a need nor a justification to establish a new section to accommodate *B. meyeri-johannis*"; loc. cit., p. 310).

There are currently 63 recognised sections and c. 1400 species in *Begonia* (mean section size 22.2; a more accurate representation of section size is in the graph in Chapter 4, Figure 4.1). In the interests of monophyly, this high number of sections should be reduced (e.g. the loss of sections *Scheidweileria*, *Squamibegonia*, *Lepsia*, *Baryandra*). In some cases this would render large and unwieldy sections even larger and less wieldy; in these cases, species level revisions are required to search for workable subdivisions (e.g. subsections). This project is limited to circumscribing suitable clades as starting points for such revisions (e.g. clades 1 to 12, as described in this chapter and in chapter 11) and to provide a framework for future workers to evaluate the phylogenetic placement and taxonomic affinities of as yet unsampled species.

12.7 Summary

While the geographic origins of the family Begoniaceae cannot be determined on the basis of these analyses, the oldest lineages in the genus *Begonia* appear to be found on the African continent. From Africa *Begonia* probably dispersed to, and radiated in, Asia and South America. In South America, it seems likely that *Begonia* arrived in Brazil, and that several lineages have migrated north and west from Brazil; independent clades have members in Mexico and in the Andean region. Within Asia there are two main clades, one predominantly continental and one predominantly on the Malesian islands. No taxa are thought to have reached Australia. Given the hypothesis that *Begonia* left Africa by a Walvis Ridge/Rio Grande Rise route to the west (the Walvis ridge would have been completely submerged by the end of the Eocene, about 35 million years ago), and into Eurasia to the north east (less than 23 million years ago), the lineage '*Begonia*' must have originated, at the latest, during the Eocene (between 56.5 and 35.4 million years ago).

13 References

Arends, J. C. (1970). Somatic chromosome numbers in 'elatior' begonias. *Mededelingen Landbouwhogeschool Wageningen Nederland* **70-20**: 1-18.

Arends, J. C. (1992). The biosystematics of *Begonia squamulosa* Hook.f. and affiliated species in section *Tetraphila* A.DC. *Wageningen Agricultural University Papers* **91-6**.

Arends, J. C. (1985). Karyology of African Begonias. *Acta Botanica Neerlandica* **34**: 230.

Azuma, T., Kajita, T., Yokoyama, J., and Ohashi, H. (2000). Phylogenetic relationships of *Salix* (Salicaceae) based on *rbcL* sequence data. *American Journal of Botany* **87**: 67-75.

Backlund, A., and Bremer, B. (1998). To be or not to be - principles of classification and monotypic plant families. *Taxon* **47**: 391-400.

Badcock, Z. (1998). A phylogenetic investigation of *Begonia* L. section *Knesbeckia* (Klotzsch) A.DC. Unpublished PhD thesis. (Glasgow: University of Glasgow): 325 pp.

Baker, W. J. (2000). pers. comm.: Royal Botanic Garden, Kew: w.baker@rbgkew.org.uk.

Baker, W. J., Dransfield, J., Harley, M. M., and Bruneau, A. (1999). Morphology and cladistic analysis of subfamily Calamoideae (Palmae). In *Evolution and classification of palms*, A. Henderson and F. Borchsenius, eds.: Memoirs of the New York Botanical Garden 83: 307-324.

Baker, W. J., Hedderson, T. A., and Dransfield, J. (2000). Molecular phylogenetics of subfamily Calamoideae (Palmae) based on nrDNA ITS and cpDNA *rps*16 intron sequence data. *Molecular Phylogenetics and Evolution* **14**: 195-217.

Baldwin, B. G. (1993). Molecular phylogenetics of *Calycadenis* (Compositae) based on ITS sequences of nuclear ribosomal DNA: chromosomal and morphological evolution reexamined. *American Journal of Botany* **80**: 222-238.

Baldwin, B. G. (1992). Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the Compositae. *Molecular Phylogenetics and Evolution* **1**: 3-16.

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S., and Donoghue, M. J. (1995). The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanic Garden* **82**: 247-277.

Barabe, D. (1980). The perianth of *Begonia* flowers. *The Begonian* **47**: 268-270.

Bateman, R. M. (1999). Integrating molecular and morphological evidence of evolutionary radiations. In *Molecular systematics and plant evolution*, P. M. Hollingsworth, R. M. Bateman and R. J. Gornall, eds. (London: Taylor & Francis): 432-471.

Bateman, R. M. (2000). pers. comm.: Natural History Museum, London: r.bateman@nhm.ac.uk.

Bawa, K. S. (1980). Evolution of dioecy in flowering plants. *Annual Review of Ecology and Systematics* **11**: 15-39.

Bennett, K. D. (1997). *Evolution and Ecology: The Pace of Life*. (Cambridge: Cambridge University Press).

Bentham, G., and Hooker, J. D. (1867). Begoniaceae. In *Genera Plantarum*, G. Bentham and J. D. Hooker, eds. (London: Reeve & Co.): 841-844.

Berg, R. G. v. d. (1983). Pollen characteristics of the genera of the Begoniaceae. *Meded. Landbouwhogeschool Wageningen* **83-9**: 55-66.

Berg, R. G. v. d. (1985). Pollen morphology of the genus *Begonia* in Africa. *Agricultural University Wageningen Papers* **84-3**: 5-94.

Bininda-Edmonds, O. R. P., Bryant, H. N., and Russell, A. P. (1998). Supraspecific taxa as terminals in cladistic analysis: implicit assumptions of monophyly and a comparison of methods. *Biological Journal of the Linnaean Society* **64**: 101-133.

Boeswinkel, F. D. (1984). Ovule and seed structure in Datiscaceae. *Acta Botanica Neerlandica* **33**: 419-429.

Bogler, D. J., and Simpson, B. B. (1996). Phylogeny of Agavaceae based on ITS rDNA sequence variation. *American Journal of Botany* **83**: 1225-1235.

Bona, C., and Alquini, Y. (1995). Morfoanatomia dos tricomas foliares de *Begonia setosa* Kl (Begoniaceae), *Leandra australis* (Cham.) Cogn. (Melastomataceae) e *Solanum fastigiatum* Willd. var. *fastigiatum* (Solanaceae). *Arquivos de Biologia e Tecnologia* **38**: 1295-1302.

Bond, J. E., and Opell, B. D. (1998). Testing adaptive radiation and key innovation hypotheses in spiders. *Evolution* **52**: 403-414.

Bonnefille, R., Roeland, J. C., and Guiot, J. (1990). Temperature and rainfall estimates for the past 40,000 years in equatorial Africa. *Nature* **346**: 347-349.

Bouman, F., and de Lange, A. (1982). Micromorphology of the seed coats in *Begonia* section *Squamibegonia* Warb. *Acta Botanica Neerlandica* **31**: 297-305.

Bouman, F., and de Lange, A. (1983). Structure, micromorphology of *Begonia* seeds. *The Begonian* **50**: 70-78, 91.

Bousquett, J., Strauss, S. H., Doerksen, A. H., and Price, R. A. (1992). Extensive variation in evolutionary rate of *rbcL* gene sequences among seed plants. *Proceedings of National Acadamy of Sciences USA* **89**: 7844-7848.

Bowden, W. M. (1945). A list of chromosome numbers in higher plants. 1. Acanathaceae to Myrtaceae. *American Journal of Botany* **32**: 81-92.

Bremer, K., and Humphries, C. (1993). Generic monograph of the Asteraceae-Anthemideae. *Bulletin of the Natural History Museum* **23**: 71-177.

Brouillet, L. (1995). pers. comm. to Mark Tebbitt: Montreal Botanic Garden

Brummitt, R. K. (1992). *Vascular Plant Families and Genera* (Whitstable, Kent: Whitstable Litho Ltd).

Bult, C. J., Sweere, J. A., and Zimmer, E. A. (1995). Cryptic sequence simplicity, nucleotide composition bias, and molecular coevolution in the large subunit of ribosomal DNA in plants: implications for phylogenetic analyses. *Annals of Misouri Botanical Garden* **82**: 235-246.

Burt-Utley, K. (1985). A revision of the Central-American species of *Begonia* section *Gireoudia* (Begoniaceae). *Tulane studies in Zoology and Botany* **25**: 1-131.

Candolle, A. de (1864). Begoniaceae. In *Prodromus Systematis Naturalis Regni Vegetabilis* (Paris: V. Masson & fil.): 266-408.

Candolle, A. de (1859). Memoire sur la famille des Begoniacees. *Ann. Sci. Nat., Bot. ser. IV* **11**: 93-115.

Carlquist, S. (1985). Wood anatomy of Begoniaceae, with comments on raylessness, paedomorphosis, relationships, vessel diameter, and ecology. *Bulletin of the Torrey Botanical Club* **112**: 59-69.

Chant, S. R. (1978). Datiscaceae. In *Flowering Plants of the World*, V. H. Heywood, ed. (Oxford: Oxford University Press): 114.

Chase, M. W., Soltis, D. E., Olmstead, R. G., Morgan, D., Les, D. H., Mishler, B. D., Duvall, M. R., Price, R. A., Hills, H. G., Qiu, Y.-L., Kron, K. A., Rettig, J. H., Conti, E., Palmer, J. D., Manhart, J. R., Sytsma, K. J., Michaels, H. J., Kress, W. J., Karol, K. G., Clark, W. D., Hedren, M., Gaut, B. S., Jansen, R. K., Kim, K.-J., Wimpee, C. F., Smith, J. F., Furnier, G. R., Strauss, S. H., Xiang, Q.-Y., Plunkett, G. M., Solits, P. S., Swensen, S. M., Williams, S. E., Gadek, P. A., Quinn, C. J., Eguiarte, L. E., Golenberg, E., Learn, G. H., Graham, S. W., Barrett, S. C., Dayanandan, S. and Albert, V. A. (1993). Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid genome *rbcL*. *Annals of the Missouri Botanical Garden* **80**: 528-580.

Chase, M. W., and Albert, V. A. (1998). A perspective on the contribution of plastid *rbcL* DNA sequences to angiosperm phylogenetics. In *Molecular Systematics of Plants II. DNA sequencing*, D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. (Boston: Klewer Academic Publishers): 488-507.

Chase, M. W., and Cox, A. V. (1998). Gene sequences, collaboration, and analysis of large data sets. *Australian Systematic Botany* **11**: 215-229.

Citérne, H., and Cronk, Q. C. B. (1999). The origin of peloric *Sinningia. The New Plantsman* **6**: 219-222.

Clark, C. G., Tague, B. W., Ware, V. C., and Gerbi, S. A. (1984). *Xenopus laevis* 28S ribosomal RNA: a secondary structure model and its evolutionary and functional importance. *Nucleic Acids Research* **12**: 6197-6220.

Clayton, W. D. (1983). The genus concept in practice. *Kew Bulletin* **38**: 149-153.

Clayton, W. D. (1974). The logarithmic distribution of Angiosperm families. *Kew Bulletin* **29**: 271-279.

Coen, E. S., and Meyerowitz, E. M. (1991). The war of the whorls: genetic interactions controlling flower development. *Nature* **353**: 31-37.

Coleman, A. W., Preparata, R. M., Mehrotra, B., and Mai, J. C. (1998). Derivation of the secondary structure of the ITS-1 transcript in Volvocales and its taxonomic correlations. *Protist* **149**: 135-146.

Cronk, Q. C. B. (1989). Measurements of biological and historical influences on plant classifications. *Taxon* **38**: 357-370.

Cronk, Q. C. B. (1990). The name of the pea: a quantitative history of legume classification. *New Phytologist* **116**: 163-175.

Cronk, Q. C. B. (2000). pers. comm.: Royal Botanic Gardens, Edinburgh; Univeristy of Edinburgh: q.cronk@rbge.org.uk.

Cronquist, A. (1981). *An integrated system of classification of flowering plants* (New York: Columbia University Press).

Cuénoud, P., del Pero Martinez, M. A., Loizeau, P. A., Spichiger, R., Andrews, S., and Manen, J. F. (2000). Molecular phylogeny and biogeography of the genus *Ilex* L. (Aquifoliaceae). *Annals of Botany* **85**: 111-122.

Dahlgren, R. (1980). A revised system of classification of the angiosperms. *Journal of the Linnaean Society, Botany* **80**: 91-124.

Diers, L. (1961). Der anteil an polyploiden in den vegetation sgurteln der westordollere Perus. *Zeitschrift fur botanik* **49**: 437-488.

Dixon, M. T., and Hillis, D. M. (1993). Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Molecular Biology and Evolution* **10**: 256-267.

Dodd, M. E., Silvertown, J., and Chase, M. W. (1999). Phylogenetic analysis of trait evolution and species diversity variation among angiosperm

families. *Evolution* **53**: 732-744.

Doorenbos, J. (1999). pers. comm.: Lijsterbeslaan 6; 6721 CW; Bennekom.

Doorenbos, J., Sosef, M. S. M., and de Wilde, J. J. F. E. (1998). The sections of *Begonia. Wageningen Agricultural University Papers* **98-2**: 1-266.

Doyle, J. J., and Davis, J. I. (1998). Homology in molecular phylogenetics: a parsimony perspective. In *Molecular systematics of plants II. DNA sequencing*, D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. (Boston: Klewer Academic Publishers): 101-131.

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf material. *Phytochemical Bulletin* **19**: 11-15.

Doyle, J. J., Doyle, J. L., Ballenger, J. A., Kajita, T., and Ohashi, H. (1997). A phylogeny of the chloroplast gene *rbcL* in the Leguminosae: taxonomic correlations and insights into the evolution of nodulation. *American Journal of Botany* **84**: 541-554.

Doyle, J. J., and Gaut, B. S. (2000). Evolution of genes and taxa: a primer. *Plant Molecular Biology* **42**: 1-23.

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of National Academy of Sciences USA* **93**: 13429-13434.

Eldredge, N. (1995). *Reinventing Darwin. The great evolutionary debate* (New York: John Wiley & Sons, Inc.).

Eldredge, N., and Gould, S. J. (1972). Punctuated Equilibria: an alternative to phyletic gradualism. In *Models in Paleobiology*, T. Schopf, ed. (San Francisco: Freeman, Cooper): 82-115.

Endress, P. K. (1994). *Diversity and evolutionary biology of tropical flowers.* (Cambridge: Cambridge University Press).

Erikkson, O., and Bremer, B. (1992). Pollination systems, dispersal modes, life forms, and diversifiaction rates in angiosperm families. *Evolution* **46**: 258-266.

Eriksson, T. (1998). AutoDecay ver. 4.0 (program distributed by the author) Department of Botany, Stockholm University, Stockholm.

Farris, J. S., Albert, V. A., Kallersjö, M., Lipscomb, D., and Kluge, A. G. (1996). Parsimony Jackknifing outperforms Neighbour-joining. *Cladistics* **12**: 99-124.

Favarger, and Huynh. (1965). *Begonia bracteosa* DC (in I.O.P.B. Chromosome number reports IV, presented by A. Löve & O.T. Solbrig). *Taxon*: 90.

Ferris, C., King, R. A., and Hewitt, G. M. (1999). Isolation within species and the history of glacial refugia. In *Molecular systematics and plant evolution*, P. M. Hollingsworth, R. M. Bateman and R. J. Gornall, eds. (London: Taylor & Francis): 20-34.

Foster, A. S., and Gifford, E. M. (1959). *Comparative Morphology of Vascular Plants* (San Francisco: WH Freeman & Co.).

Gatesy, J., DeSalle, R., and Wheeler, W. (1993). Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Molecular Phylogenetics and Evolution* **2**: 152-157.

Gaut, B. S. (1998). Molecular clocks and nucleotide substitution rates in higher plants. In *Evolutionary Biology*, M. K. Hecht et al., eds. (New York: Plenum Press): 93-120.

Gaut, B. S., Muse, S. V., Clark, W. D., and Clegg, M. T. (1992). Relative rates of nucleotide substitution at the *rbcL* locus of monocotyledonous plants. *Journal of Molecular Evolution* **35**: 292-303.

Gibby, M. (2000). pers. comm.: Royal Botanic Garden, Edinburgh: m.gibby@rbge.org.uk.

Gilbert, D. G. (1992). LoopDloop, a Macintosh program for visualising RNA secondary structure.

Gilbert, D. G. (1995). SeqPup, a Macintosh computer program for sequence alignments.

Goloboff, P. A. (1999). Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* **15**: 415-428.

Goulet, I., Barabe, D., and Brouillet, L. (1994). Analyse structurale et architecture de l'inflorescence des Begoniaceae. *Canadian Journal of Botany* **72**: 897-914.

Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology* **47**: 9-17.

Guo, Q., and Ricklefs, R. E. (2000). Species richness in plant genera disjunct between temperate eastern Asia and North America. *Botanical Journal of the Linnaean Society* **134**: 401-423.

Guyer, C., and Slowinski, J. B. (1993). Adaptive radiation and the topology of large phylogenies. *Evolution* **47**: 253-263.

Hall, R. (1988). The plate tectonics of Cenozoic SE Asia and the distribution of land and sea. In *Biogeography and Geological Evolution of SE Asia*, R. Hall and J. D. Holloway, eds. (Leiden: Backhuys Publishers): 99-131.

Hallam, A. (1994). *An Outline of Phanerozoic Biogeography*, 1st Edition, Volume 10, A. Hallam, B. R. Rosen and T. Whitmore, eds. (Oxford: Oxford University Press).

Hallam, A. (1992). *Phanerozoic sea-level changes* (New York: Columbia University Press).

Hamel, J. L. (1937). Etudes caryologiques sur quelques begoniacees. *Revue Cytol. Biol. veg.* **2**: 392-413.

Hancock, J. M., and Dover, G. A. (1990). 'Compensatory slippage' in the evolution of ribosomal RNA genes. *Nucleic Acids Research* **18**: 5949-5954.

Hancock, J. M., and Dover, G. A. (1988). Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. *Molecular Biology and Evolution* **5**: 377-391.

Hassouna, N., Michot, B., and Bachelleria, J. P. (1984). The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. *Nucleic Acids Research* **12**: 3563-3574.

Heard, S. B. (1996). Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution* **50**: 2141-2148.

Heard, S. B. (1992). Patterns of tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution* **46**: 1818-1826.

Heitz, E. (1927). Ueber multiple und aberrante chromosomenzahlen. *Abh. Naturw. Ver. Hamburg* **21**: 48-58.

Hershkovitz, M. A., and Zimmer, E. A. (1996). Conservation patterns in angiosperm rDNA ITS2 sequences. *Nucleic Acids Research* **24**: 2857-2867.

Hershkovitz, M. A., Zimmer, E. A., and Hahn, W. J. (1999). Ribosomal DNA sequences and angiosperm systematics. In *Molecular systematics and plant evolution*, P. M. Hollingsworth, R. M. Bateman and R. J. Gornall, eds. (London: Taylor & Francis): 268-326.

Heslop-Harrison, J. (1953). *New concepts in flowering-plant taxonomy* (London: Heineman).

Hewitt, G. M. (1996). Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnaean Society* **58**: 247-276.

Hillis, D. M., and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**: 182-192.

Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R., and Molineux, I. J. (1992). Experimental phylogenetics: generation of a known phylogeny. *Science* **255**: 589-592.

Hooker, J. D. (1881). Figure 1. *Gardeners Chronicle* **8**.

Hoover, W. S. (1986). Stomata and stomatal clusters in *Begonia*: ecological response in two Mexican species. *Biotropica* **18**: 16-21.

Hsu, C. C. (1967). Preliminary chromosome studies on the vasular plants of Taiwan (I). *Taiwania* **13**.

Hu, S.-Y. (1967). The evolution and distruibution of the species of Aquifoliaceae in the Pacific area. *Journal of Japanese Botany* **42**: 13-27, 49-59.

Hu, S.-Y. (1949). The genus *Ilex* in China. *Journal of Arnold Arboretum* **30**: 233-344, 348-387.

Hu, S.-Y. (1950). The genus *Ilex* in China. *Journal of Arnold Arboretum* **31**: 39-80, 214-240, 241-263.

Huelsenbeck, J. P., and Kirkpatrick, M. (1996). Do phylogenetic methods produce trees with biased shapes? *Evolution* **50**: 1418-1424.

Hughes, M. (2000). pers. comm.: Royal Botanic Garden, 20A, Inverleith Row, Edinburgh: m.hughes@rbge.org.uk.

Humphries, C. J., and Parenti, L. R. (1999). *Cladistic Biogeography, interpreting patterns of plant and animal distributions*, A. Hallam, B. R. Rosen and T. C. Whitmore, eds. (Oxford: Oxford University Press).

Hutchinson, J. (1959). The Families of Flowering Plants volume 1: Dicotyledons arranged according to a new system based on their probable phylogeny. 2nd Edition. (Oxford: Clarendon Press).

Hutchinson, J. (1964). Order 7, Leguminales. In *The Genera of Flowering Plants* (Oxford: Clarendon Press): 221-489.

Hutchinson, J. H. (1982). Turtle, crocodile, and champsosaur diveristy changes in the Cenozoic of the north-central region of western United States. *Palaeobeography, Paleaoclimatology, Palaeoecology* **37**: 149-164.

Huynh, K. L. (1965). Contribution a l'etude caryologique et embryologique des phanerogames du Peru. *Denkschr. Schweiz. Nat. Ges.* **85**: 1-178.

Irmscher, E. (1925). Begoniaceae. In *Die Naturlichen Pflanzenfamilien*, A. Engler and K. Prantl, eds. (Leipzig: Wilhelm Engelmann): 548-588.

Irmscher, E. (1939). Die begoniaceen chinas. *Mitt. Inst. Allg. Bot. Hamburg* **10**: 427-557.

Irmscher, E. (1959). *Begonia masoniana* Irmscher. *The Begonian* **1959**: 202-203.

Irmscher, E. (1961). Monographische revision der Begoniaceen Afrikas 1. Sekt. *Augustia* und *Rostrobegonia* some einige neue sippen aus anderen sektionen. *Botanische Jarhbucher* **81**: 106-188.

Jaeger, J. A., Turner, D. H., and Zuker, M. (1989). Improved predictions of secondary structure for RNA. *Proceedings of the National Academy of Sciences USA, BIOCHEMISTRY* **86**: 7706-7710.

Jaeger, L. A., Turner, H., and Zuker, M. (1989). Predicting optimal and suboptimal secondary structure for RNA. In *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences; Methods in Enzymology 183*, R. F. Doolittle, ed.: 281-306.

Jeandroz, S., Roy, A., and Bousquet, J. (1997). Phylogeny and phylogeography of the circumpolar genus *Fraxinus* (Oleaceae) based on internal transcribed spacer sequences of nuclear ribosomal DNA. *Molecular Phylogenetics and Evolution* **7**: 241-251.

Jin, X., and Wang, F.-H. (1994). Style and ovary anatomy of Chinese *Begonia* and its taxonomic and evolutionary implications. *Cathaya* **6**: 125-144.

Jobst, J., King, K., and Hemleben, V. (1998). Molecular evolution of the Internal Transcribed Spacers (ITS 1 and ITS 2) and phylogenetic relationships among species of the family Cucurbitaceae. *Molecular Phylogenetics and Evolution* **9**: 204-219.

Johnson, L. A., Soltis, D. E., and Soltis, P. S. (1999). Phylogenetic relationships of Polemoniaceae inferred from 18S ribosomal DNA sequences. *Plant Systematics and Evolution* **214**: 65-89.

Jong, K. (1997). *Laboratory manual of plant cytological techniques*: (Royal Botanic Garden Edinburgh).

Jorgensen, T. H., and Frydenberg, J. (1999). Diversification in insular plants: inferring the phylogenetic relationship in *Aeonium* (Crassulaceae) using ITS sequences of nuclear ribosomal DNA. *Nordic Journal of Botany* **19**: 613-621.

Kallersjö, M., Albert, V. A., and Farris, J. S. (1999). Homoplasy *increases* phylogenetic structure. *Cladistics* **15**: 91-93.

Kapoor. (1966). *Hillebrandia sandwichensis* Oliver (in I.O.P.B. chromosome number reports VIII, presented by A. Love). *Taxon*: 284.

Karegeannes, C. (1981). *B. listada*: where the misspelling came from. *The Begonian* **49**: 157, 159.

Keraudren-Aymonin, M. (1983). Begonicees. In *Flore du Madagascar*, H. Humbert, ed. (Paris: Museum National d'Historie Naturelle): 7-108.

Keraudren-Aymonin, M. (1983). Caracteres et interet taxinomique de l'ornementation tegumentaire des graines de quelques especes de *Begonia* de Madagascar observees en M.E.B. *Bull. Soc. Bot. Fr.* **130**: 329-338.

Kim, H.-G., Keeley, S. C., Vroom, P. S., and Jansen, R. K. (1998). Molecular evidence for an African origin of the Hawaiian endemic *Hesperomannia* (Asteraceae). *Proceedings of the National Academy of Sciences USA* **95**: 15440-15445.

Kim, J. (1996). General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Systematic Biology* **45**: 363-374.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**: 624-626.

Kirkpatrick, M., and Slatkin. (1993). Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* **44**: 1671-1681.

Kitching, I. J., Forley, P. L., Humphries, C. J., and Williams, D. M. (1998). *Cladistics, the theory and practice of parsimony analysis*, 2nd Edition (Oxford: Oxford University Press).

Klazenga, N., de Wilde, J. J. F. E., and Quene, R. J. (1994). *Begonia* sect. *Mezieria* (Gaud.) Warb., a taxonomic revision. *Bull. Jard. Bot. Nat. Belg.* **63**: 263-312.

Kondo, K. (1973). Chromosome numbers of five taxa. *Chromosome Information Service* **15**: 33-34.

Kopperud, C., and Einset, J. W. (1995). DNA isolation from *Begonia* leaves. *Plant Molecular Biology Reporter* **13**: 129-130.

Kuzoff, R. K., Sweere, J. A., Soltis, D. E., Soltis, P. S., and Zimmer, E. A. (1998). The phylogenetic potential of entire 26S rDNA sequences in plants. *Molecular Biology and Evolution* **15**: 251-263.

Lange, A. d. (1988). Begonia seeds and their adaptations to dispersal. *Acta Botanica Neerlandica* **37**: 322.

Lange, A. d., and Bouman, F. (1986). Micromorphology of the seeds in *Begonia* section *Solananthera* A.DC. *Acta Botanica Neerlandica* **35**: 489-495.

Lange, A. d., and Bouman, F. (1999). *Seed micromorphology of Neotropical begonias*, Volume 90 (Washington DC: Smithsonian Institute Press).

Lange, A. d., and Bouman, F. (1992). Seed micromorphology of the genus *Begonia* in Africa: taxonomic and ecological implications. *Wageningen Agricultural Univeristy Papers* **91-4**: 1-82.

Lapointe, F.-J., and Cucumel, G. (1997). The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology* **46**: 306-312.

Lawrence, G. H. M. (1951). *Taxonomy of vascular plants* (New York: The McMillan Company).

Lee, Y. S. (1974). A study of stem anatomy in *Begonia* L. *Phytologia* **27**.

Legro, R. A. H., and Doorenbos, J. (1969). Chromosome numbers in *Begonia. Netherland Journal of Agricultural Science* **17**: 189-202.

Legro, R. A. H., and Doorenbos, J. (1971). Chromosome numbers in *Begonia* 2. *Netherland Journal of Agricultural Science* **19**: 176-183.

Legro, R. A. H., and Doorenbos, J. (1973). Chromosome numbers in *Begonia* 3. *Netherland Journal of Agricultural Science* **21**: 167-170.

Li, J., Bogle, A. L., and Klein, A. S. (1999). Phylogenetic relationships of the Hammelidaceae inferred from sequences of Internal Transcribed Spacers (ITS) of nuclear ribosomal DNA. *American Journal of Botany* **86**: 1027-1037.

Li, W. H. (1997). *Molecular Evolution* (Sunderland, Massachusetts: Sinaeur).

Lincoln, R. J., Boxshall, G. A., and Clark, P. F. (1982). *A dictionary of ecology, evolution and systematics.* (Cambridge: Cambridge Univeristy Press).

Lindley, J. (1846). *The Vegetable Kingdom, etc.* (London: Bradbury & Evans).

Linnaeus, C. (1753). *Species Plantarum* (Stockholm).

Liston, A., Rieseberg, L. H., and Elias, T. S. (*1990*). Functional androdioecy in the flowering plant *Datisca glomerata. Nature* **343**: 641-642.

Liston, A., Rieseberg, L. H., and Elias, T. S. (1989). Morphological stasis and molecular divergence in the intercontinental disjunct genus *Datisca* (Datiscaceae). *Aliso* **12**: 525-542.

Liston, A., Rieseberg, L. H., and Hanson, M. A. (1992). Geographic partitioning of chloroplast DNA variation in the genus *Datisca* (Datiscaceae). *Plant Systematics and Evolution* **181**: 121-132.

Liston, A., Robinson, W. A., Pinero, D., and Alvarez-Buylla, E. R. (1999). Phylogenetics of Pinus (Pinaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. *Molecular Phylogenetics and Systematics* **11**: 95.

Loesener, T. (1942). Aquifoliaceae. In *Die Naturlichen Pflantzenfamilien*, A. Engler and K. Prantl, eds. (Eigelmann: Leipzig): 36-86.

Mabberley, D. J. (1997). *The plant-book: a portable dictionary of the vascular plants*, 2nd Edition (Cambridge: Cambridge University Press).

Maddison, W. P., and Maddison, D. R. (1992). MacClade (Sunderland, Massachusetts: Sinauer Associates, Inc., Publishers).

Maddison, W. P., and Maddison, D. R. (1997). MacClade 3.07 (Sunderland, Massachusetts: Sinauer Associates, Inc., Publishers).

Mai, J. C., and Coleman, A. W. (1997). The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *Journal of Molecular Evolution* **44**: 258-271.

Manos, P. S. (1997). Systematics of *Nothofagus* (Nothofagaceae) based on rDNA spacer sequences (ITS): taxonomic congruence with morphology and plastid sequences. *American Journal of Botany* **84**: 1137-1155.

Mason, L. M. (1957). Begonias at Fincham. *The Begonian* **24**: 3-4.

Matolweni, L. O., Balkwill, K., and McLellan, T. (2000). Genetic diversity and gene flow in the morphologically variable, rare endemics *Begonia dregei* and *Begonia homonyma* (Begoniaceae). *American Journal of Botany* **87**: 431-439.

Matsuura, H., and Okuno, S. (1936). Cytogenetical studies on *Begonia.* I. The chromosome numbers (a preliminary note). *Japanese Journal of Genetics* **12**: 42-43.

McCall, R. A. (1997). Implications of recent geological investigations of the Mozambique Channel for the mammalian colonization of Madagascar. *Proceedings of the Royal Society of London B* **264**: 663-665.

McDade, L. A., Masta, S. E., Moody, M. L., and Waters, E. (2000). Phylogenetic relationships among Acanthaceae: evidence from two genomes. *Systematic Botany* **25**: 106-121.

McGregor. (1969). Chromosomes in *Begonia. The Begonian* **36**: 230-232.

McLellan, T. (1990). Development of differences in leaf shape in *Begonia dregei* (Begoniaceae). *American Journal of Botany* **77**: 323-337.

McLellan, T. (2000). pers. comm.: Molecular and Cell Biology, University of the Witwatersrand, Private Bag 3, WITS 2050, South Africa: 108trm@cosmos.wits.ac.za.

Mereminski, M. H. (1936). Zrozwoju woreczka zalqzkowego ukosnicy *Begonia incana* Lindl. - Uber embryosackentwicklung bei *Begonia incana* Lindl. (ein beitrag zur embryologie der gattung *Begonia*). *Extr. Bull. Acad. pol. Sci.* **III B I**: 53-92.

Miller, R. E., Rausher, M. D., and Manos, P. S. (1999). Phylogenetic systematics of *Ipomoea* (Convolvulaceae) based on ITS and *Waxy*

sequences. *Systematic Botany* **24**: 209-227.

Minelli, A. (1993). *Biological Systematics - The State of the Art* (London: Chapman & Hall).

Mishler, B. D. (1994). Cladistic analysis of molecular and morphological data. *American Journal of Physical Anthropology* **94**: 143-156.

Mishler, B. D., Soltis, P. S., and Soltis, D. E. (1998). *Compartmentalization in phylogeny reconstruction: philosophy and practice.* (Princeton, New Jersey: DIMACS).

Möller, M., and Cronk, Q. C. B. (2001). Phylogenetic studies in *Streptocarpus*: reconstruction of biogeographic history and distribution patterns in *Streptocarpus* (Gesneriaceae). in prep.

Möller, M., and Cronk, Q. C. B. (1997). Origin and relationships of *Saintpaulia* (Gesnerianceae) based on ribosomal DNA internal transcribed spacer (ITS) sequences. *American Journal of Botany* **84**: 956-965.

Molnar, P., and Tapponnier, P. (1975). Cenozoic tectonics of Asia: effects of a continental collision. *Science* **189**: 419-426.

Mooers, A. O., and Heard, S. B. (1997). Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology* **72**: 31-54.

Morley, R. J. (1998). Palynological evidence for Tertiary plant dispersals in the SE Asian region in relation to plate tectonics and climate. In *Biogeography and Geological Evolution of SE Asia*, R. Hall and J. D. Holloway, eds. (Leiden: Backhuys Publishers): 211-234.

Morrison, D. A., and Ellis, J. T. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. *Molecular Biology and Evolution* **14**: 428-441.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., de Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* **403**: 853-858.

Nee, S., and Harvey, P. H. (1994). Getting to the roots of flowering plant diversity. *Science* **264**: 1549-1550.

Niklas, K. J. (1997). *The Evolutionary Biology of Plants*, 1st Edition (Chicago: The University of Chicago Press).

Nixon, K. C. (1999). The Parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics* **15**: 407-414.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* **246**: 96-97.

Oxelman, B., and Liden, M. (1995). The position of *Circaeaster* - evidence from nuclear ribosomal DNA. *Plant Systematics and Evolution [Suppl.]* **9**: 189-193.

Page, R. D. M., and Holmes, E. C. (1998). *Molecular Evolution: a Phylogenetic Approach*, 1st Edition (Oxford: Blackwell Science).

Panda, S., and de Wilde, J. J. F. E. (1995). Diversity and taxonomic value of stigmatic surfaces in Begoniaceae: SEM analysis. *Acta Botanica Neerlandica* **44**: 139-150.

Parrish, J. T. (1993). The palaeogeography of the opening South Atlantic. In *The Africa-South America connection*, W. George and R. Lavocat, eds. (Oxford: Clarendon Press): 8-27.

Pearson, P. N. (1999). Apomorphy distribution is an important aspect of cladogram symmetry. *Systematic Biology* **48**: 399-406.

Peng, C.-I., and Chen, Y.-K. (1991). Hybridity and parentage of *Begonia buimontana* Yamamoto (Begoniaceae) from Taiwan. *Annals of the Missouri Botanical Garden* **78**: 995-1001.

Pennington, R. T. (1995). Cladistic analysis of chloroplast DNA restriction site characters in *Andira* (Leguminosae: Dalbergieae). *American Journal of Botany* **82**: 526-534.

Pfosser, M., and Speta, F. (1999). Phylogenetics of Hyacinthaceae based on plastid DNA sequences. *Annals of Missouri Botanical Garden* **86**: 852-875.

Powell, C. M., and Conaghan, P. J. (1973). Plate tectonics and the Himalayas. *Earth and planetary science letters* **20**: 1-12.

Prather, L. A., and Jansen, R. K. (1998). Phylogeny of *Cobaea* (Polemoniaceae) based on sequence data from the ITS region of nuclear ribosomal DNA. *Systematic Botany* **23**: 57-72.

Raikow, R. J. (1986). Why are there so many kinds of passarine birds? *Systematic Zoology* **35**: 255-259.

Reitsma, J. M. (1984). Placentation in Begonias from the African continent. *Meded. Landbouwhogeschool Wageningen* **83-9**: 21-53.

Rice, K. A., Donoghue, M. L., and Olmstead, R. G. (1997). Analysing large data sets: *rbcL* 500 revisited. *Systematic Biology* **46**: 554-563.

Richardson, I. B. K. (1993). Begoniaceae. In *Flowering Plants of the World*, V. H. Heywood, ed. (London: B.T. Batsford Ltd.): 113-114.

Richardson, J. E. (1999). Molecular systematics of the genus *Phylica* L. with an emphasis on the island species. Unpublished PhD thesis (Edinburgh: University of Edinburgh): 253 pp.

Richardson, J. E., Fay, M. F., Cronk, Q. C. B., Bowman, D., and Chase, M. W. (2000). A phylogenetic analysis of Rhamnaceae using *rbcL* and *trnL-F* plastid DNA sequences. *American Journal of Botany* **87**: 1309-1324.

Rieseberg, L. H., Hanson, M. A., and Philbrick, C. T. (1992). Androdioecy is derived from dioecy in Datiscaceae: evidence from restriction site mapping of PCR-amplified chloroplast DNA fragments. *Systematic Botany* **17**: 324-336.

Rieseberg, L. H., and Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants* **5**: 65-84.

Ro, K. E., Keener, C. S., and McPheron, B. A. (1997). Molecular phylogenetic study of the Ranunculaceae: utility of the nuclear 26S ribosomal DNA in inferring intrafamilial relationships. *Molecular Phylogenetics and Evolution* **8**: 117-127.

Salisbury, B. A. (1999). Misinformative characters and phylogeny shape. *Systematic Biology* **48**: 153-169.

Sanderson, M. J. (1990). Estimating rates of speciation and evolution: a bias due to homoplasy. *Cladistics* **6**: 387-391.

Sanderson, M. J., Purvis, A., and Henze, C. (1998). Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* **13**: 105-109.

Sanderson, M. J., and Wojciechowski, M. F. (1996). Diversification rates in a temperate legume clade: are there "so many species" of *Astragalus* (Fabaceae)? *American Journal of Botany* **83**: 1488-1502.

Sands, M. (2000). pers. comm.: Royal Botanic Gardens, Kew: m.sands@rbgkew.org.uk.

Sang, T., Crawford, D. J., Kim, S.-C., and Stuessy, T. (1994). Radiation of the endemic genus *Dendroseris* (Asteraceae) on the Juan Fernandos

islands: evidence from sequences of the ITS regions of nuclear ribosomal DNA. *American Journal of Botany* **81**: 1494-1501.

Sang, T., Crawford, D. J., and Stuessy, T. (1995). ITS sequences and the phylogeny of the genus *Robinsonia* (Asteraceae). *Systematic Botany* **20**: 55-64.

Sang, T., and Zhang, D. (1999). Reconstructing hybrid speciation using sequences of low copy nuclear genes: hybrid origins of five *Paeonia* species based on *Adh* gene phylogenies. *Systematic Botany* **24**: 148-163.

Sarkar, A. K. (1974). Evolution of species in the genus *Begonia. Proceedings Indian Sci. Congr. Assoc.* **61 (III B)**: 32-33.

Sarkar, A. K. (1970). Cytotaxonomy of Angiosperms II. Dicotyledons: Cucurbitales; Chromosome number reports of plants. In *The Research Bulletin volume 2, 1967-68, Cytogenetics Lab., Dept. of Botany, Univeristy of Calcutta*, A. Sharma and A. K. Sarkar, eds.: 28, 39.

Sarkar, A. K. (1989). Taxonomy of *Begonia* L. (Begoniaceae) as judged through cytology. *Feddes Repertorium* **100**: 241-250.

Savolainen, V. (2000). pers. comm. (lecture at RBGE): Royal Botanic Gardens, Kew.

Savolainen, V., Chase, M. W., Hoot, S. B., Morton, C. M., Soltis, D. E., Bayer, C., Fay, M. F., de Bruijn, A. Y., Sullivan, S., and Qiu, Y.-L. (2000a). Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. *Systematic Biology* **49**: 306-362.

Savolainen, V., Fay, M. F., Albach, D. C., Backlund, A., van der Bank, M., Cameron, K. M., Johnson, S. A., Llédo, M. D., Pintaud, J.-C., Powell, M., Sheahan, M. C., Soltis, D. E., Soltis, P. S., Weston, P., Whitten, W. M., Wurdack, K. J., and Chase, M. W. (2000b). Phylogeny of the eudicots: a nearly complete familial analysis based on *rbcL* gene sequences. *Kew*

Bulletin **55**: 257-309.

Schemske, D. W., Agren, J., and le Corff, J. (1996). Deceit pollination in the monoecious, neotropical herb *Begonia oaxacana* (Begoniaceae). In *Floral biology*, D. G. Lloyd and S. C. H. Barrett, eds.: Chapman and Hall): 292-318.

Schmidt, G. J., and Schilling, E. E. (2000). Phylogeny and Biogeography of *Eupatorium* (Asteraceae: Eupatorieae) based on nuclear ITS sequence. *American Journal of Botany* **87**: 716-726.

Scotese, C. R., Gahagen, L. M., and Larsen, R. L. (1988). Plate tectonic reconstructins of the Cretaceous and Cenozoic ocean basins. *Tectonophysics* **155**: 27-48.

Seelanan, T., Schnabel, A., and Wendel, J. F. (1997). Congruence and consensus in the cotton tribe (Malvaceae). *Systematic Botany* **22**: 259-290.

Seitner, P. G. (1972). Some observations on *Begonia* seeds. *The Begonian* **39**: 47-55.

Sessions, S. K. (1996). 5. Chromosomes: molecular cytogenetics. In *Molecular Systematics*, D. M. Hillis, C. Moritz and B. Mable, eds. (Sunderland, Massachusetts, USA: Sinauer Associates, Inc.): 121-168.

Setoguchi, H., Osawa, T. A., Pintaud, J. C., Jaffre, T., and Veillon, J. M. (1998). Phylogenetic relationships within Araucariaceae based on *rbcL* gene sequences. *American Journal of Botany* **85**: 1507-1516.

Sharma, A. K., and Bhattacharyya, U. C. (1961). Cytological studies in *Begonia* - II. *Caryologia* **14**: 279-301.

Sharma, A. K., and Bhattacharyya, U. C. (1957). Cytological studies in *Begonia* - I. *La Cellule* **58**: 307-329.

Sharp, P. (2000). How old are Begonias? *Begonia Australis* **10**: 30-31.

Sharp, P. (1996). How old? *The Begonian* **63**: 97-99.

Shaw, H. K. A. (1966). *Dictionary of the Flowering Plants and Ferns*, 7th Edition.

Shui, Y.-M., Li, Q.-R., and Huang, S.-H. (1999). Observation of leaf epidermis and its hair of *Begonia* from Yunnan. *Acta Botanica Yunnanica* **21**: 309-316.

Siddall, M. E., and Kluge, A. G. (1997). Probabilism and phylogenetic inference. *Cladistics* **13**: 313-336.

Siddall, M. E., and Whiting, M. F. (1999). Long-branch abstractions. *Cladistics* **15**: 9-24.

Siebert, D. J. (1992). Tree statistics; trees and 'confidence'; consensus trees; alternatives to parsimony; character weighting; character conflict and its resolution. In *Cladistics: a practical course in systematics*, P. L. Forey, C. J. Humphries, I. J. Kitching, R. W. Scotland, D. J. Siebert and D. M. Williams, eds. (Oxford: Oxford University Press): 72-88.

Simmons, M. P., and Ochoterena, H. (2000). Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* **49**: 369-381.

Sluiman, H. (1998). pers. comm.: Royal Botanic Garden, Edinburgh: h.sluiman@rbge.org.uk.

Smith, L. B., and Wasshausen, D. C. (1981). The *Begonia* on the cover: *B. listada*. *The Begonian* **49**: 155-156.

Smith, L. B., Wasshausen, D. C., Golding, J., and Karegeannes, C. E. (1986). Begoniaceae. Part I: Illustrated key. Part II: Annotated species list. *Smithsonian Contributions to Botany* **60**: 1-584.

Snow, R. (1959). Chromosome numbers of Californian plants, with notes on some cases of cytological interest. *Madroso* **15**: 81-89.

Soltis, D. E., Hibsch-Jetter, C., Solis, P. S., Chase, M. W., and Farris, J. S. (1997). Molecular phylogenetic relationships among angiosperms: an overview based on *rbcL* and 18S rDNA sequences. In *Evolution and diversification of land plants*, K. Iwatsuki and P. H. Raven, eds. (Tokyo: Springer-Verlag): 157-178.

Soltis, D. E., Johnson, L. A., and Looney, C. (1996). Discordance between ITS and chloroplast topologies in the *Boykinia* group. *Systematic Botany* **21**: 169-187.

Soltis, D. E., and Soltis, P. S. (2000). Phylogenetic analysis of large data sets. In *The Biology of Biodiversity*, M. Kato, ed. (Tokyo: Springer-Verlag): 91-103.

Soltis, D. E., and Soltis, P. S. (1998). Chosing an approach and an appropriate gene for phylogenetic analysis. In *Molecular systematics of plants II. DNA sequencing.*, D. E. Soltis, P. S. Soltis and J. J. Doyle, eds. (Boston: Klewer Academic Publishers): 1-42.

Soltis, D. E., Soltis, P. S., Nickrent, D. L., Johnson, L. A., Hahn, W. J., Hoot, S. B., Sweere, J. A., Kuzoff, R. K., Kron, K. A., Chase, M. W., Swensen, S. M., Zimmer, E. A., Chaw, S.-M., Gillespie, L. J., Kress, W. J., and Sytsma, K. J. (1997). Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *Annals of the Missouri Botanical Garden* **84**: 1-49.

Soltis, P. S., and Soltis, D. E. (1995). 4. Plant Molecular Systematics; inferences of phylogeny and evolutionary processes. In Evolutionary Biology, M. K. Hecht, R. J. MacIntyre and M. T. Clegg, eds. (New York: Plenum Press): 139-194.

Soltis, P. S., Soltis, D. E., and Chase, M. W. (1999). Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402**:

402-404.

Sosef, M. S. M. (1994). Refuge begonias: taxonomy, phylogeny and historical biogeography of *Begonia* sect. *Loasibegonia* and sect. *Scutobegonia* in relation to glacial rain forest refuges in Africa (Studies in Begoniaceae 5). *Wageningen Agricultural University papers*: 1-306.

Stoebe, B., Hasman, S., Goremykin, V., Kowallik, K. V., and Martin, W. (1999). Proteins encoded in sequenced chloroplast genomes: an overview of gene content, phylogenetic information and endosymbiotic gene transfer to the nucleus. In *Molecular systematics and plant evolution*, P. M. Hollingsworth, R. M. Bateman and R. J. Gornall, eds. (London: Taylor & Francis): 327-352.

Sun, Y., Skinner, D. Z., Lang, G. H., and Hulbert, S. H. (1994). Phylogenetic analysis of sorghum and related taxa using internal transcribed spacers of nuclear ribosomal DNA. *Theoretical and Applied Genetics* **89**: 26-32.

Swensen, S. (1996). The evolution of actinorhizal symbioses: evidence for multiple origins of the symbiotic association. *American Journal of Botany* **83**: 1503-1512.

Swensen, S. M., Luthi, J. N., and Rieseberg, L. H. (1998). Datiscaceae revisited: monophyly and the sequence of breeding system evolution. *Systematic Botany* **23**: 157-169.

Swensen, S. M., Mullin, B. C., and Chase, M. W. (1994). Phylogenetic affinities of Datiscaceae based on an analysis of nucleotide sequences from the plastid *rbcL* gene. *Systematic Botany* **19**: 157-168.

Swofford, D. L. (1993). *PAUP: phylogenetic analysis using parsimony (version 3.1) users manual* (Illinois: Illinois Natural History Survey).

Swofford, D. L. (1998). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) (Sunderland, Massachusetts: Sinauer Associates).

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. In *Molecular Systematics*, D. M. Hillis, C. Moritz and B. K. Mable, eds. (Massachusetts: Sinauer): 407-514.

Takhtajan, A. (1980). Outline of the classification of flowering plants (Magnoliophyta). *Botanical Review (Lancaster)* **46**: 225-359.

Tautz, D., Hancock, J. M., Webb, D. A., Tautz, C., and Dover, G. A. (1988). Complete sequences of the rRNA genes of *Drosphila melanogaster. Molecular Biology and Evolution* **5**: 366-376.

Tebbitt, M. C. (1997). A systematic investigation of *Begonia* section *Sphenanthera* (Hassk.) Benth. & Hook.f. Unpublished PhD thesis (Glasgow: University of Glasgow): 231 pp.

Tebbitt, M. C., and MacIver, C. M. (1999). The systematic significance of the endothecium in Begoniaceae. *Botanical Journal of the Linnaean Society* **131**: 203-221.

Teo, L.-L., and Kiew, R. (1999). First record of a natural *Begonia* hybrid in Malaysia. *Gardens' Bulletin Singapore* **51**: 103-118.

Thiede, J. (1977). Subsidence of aseimic ridges: evidence from sediments on Rio Grande Rise (south-west Atlantic Ocean). *Bulletin of the American Association of Petroleum Geologists* **61**: 939-940.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1997). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673-4680.

Thorne, R. T. (1992). Classification and geography of the flowering plants. *Botanical Review (Lancaster)* **58**: 225-348.

Vbra, E. S. (1985). Environment and evolution: alternative causes of the temporal distribution of evolutionary events. *South African Journal of Science* **81**: 229-236.

Veevers, J. J., Powell, C. M., and Johnson, B. D. (1980). Seafloor constraints on the reconstruction of Gondwanaland. *Earth and Planetary Science Letters* **51**: 435-444.

Vogel, S. (1998). Remarkable nectaries: structure, ecology, organophyletic perspectives IV. Miscellaneous cases. *Flora* **193**: 225-248.

Wagstaff, S. J., and Dawson, M. I. (2000). Classification, origin, and patterns of diversification of *Corynocarpus* (Corynocarpaceae) inferred from DNA sequences. *Systematic Botany* **25**: 134-149.

Walters, S. M. (1986). The name of the rose: a review of ideas on the European bias in angiosperm classification. *New Phytologist* **104**: 527-546.

Warburg, O. (1894). Begoniaceae. In *Die Naturlichen Pflanzenfamilien*, A. Engler and K. Prantl, eds. (Leipzig: Wilhelm Engelmann): 121-150.

Watson, L. E., Evans, T. M., and Boluarte, T. (2000). Molecular phylogeny and biogeography of tribe Antheimidae (Asteraceae), based on chloroplast gene *ndhF*. *Molecular Phylogenetics and Evolution* **15**: 59-69.

Wenzel, J. W., and Siddall, M. E. (1999). Noise. *Cladistics* **15**: 51-64.

Wheeler, W. C. (1995). Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Systematic Biology* **44**: 321-331.

Wheeler, W. C., Gatesy, J., and DeSalle, R. (1995). Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Molecular Phylogenetics and Evolution* **4**: 1-9.

Wheeler, W. C., and Gladstein, D. S. (1992). MALIGN (New York: American Museum of Natural History).

Wheeler, W. C., and Honeycut, R. L. (1988). Paired sequence difference in ribosomal RNAs: evolution and phylogenetic implications. *Molecular Biology and Evolution* **5**: 90-96.

White, O. E., Taylor, J. H., and Speese, B. M. (1946). *Begonia* species hybrids. *Journal of Heredity* **37**: 66-70.

White, T. J., Bruns, T., Lee, S., and Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In *PCR Protocols: a guide to methods and applications*: (Academic Press): 315-322.

Wilde, J. J. F. E. de (1985). *Begonia* section *Cristasemen* J.J. de Wilde, sect. nov. *Agricultural Univeristy Wageningen Papers* **84-3**: 113-129

Wilde, J. J. F. E. de, and Arends, J. C. (1979). *Begonia loranthoides* Hook. f. (sect. *Tetraphila* A. DC.). *Acta Botanica Neerlandica* **28**: 357-374

Wilde, J. J. F. E. de, and Arends, J. C. (1989). *Begonia salaziensis* (Gaud.) Warb., taxonomy and placentation. *Acta Botanica Neerlandica* **38**: 31-39

Wilde, J. J. F. E. de, and Arends, J. C. (1980). *Begonia* section *Squamibegonia* Warb. a taxonomic revision. *Misc. Papers (Landbouwhogeschool Wageningen)* **19**: 377-421

Wilkinson, M., and Thorley, J. L. (1998). Reduced supertrees. *Trends in Ecology and Evolution* **13**: 283

Willis, J. C. (1922). *Age and Area* (Cambridge: Cambridge University Press).

Windley, B. F. (1995). *The evolving continents*, 3rd Edition (Chichester: John Wiley & Sons).

Wojciechowski, M. F., Sanderson, M. J., and Hu, J. M. (1999). Evidence on the monophyly of *Astragalus* (Fabaceae) and its major subgroups based on nuclear ribosomal DNA ITS and chloroplast DNA *trnL* intron data. *Systematic Botany* **24**: 409-437

Wu, C.-I., and Li, W.-H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of National Academy of Sciences USA* **82**: 1741-1745

Yang, Z. (1996). Among-site variation and its impact on phylogenetic analysis. *Trends in Ecology and Evolution* **11**: 367-371

Yokoyama, J., Suzuki, M., Iwatsuki, K., and Hasabe, M. (2000). Molecular phylogeny of *Coriaria*, with special emphasis on the disjunct distribution. *Molecular Phylogenetics and Evolution* **14**: 11-19

Yuan, Y. M., and Kupfer, P. (1997). The monophyly and rapid evolution of *Gentiana* sect. *Chondrophyllae* Bunge *s.l.* (Gentianaceae): evidence from the nucleotide sequences of the internal transcribed spacers of nuclear ribosomal DNA. *Botanical Journal of the Linnaean Society* **123**: 25-43

Zeilinga, A. E. (1962). Cytological investigation of hybrid varieties of *Begonia semperflorens*. *Euphytica* **11**: 126-136

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science* **244**: 48-52

14 Appendices

14.1A: Large Vascular Plant Genera, from Minelli 1993, arranged according to Species Number

FAMILY	GENUS	SPP No.	FAMILY	GENUS	SPP No.
Cyperaceae	*Carex*	2000	Poaceae	*Panicum*	500
Euphorbiaceae	*Euphorbia*	2000	Polygonaceae	*Polygala*	500
Piperaceae	*Piper*	2000	Rosaceae	*Potentilla*	500
Fabaceae	*Astragalus*	1750	Eriocaulaceae	*Paepalanthus*	485
Solanaceaea	*Solanum*	1700	Fabaceae	*Mimosa*	480
Begoniaceae	*Begonia*	1400	Ebenaceae	*Diospyros*	475
Asteraceae	*Senecio*	1250	Ericaceae	*Vaccinium*	450
Fabaceae	*Acacia*	1200	Fabaceae	*Desmodium*	450
Orchidaceae	*Pleurothallis*	1120	Euphorbiaceae	*Acalypha*	430
Melastomataceae	*Miconia*	1000	Passifloraceae	*Passiflora*	430
Myrtaceae	*Syzygium*	1000	Primulaceae	*Primula*	425
Orchidaceae	*Bulbophyllum*	1000	Aquifoliaceae	*Ilex*	400
Piperaceae	*Peperomia*	1000	Dryopteridaceae	*Diplazium*	400
Rubiaceae	*Psychotria*	800-1500	Eriocaulaceae	*Eriocaulon*	400
Lamiaceae	*Salvia*	900	Fabaceae	*Tephrosia*	400
Orchidaceae	*Dendrobium*	900	Fagaceae	*Quercus*	400
Balsaminaceae	*Impatiens*	850	Lamiaceae	*Clerodendrum*	400
Dioscoriaceae	*Dioscorea*	850	Lauraceae	*Litsea*	400
Ericaceae	*Rhododendron*	850	Lomariopsidaceae	*Elaphoglossum*	400
Orchidaceae	*Epidendrum*	800	Melastomataceae	*Medinilla*	400
Euphorbiaceae	*Croton*	750	Rubiaceae	*Pavetta*	400
Moraceae	*Ficus*	750	Salicaeae	*Salix*	400
Ericaceae	*Erica*	735	Violaceae	*Viola*	400
Aspleniaceae	*Asplenium*	720	Bromeliaceae	*Tillandsia*	380
Araceae	*Anthurium*	700	Clusiaceae	*Hypericum*	370
Caryophylaceae	*Silene*	700	Gentianaceae	*Gentiana*	361
Fabaceae	*Indigofera*	700	Asteraceae	*Artemesia*	350
Oxalidaceae	*Oxalis*	700	Scrophulariacae	*Pedunculoris*	350
Pandanaceae	*Pandanus*	700	Asteraceae	*Saussurea*	300
Selaginellaceae	*Selaginella*	700	Cyperaceae	*Cyperus*	300
Alliaceae	*Allium*	690	Geraniaceae	*Geranium*	300
Orchidaceae	*Oncidium*	680	Lamiaceae	*Hyptis*	300
Convolvulaceae	*Ipomoea*	650	Lycopodiaceae	*Huperzia*	300
Cyatheaceae	*Cyathea*	620	Rubiaceae	*Galium*	300
Acanthaceae	*Justicia*	600	Rubiaceae	*Ixora*	300
Asteraceae	*Helichrysum*	600	Asteraceae	*Aster*	250
Euphorbiaceae	*Phyllanthis*	600	Myrsinaceae	*Ardisia*	250
Fabaceae	*Crotalaria*	600	Myrtaceae	*Myricia*	250
Myrtaceae	*Eucalyptus*	600	Guttiferae	*Garcinia*	200
Orchidaceae	*Habenaria*	600	Oleaceae	*Jasminum*	200
Ranunculaceae	*Ranunculus*	600	Urticaceae	*Pilea*	200
Myrtaceae	*Eugenia*	550			
Asteraceae	*Veronina*	500			
Asteraceae	*Cousinia*	500			
Asteraceae	*Centaurea*	500			
Berberidaceae	*Berberis*	500			

Authorities as given in table 14.3

14.1B: Large Vascular Plant Genera, from Minelll, 1993, arranged according to Family

FAMILY	GENUS	SPP No.	FAMILY	GENUS	SPP No.
Acanthaceae	*Justicia*	600	Lamiaceae	*Clerodendrum*	400
Alliaceae	*Allium*	690	Lamiaceae	*Hyptis*	300
Aquifoliaceae	*Ilex*	400	Lauraceae	*Litsea*	400
Araceae	*Anthurium*	700	Lomariopsidaceae	*Elaphoglossum*	400
Aspleniaceae	*Asplenium*	720	Lycopodiaceae	*Huperzia*	300
Asteraceae	*Senecio*	1250	Melastomataceae	*Miconia*	1000
Asteraceae	*Helichrysum*	600	Melastomataceae	*Medinilla*	400
Asteraceae	*Centaurea*	500	Moraceae	*Ficus*	750
Asteraceae	*Cousinia*	500	Myrslnaceae	*Ardisia*	250
Asteraceae	*Veronina*	500	Myrtaceae	*Syzygium*	1000
Asteraceae	*Artemesia*	350	Myrtaceae	*Eucalyptus*	600
Asteraceae	*Saussurea*	300	Myrtaceae	*Eugenia*	550
Asteraceae	*Aster*	250	Myrtaceae	*Myricia*	250
Balsaminaceae	*Impatiens*	850	Oleaceae	*Jasminum*	200
Begoniaceae	*Begonia*	1400	Orchidaceae	*Pleurothallis*	1120
Berberidaceae	*Berberis*	500	Orchidaceae	*Bulbophyllum*	1000
Bromeliaceae	*Tillandsia*	380	Orchidaceae	*Dendrobium*	900
Caryophylaceae	*Silene*	700	Orchidaceae	*Epidendrum*	800
Clusiaceae	*Hypericum*	370	Orchidaceae	*Oncidium*	680
Convolvulaceae	*Ipomoea*	650	Orchidaceae	*Habenaria*	600
Cyatheaceae	*Cyathea*	620	Oxalidaceae	*Oxalis*	700
Cyperaceae	*Carex*	2000	Pandanaceae	*Pandanus*	700
Cyperaceae	*Cyperus*	300	Passifloraceae	*Passiflora*	430
Dioscoriaceae	*Dioscorea*	850	Piperaceae	*Piper*	2000
Dryopteridaceae	*Diplazium*	400	Piperaceae	*Peperomia*	1000
Ebenaceae	*Diospyros*	475	Poaceae	*Panicum*	500
Ericaceae	*Rhododendron*	850	Polygonaceae	*Polygala*	500
Ericaceae	*Erica*	735	Primulaceae	*Primula*	425
Ericaceae	*Vaccinium*	450	Ranunculaceae	*Ranunculus*	600
Eriocaulaceae	*Paepalanthus*	485	Rosaceae	*Potentilla*	500
Eriocaulaceae	*Eriocaulon*	400	Rubiaceae	*Psychotria*	800-1500
Euphorbiaceae	*Euphorbia*	2000	Rubiaceae	*Pavetta*	400
Euphorbiaceae	*Croton*	750	Rubiaceae	*Galium*	300
Euphorbiaceae	*Phyllanthis*	600	Rubiaceae	*Ixora*	300
Euphorbiaceae	*Acalypha*	430	Salicaeae	*Salix*	400
Fabaceae	*Astragalus*	1750	Scrophulariacae	*Peduncularis*	350
Fabaceae	*Acacia*	1200	Selaginellaceae	*Selaginella*	700
Fabaceae	*Indigofera*	700	Solanaceaea	*Solanum*	1700
Fabaceae	*Crotalaria*	600	Urticaceae	*Pilea*	200
Fabaceae	*Mimosa*	480	Violaceae	*Viola*	400
Fabaceae	*Desmodium*	450			
Fabaceae	*Tephrosia*	400			
Fagaceae	*Quercus*	400			
Gentianaceae	*Gentiana*	361			
Geraniaceae	*Geranium*	300			
Guttiferae	*Garcinia*	200			
Lamiaceae	*Salvia*	900			

Authorities as given in table 14.3

14.2: Families which contain large genera (as listed in Minelli, 1993), arranged according to total number of species (from Mabberley, 1997).

GRADE	FAMILY	No. GENERA	No. SPP	No.SPP / No. GENERA
DICOT	Asteraceae	1528	22750	14.89
MONOCOT	Orchidaceae	788	18500	23.48
DICOT	Fabaceae	642	18000	28.04
DICOT	Rubiaceae	630	10200	16.19
MONOCOT	Poaceae	668	9500	14.22
DICOT	Euphorbiaceae	313	8100	25.88
DICOT	Lamiaceae	252	6700	26.59
DICOT	Scrophulariacae	269	5100	18.96
DICOT	Melastomataceae	188	4950	26.33
DICOT	Myrtaceae	129	4620	35.81
MONOCOT	Cyperaceae	98	4350	44.39
DICOT	Acanthaceae	229	3450	15.07
DICOT	Ericaceae	107	3400	31.78
DICOT	Piperaceae	8	3000	375.00
DICOT	Solanaceaea	94	2950	31.38
DICOT	Lauraceae	52	2850	54.81
DICOT	Rosaceae	95	2825	29.74
DICOT	Ranunculaceae	62	2450	39.52
MONOCOT	Bromeliaceae	59	2400	40.68
DICOT	Caryophylaceae	87	2300	26.44
FERN	Dryopteridaceae	47	1700	36.17
DICOT	Convolvulaceae	56	1600	28.57
DICOT	Clusiaceae	45	1370	30.44
MONOCOT	Araceae	47	1325	28.19
DICOT	Gentianaceae	78	1225	15.71
DICOT	Myrsinaceae	33	1225	37.12
DICOT	Polygonaceae	46	1100	23.91
DICOT	Moraceae	38	1100	28.95
DICOT	Urticaceae	48	1050	21.88
MONOCOT	Eriocaulaceae	9	1000	111.11
DICOT	Begoniaceae	2	900	450.00
DICOT	Dioscoriaceae	8	880	110.00
MONOCOT	Pandanaceae	3	875	291.67
MONOCOT	Alliaceae	30	850	28.33
DICOT	Primulaceae	22	825	37.50
DICOT	Balsaminaceae	2	820	410.00
DICOT	Violaceae	20	800	40.00
DICOT	Oxalidaceae	6	775	129.17
FERN	Aspleniaceae	1	720	720.00
DICOT	Geraniaceae	11	700	63.64
	Selaginellaceae	1	700	700.00
DICOT	Berberidaceae	15	680	45.33
TREE FERN	Cyatheaceae	1	620	620.00
DICOT	Oleaceae	24	615	25.63
DICOT	Passifloraceae	17	575	33.82
FERN	Lomariopsidaceae	6	525	87.50
DICOT	Ebenaceae	2	485	242.50
DICOT	Salicaeae	2	435	217.50
DICOT	Aquifoliaceae	4	420	105.00
CLUBMOSS	Lycopodiaceae	4	380	95.00

14.3: Large vascular plant genera which appear in the Plant Fossil Record, arranged by genus size

FAMILY	GENUS	No. SPP		FOSSIL RECORD
Cyperaceae	*Carex* L.	2000		Pliocene
Euphorbiaceae	*Euphorbia* L.	2000		
Piperaceae	*Piper* L.	2000		
Fabaceae	*Astragalus* L.	1750		Pleistocene
Solanaceaea	*Solanum* L.	1700		
Begoniaceae	*Begonia* L.	1400		
Asteraceae	*Senecio* L.	1250		
Fabaceae	*Acacia* Miller	1200		Oligocene
Rubiaceae	*Psychotria* L.	1200		
Orchidaceae	*Pleurothallis* R.Br.	1120		
Orchidaceae	*Bulbophyllum* Thouars.	1000		
Melastomataceae	*Miconia* Ruiz & Pavon	1000		
Piperaceae	*Peperomia* Ruiz & Pavon	1000		
Myrtaceae	*Syzygium* Gaertner	1000		
Orchidaceae	*Dendrobium* Sw.	900		
Lamiaceae	*Salvia* L.	900		
Dioscoriaceae	*Dioscorea* L.	850		
Balsaminaceae	*Impatiens* L.	850		
Ericaceae	*Rhododendron* L.	850		Eocene
Orchidaceae	*Epidendrum* L.	800		
Euphorbiaceae	*Croton* L.	750		
Moraceae	*Ficus* L.	750		Cretaceous
Ericaceae	*Erica* L.	735		Oligocene
Aspleniaceae	*Asplenium* L.	720		Cretaceous
Araceae	*Anthurium* Schott	700		
Fabaceae	*Indigofera* L.	700		Miocene
Oxalidaceae	*Oxalis* L.	700		
Pandanaceae	*Pandanus* Parkinson	700		
Selaginellaceae	*Selaginella* Pal.	700		Cretaceous
Caryophylaceae	*Silene* L.	700		
Alliaceae	*Allium* L.	690		Pleistocene
Orchidaceae	*Oncidium* Sw.	680		
Convolvulaceae	*Ipomoea* L.	650		
Cyatheaceae	*Cyathea* Sm.	620		
Fabaceae	*Crotalaria* L.	600		
Myrtaceae	*Eucalyptus* L'Herit.	600		Cretaceous
Orchidaceae	*Habenaria* Willd.	600		
Asteraceae	*Helichrysum* Miller	600		
Acanthaceae	*Justicia* L.	600		
Euphorbiaceae	*Phyllanthis* L.	600		
Ranunculaceae	*Ranunculus* L.	600		Oligocene
Myrtaceae	*Eugenia* L.	550		Cretaceous
Berberidaceae	*Berberis* L.	500		Pliocene
Asteraceae	*Centaurea* L.	500		
Asteraceae	*Cousinia* Cass.	500		
Poaceae	*Panicum* L.	500		Oligocene
Polygonaceae	*Polygala* L.	500		
Rosaceae	*Potentilla* L.	500		
Asteraceae	*Veronina* L.	500		
Eriocaulaceae	*Paepalanthus* Kunth	485		
Fabaceae	*Mimosa* L.	480		Pliocene
Ebenaceae	*Diospyros* L.	475		Cretaceous
Fabaceae	*Desmodium* Desv.	450		Miocene
Poaceae	*Festuca* L.	450	*	
Ericaceae	*Vaccinium* L.	450		Oligocene
Euphorbiaceae	*Acalypha* L.	430		
Asteraceae	*Mikania* Willd.	430	*	
Passifloraceae	*Passiflora* L.	430		
Primulaceae	*Primula* L.	425		
Lamiaceae	*Clerodendrum* L.	400		
Dryopteridaceae	*Diplazium* Sw.	400		

Lomariopsidaceae	*Elaphoglossum* Schott ex J.Sm.	400		
Eriocaulaceae	*Eriocaulon* L.	400		
Aquifoliaceae	*Ilex* L.	400		Cretaceous
Lauraceae	*Litsea* Lam.	400		Cretaceous
Melastomataceae	*Medinilla* Gaudich	400		
Rubiaceae	*Pavetta* L.	400		
Fagaceae	*Quercus* L.	400		Palaeocene
Salicaeae	*Salix* L.	400		Miocene
Fabaceae	*Tephrosia* Pers.	400		
Violaceae	*Viola* L.	400		
Bromeliaceae	*Tillandsia* L.	380		
Araceae	*Philodendron* Schott	375	*	
Clusiaceae	*Hypericum* L.	370		
Gentianaceae	*Gentiana* L.	361		
Asteraceae	*Artemesia* L.	350		Neogene
Fabaceae	*Inga* Miller	350	*	
Orchidaceaae	*Liparis* Rich.	350	*	
Lauraceae	*Ocotea* Aublet	350	*	
Scrophulariacae	*Peduncularis* L.	350		
Cyperaceae	*Cyperus* L.	300		
Rubiaceae	*Galium* L.	300		
Geraniaceae	*Geranium* L.	300		
Lycopodiaceae	*Huperzia* Bernh.	300		
Lamiaceae	*Hyptis* Jacq.	300		
Rubiaceae	*Ixora* L.	300		
Juncaceae	*Juncus* L.	300	*	
Campanulaceae	*Lobelia* L.	300	*	
Actinidiaceae	*Saurauia* Willd.	300	*	
Asteraceae	*Saussurea* DC	300		
Ranunculaceae	*Clematis* L.	295	*	
Geraniaceae	*Pelargonium*	270	*	
Myrsinaceae	*Ardisia* Sw.	250		Oligocene
Asteraceae	*Aster* L.	250		Pleistocene
Capparidaceae	*Capparis* L.	250	*	
Myrtaceae	*Myrcia* DC ex Guillemin	250		
Apiaceae	*Eryngium* L.	240	*	
Poaceae	*Digitaria* Haller	220	*	
Iridaceae	*Iris* L.	210	*	
Brassicaceae	*Erysimum* L.	200	*	
Clusiaceae	*Garcinia* L.	200		
Oleaceae	*Jasminum* L.	200		
Verbeniaceae	*Lippia* L.	200	*	
Cactaceae	*Opuntia* Miller	200	*	
Hyacinthaceae	*Ornithogalum* L.	200	*	
Urticaceae	*Pilea* Lindley	200		
Lamiaceae	*Plectranthus* L'Herit.	200	*	
Dryopteridaceae	*Polystrichum* Roth	200	*	
Anacardiaceae	*Rhus* L.	200	*	

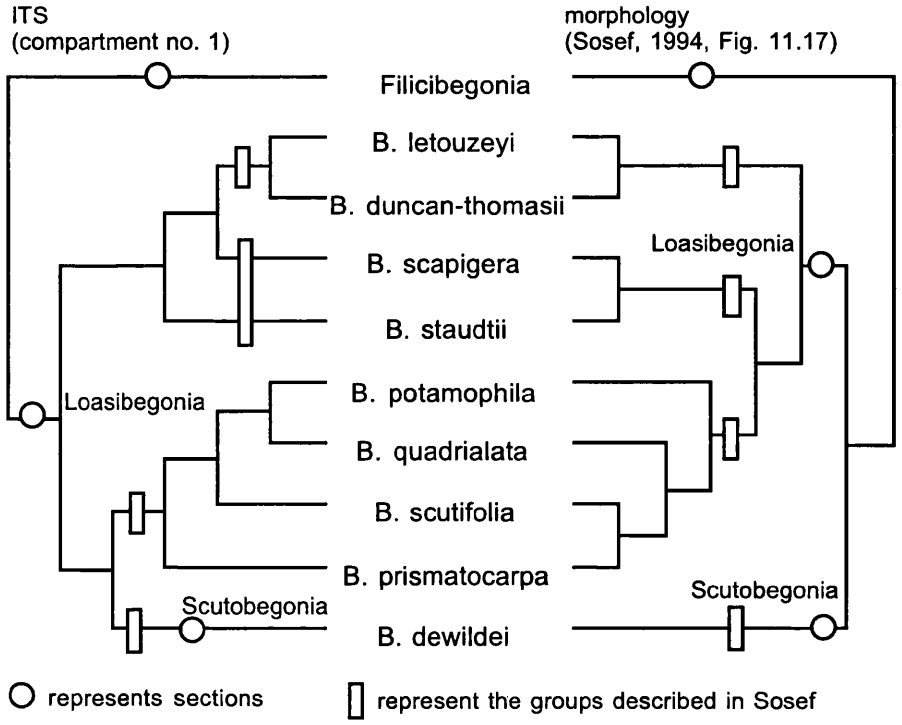
* denotes genera not in Minelli's list (1993) but found in a quick look through Mabberley (1997). List therefore appears to be less reliable for the 'smaller' genera; those of less than 400 species are not included in any further discussion.

Fossil record data from the Plant Fossil Record (http://ibs.uel.ac.uk/palaeo/pfr2/pfr.htm) from searching by extant genus name. Therefore form genera will not have been found.

14.4: Comparison between the ITS tree, *Loasibegonia / Scutobegonia,* and Sosef's (1994) tree

Sosef (1994) produced a cladogram for the sections *Loasibegonia* and *Scutobegonia*. This cladogram was based on an analysis of 132 morphological and anatomical characters, for 44 taxa (four of which were regarded as outgroups). Of all these ingroup taxa, only nine were sequenced for ITS. Still, a quick comparison can be made between the relationships suggested by Sosef's analysis, and those suggested by the ITS analysis. The branching diagram on the left is the most parsimonious tree for the ITS data set (from Chapter 7, Figure 7.7, rooted on *B. aspleniifolia*); the branching diagram on the right is not a most parsimonious tree for Sosef's data, but represents the relationships suggested by pruning all the extraneous taxa from his "conclusive cladogram" (Sosef, 1994, Figure 11.17, p. 111, rooted on section *Filicibegonia*).

Figure 14.1: Comparison between ITS MPT & Sosef's analysis



The only species from section *Scutobegonia* which was included in the ITS analysis is *B. dewildei*. Its position changes radically between the two trees: in the ITS tree, section *Loasibegonia* is paraphyletic, and includes section *Scutobegonia*; Sosef's tree is consistent with both sections being monophyletic (as is shown in his unpruned tree). It may be that the inclusion of more taxa from section *Scutobegonia* in the ITS analyses will pull *B. dewildei* out of section *Loasibegonia*. Unfortunately the other taxon from this section which is in cultivation, *B. hirsutula*, did not amplify for ITS. It would also be useful to rerun Sosef's analyses using only the taxa included in the ITS analysis.

Sosef breaks the taxa he has examined into seven monophyletic groups. ITS shows a sister-group relationships between *B. letouzeyi* and *B. duncan-thomasii*. In Sosef's cladogram there are six taxa in this clade; he calls it 'the *B. letouzeyi* group'. *B. scapigera* and *B. staudtii* belong to Sosef's '*B. scapigera* group'. This is not supported by the ITS data, which resolve this group as paraphyletic with the '*B. letouzeyi* group'. Sosef's '*B. potamophila* group' is represented here by *B. potamophila*, *B. prismatocarpa*, *B. scutifolia* and *B. quadrialata*. Although ITS supports the monophyly of this group, it resolves the relationships within it differently. *B. dewildei* is the only included representative from Sosef's '*B. ferramica* group'.

For an example of the morphological characters which hold together one of Sosef's groups, see his Figure 11.11 (p. 105), the *'B. scapigera* group'. This is supported by three synapomorphies:

Ch. 65: Ovary shape (narrowly elliptical - obovate TO narrowly oblong - narrowly elliptic)
Ch. 71: Wing shape (linear - obovate TO linear)
Ch. 114:Placenta shape (lobed, thickened TO not or weakly lobed, strongly thickened).

Each of these characters has overlapping states; further, character 114 has a reversal in *B. staudtii*. In the ITS analysis, the clade with *B. scapigera*, *B. duncan-thomasii* and *B. letouzeyi* has Bremer support value 12, and 100% bootstrap support. The ITS analysis is not necessarily more reliable, but caution needs to be taken with morphology, especially when analyses include quantitative or overlapping characters.

14.5 Herbarium specimens included in morphological analyses (from E).

B. aequata A. Gray. Wilkie, P., Argent, C.G.C., Mendum, M., Pennington, R.T., Romero, E.M. & Fuentes, R.E. Philippines RBGE accession 1997 2515: Luzon Island: Camarines Sur.: Naga Province. Barangay Panicuason: Mt. Isarog, west slope. On tree in lower submontane forest, 1200 m. 13°39' N,123°21'E. Climber.

B. formosana (Hayata) Masamune. Edinburgh Taiwan Expedition (1993) no. 24, 31 x 1993. Taiwan: Maioli Hsian, Tahsueh Shan, line 210 at 25 km. Warm temperate coniferous forest, codominant with Fagaceae. Shady moist woodland slopes in very organic soil. 2145 m. 24°15' N, 121°5'E.

B. oxysperma A.DC. Wilkie, P., Argent, C.G.C., Mendum, M., Pennington, R.T., Romero, E.M. & Fuentes, R.E. No. 29142. Philippines: Luzon Island: Camarines Sur.: Naga Province. Barangay Panicuason: Mt. Isarog, west slope. On tree in lower submontane forest, mostly on tree ferns, 1200 m. 13°39' N, 123°21'E. Epiphytic climber. RBGE accession 1997 2519.

B. rufo-serica Toledo C11195, 4 v 1977. RBGE accession no. 1964 3108, G35. RBGE cultivated plants.

B. serratipetala Irmsch. Reeves no. 588, vii 1983. Waimeram, Paiela Census Division, Porgera District, Enga Province. Terrestrial - old garden, also planted near houses by local people. 1800 m.

B. sp. '*exotica*' (= *B.* cf. *brevirimosa*). T.M. Reeves no. 142, xii 1981. Korombi: Paiela Census Division, Purgera District, Enga Province. Terrestrial in shaded forest. 1500 m.

B. sp., Sulawesi 252. Argent, G., Mendum, M. & Hendrian no. 00116, 20 ii 2000. Lake Poso, south Sulawesi, c. 2°24' S, 120°48' E. Roadside ditch in shade in disturbed rain forest. c. 1150 m.

B. sp., Sulawesi 253. Argent, G., Mendum, M. & Hendrian no. 00151, 25 ii 2000. Mt Sojol, Central Sulawesi, c. 0° 40' S, 120°10' E. Valley bottom in shade of rain forest. c. 600 m.

B. sp., Sulawesi 254. Argent, G., Mendum, M. & Hendrian no. 00152, 25 ii 2000. Mt Sojol, Central Sulawesi, c. 0° 40' S, 120°10' E. Valley bottom in shade of rain forest. c. 600 m.