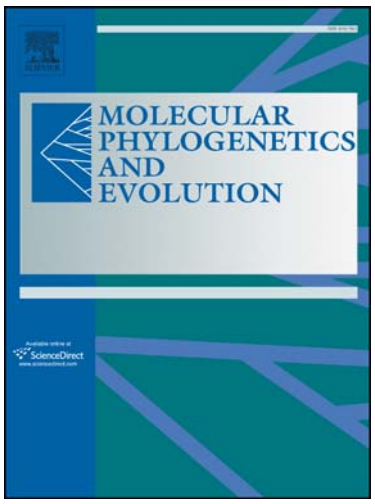
Accepted Manuscript



Species trees from consensus Single Nucleotide Polymorphism (SNP) data: testing phylogenetic approaches with simulated and empirical data

Alexander N. Schmidt-Lebuhn, Nicola C. Aitken, Aaron Chuah

PII: S1055-7903(17)30540-7
DOI: http://dx.doi.org/10.1016/j.ympev.2017.07.018
Reference: YMPEV 5881

To appear in: *Molecular Phylogenetics and Evolution*

Received Date: 18 July 2016
Revised Date: 21 February 2017
Accepted Date: 22 July 2017

Please cite this article as: Schmidt-Lebuhn, A.N., Aitken, N.C., Chuah, A., Species trees from consensus Single Nucleotide Polymorphism (SNP) data: testing phylogenetic approaches with simulated and empirical data, *Molecular Phylogenetics and Evolution* (2017), doi: http://dx.doi.org/10.1016/j.ympev.2017.07.018

Species trees from consensus Single Nucleotide Polymorphism (SNP) data: testing phylogenetic approaches with simulated and empirical data

Alexander N. Schmidt-Lebuhn[a,]*, Nicola C. Aitken[b], Aaron Chuah[c]

[a]*CSIRO, Australian National Herbarium, Clunies Ross Street, Canberra ACT 2601, Australia*

[b]*Research School of Biology, Australian National University, Canberra ACT 2601, Australia*

[c]*The John Curtin School of Medical Research, Australian National University, Canberra ACT 2601, Australia*

*Corresponding author at: CSIRO, Clunies Ross Street, Canberra ACT 2601, Australia.

E-mail address: Alexander.S-L@csiro.au

ABSTRACT

Datasets of hundreds or thousands of SNPs (Single Nucleotide Polymorphisms) from multiple individuals per species are increasingly used to study population structure, species delimitation and shallow phylogenetics. The principal software tool to infer species or population trees from SNP data is currently the BEAST template SNAPP which uses a Bayesian coalescent analysis. However, it is computationally extremely demanding and tolerates only small amounts of missing data. We used simulated and empirical SNPs from plants (Australian *Craspedia*, Asteraceae, and *Pelargonium*, Geraniaceae) to compare species trees produced (1) by SNAPP, (2) using SVD quartets, and (3) using Bayesian and parsimony analysis with several different approaches to summarising data from multiple samples into one set of traits per species. Our aims were to explore the impact of tree topology and missing data on the results, and to test which data summarising and analyses approaches would best approximate the results obtained from SNAPP for empirical data. SVD quartets retrieved the correct topology from simulated data but with very divergent branch length distributions, as did SNAPP except in the case of a very unbalanced phylogeny. Both methods failed to retrieve the correct topology when large amounts of data were missing. Bayesian analysis of species

level summary data scoring the two alleles of each SNP as independent characters and parsimony analysis of data scoring each SNP as one character produced trees with branch length distributions closest to the true trees on which SNPs were simulated. For empirical data, Bayesian inference and Dollo parsimony analysis of data scored allele-wise produced phylogenies most congruent with the results of SNAPP. In the case of study groups divergent enough for missing data to be phylogenetically informative (because of additional mutations preventing amplification of genomic fragments or bioinformatic establishment of homology), scoring of SNP data as a presence/absence matrix irrespective of allele content might be an additional option. As this depends on sampling across species being reasonably even and a random distribution of non-informative instances of missing data, however, further exploration of this approach is needed. Properly chosen data summary approaches to inferring species trees from SNP data may represent a potential alternative to currently available individual-level coalescent analyses especially for quick data exploration and when dealing with computationally demanding or patchy datasets.

Keywords:

Coalescent model

Craspedia

Genotyping-by-Sequencing

Parsimony

Pelargonium

Single Nucleotide Polymorphism

1. Introduction

1.1. Species trees

Over the past decade, it has become feasible to produce genomic data for ever larger numbers of specimens (McCormack et al., 2013). This wealth of data enables phylogeneticists, population geneticists and evolutionary biologists to address evolutionary questions and to resolve phylogenetic relationships that had remained intractable with the limited numbers of independent markers provided by Sanger Sequencing or traditional genotyping approaches such as microsatellites. However, researchers are now faced with new challenges of analysing their data in an appropriate way (Freudenstein et al., 2003; Lemmon and Lemmon, 2013; Misof et al., 2013). These challenges are both methodological and computational, as available software tools may struggle with large datasets and large numbers of samples (Raj et al., 2014).

Among the methodological challenges is the potential for genomic data to show conflicting signals. In the case of datasets comprising a limited number of sequence regions, the observations of incongruent gene phylogenies and of gene trees showing the sequence copies from individual species as non-monophyletic are now well understood. They can be explained by ancestral polymorphism, also known as incomplete lineage sorting, or by recent hybridisation (Knowles, 2009; Maddison, 1997; Maddison and Knowles, 2006; Szöllősi et al., 2015). Numerous analytic tools have been developed to infer species phylogenies from multiple gene trees, especially under the assumption of ancestral polymorphism.

The most important parsimony approaches are Minimising Deep Coalescences (MDC) and Minimising Gene Duplications and Gene Losses (Maddison, 1997), implemented in software such as Mesquite (Maddison and Maddison, 2011) or iGTP (Chaudhary et al., 2010). Other tools use distance-based (Shaw et al., 2013), Likelihood (Liu, 2008) or Bayesian coalescent methods (Heled and Drummond, 2010). Efficient algorithms using the coalescent model such as ASTRAL-II (Mirarab and Warnow, 2015) allow the analysis of large datasets with hundreds of taxa and hundreds of genes. Progress has also been made in including both ancestral polymorphism and recent hybridisation into the same analysis, e.g. with Most Parsimonious Reconciliations (Doyon et al., 2010). At the same time, the value of concatenating all data continues to be explored (Gadagkar et al., 2005; Tonini et al., 2015), and an approach has been suggested for

deciding when this approach is preferable (McVay and Carstens, 2013). Consequently, phylogeneticists are now well equipped to use sequence data from multiple loci to infer phylogenies in the genomic era with the method of their preference.

1.2. Genome-wide SNPs

An increasingly popular approach at the shallowest phylogenetic levels and in population genetics is the generation of thousands of Single Nucleotide Polymorphisms (SNPs) from across the genome using methods such as reduced representation shotgun sequencing (Altshuler et al., 2000), RAD-seq (Baird et al., 2008) and Genotyping-by-Sequencing (Elshire et al., 2011). The resulting data are often used to infer population structure or delimit species, but can also be employed to study the phylogenetic relationships within species complexes too recently diverged to be resolved on the basis of DNA sequences of single loci each of which contain only few polymorphisms (Lambert et al., 2013; Rheindt et al., 2014).

In reduced representation, missing data can be caused by several factors with very different implications (Davey et al., 2011). Due to low sequencing depth and the additional vagaries of fragment amplification, any marker may not be scored for some samples for purely stochastic reasons. On the other hand, missing data can be caused by mutations in restriction sites for RAD-seq and Genotyping-by-Sequencing, which will prevent a fragment from amplifying or passing size selection, or by additional mutations in the fragment, which may lead to homology rejection by bioinformatic analysis pipelines. The frequency of the latter processes is correlated with the phylogenetic depth of the study group (Cariou et al., 2013), suggesting the possibility that present or absent data may carry a phylogenetic signal, at least assuming that missing data caused by the former processes are distributed randomly.

In contrast to multi-gene datasets, at present there is only a limited number of tools for the inference of species phylogenies based on SNP data from multiple samples per species. The most commonly used software package is BEAST (Bouckaert et al., 2014) with its template SNAPP (Bryant et al., 2012), which implements a Bayesian coalescent analysis. Unfortunately, it is computationally very demanding for larger datasets (Yoder et al., 2013) and will tolerate only small amounts of missing data (Jason Bragg, pers. comm.; A.N.S.-L., pers. obs.). An alternative software is PoMo (Maio et al., 2015), but its documentation indicates

that no missing data are allowed, severely limiting its utility with reduced representation data. A faster alternative also using the coalescent approach was developed by Chifman and Kubatko (2014) under the name SVDquartets and has been implemented in PAUP* 4a149 (http://people.sc.fsu.edu/~dswofford/paup_test/). Other likelihood models for species trees from SNPs exist (RoyChoudhury et al., 2008), but to our knowledge no other software appears to be available at this time.

1.3. Motivation

To resolve phylogenetic relationships at the level of young genera or species complexes, sequence data may be insufficient due to their low divergence. Marker systems such as SNPs from Genotyping-by-Sequencing represent an attractive alternative because they can provide numerous characters for non-model organisms. We have generated SNP data for difficult-to-resolve groups (*Craspedia* G.Forst., see below; *Ozothamnus ledifolius* complex, M. de Salas & A.N. Schmidt-Lebuhn, unpubl. data) to ultimately infer phylogenetic relationships at the species level. However, due to the large size of the datasets and significant amounts of missing data, we found it impossible to make use of SNAPP, prompting us to explore other options.

1.4. Aims of the study

In the present paper, we investigate several distinct ways of summarising SNP data from multiple samples per species into one set of traits per species, and Bayesian and several distinct parsimony approaches to infer species trees from the summarised data. Using both empirical and simulated data, and comparing resulting trees against each other and against information from other sources, we aim to: (1) explore the performance of different approaches when analysing simulated datasets, (2) explore which Bayesian and parsimony approaches using species-level summarised data provide results congruent with SNAPP, (3) compare scoring of SNP data as present or absent (missing) against scoring SNPs by their allele values, and thus (4) find the most defensible and useful summary method for species trees from SNP data that for application to large or patchy datasets that are intractable with currently available individual level,

coalescent-based approaches.

2. Materials and methods

2.1. SNP data sets

To test methods on simulated datasets where the true phylogeny is known, we constructed three artificial trees of eight species and used SIMCOAL 2 (Laval and Excoffier, 2004) to simulate 500 SNPs in each case. Each population had a constant effective size of 5,000, there was no migration between populations, and recombination rates were set to 100%. Lineage splits were designed to occur after varying times with a minimum of 5,000 generations. One tree was designed to be completely balanced (Fig. 1A), the second to be completely unbalanced (Fig. 1G), and the third to show a mixture of isolated and closely related lineages with varying divergence times (Fig. 1M). We then randomly deleted 5%, 25% and 75% of the individual-level data matrix simulated on the mixed tree to explore the effect of different amounts of randomly distributed missing data. To make the run time of SNAPP less prohibitive, we restricted the samples to three per species.

We generated empirical SNP data used in this study with Genotyping-by-Sequencing (Elshire et al., 2011) as described in detail by Nicotra et al. (2016). Briefly, genomic DNA was digested with PstI and ligated to uniquely barcoded sequencing adaptor pairs. Samples were then individually PCR amplified and pooled in an equimolar manner. Library amplicons between 250-600 bp were extracted from an agarose gel and sequenced in a HiSeq2000 using a 100 bp Paired End protocol (at the Biomolecular Resource Facility at the Australian National University). SNP calling was conducted by the BRF Genome Discovery Unit using the TASSEL UNEAK approach (Lu et al., 2013).

We used two empirical datasets to test analysis methods. The first included 8,958 SNPs for 240 samples of the daisy genus *Craspedia* (Asteraceae, Gnaphalieae). It comprises ca. 20 described species distributed across the southern half of Australia with the centre of diversity in alpine areas and an unknown number of species in New Zealand (Schmidt-Lebuhn and Milner, 2013). With the exception of the arid zone annual *C. haplorrhiza* J.Everett & Doust, all species are perennial rosette plants with yellow, white or rarely

orange compound heads borne individually on leafy stalks.

Molecular phylogenetic analyses using traditional nuclear ribosomal and chloroplast markers have established that *C. haplorrhiza* has a phylogenetically isolated position in the genus, and that the New Zealand species form a clade nested within the Australian species, presumably the result of a single dispersal event (Ford et al., 2007). However, relationships among the perennial Australian species remain unresolved due to low sequence divergence and the presence of several species in both major nuclear ribosomal DNA clades (Schmidt-Lebuhn, 2013).

Samples were assigned to species by *a priori* identification based on morphology. However, we treated thirteen New Zealand samples as one lineage because all local species are known to form a monophyletic group nested within Australian *Craspedia*, and because species delimitation in the New Zealand clade is insufficiently understood (Breitwieser et al., 2010). The *Craspedia* dataset was characterised by large amounts of missing data, with individual samples scored for 103 to 2,708 SNPs (average 911, median 828), potentially in part because the species were more divergent than anticipated. Species were also sampled to very varying degrees, from only one sample in the case of locally endemic *C. preminghana* Rozefelds to 31 samples in the case of *C. jamesii* J.Everett & Joy Thomps.

We also used a published dataset of Australian native *Pelargonium* L'Her. ex Aiton (Geraniaceae) (Nicotra et al., 2016) which provides an accepted phylogeny inferred with SNAPP to compare against other approaches. The dataset comprised 463 SNPs for 23 samples representing eight species, sub-sampled from an original 29,531 SNPs for 177 samples to decrease the amount of missing data and computation time.

2.2. Character scoring

To produce character matrices of the species for parsimony analysis, we summarised the SNP data from the individual samples in three different ways.

1. Locus-wise. If the species contained only the major allele the SNP was scored as state 0, if it contained only the minor allele as state 1, if it contained both as (01) polymorphic, otherwise as missing data (Fig. 2B).

2. Allele-wise. Each SNP was transformed into two individual characters, so that each allele was

scored as present (1) or absent (0) in a species (Fig. 2C, D).

3. Presence/absence (*Craspedia* only). The individual alleles were ignored, and a SNP was scored as present (1) if data were present for at least one sample from the species, otherwise as absent (0).

A Python 2.7 script (Python Software Foundation, 2016) was custom-written to automate summarising the SNP data. It requires a standard comma or tab separated text file with samples in columns and SNPs scored as 0/1/2 (homozygous/heterozygous/homozygous) in rows. It outputs three different nexus files for use in PAUP. (Note to referees: For review purposes the script is Supplementary Data S3, and it will ultimately be made available on the first author's institutional website.)

2.3. Phylogenetic analyses

Data scored locus-wise were treated as standard Wagner characters (Kluge and Farris, 1969), with multistate characters specified as polymorphic (Fig. 2B). The matrices resulting from allele-wise and presence/absence scoring were subjected to parsimony analyses under three different character optimisations: Wagner parsimony (Fig. 2C), Dollo parsimony (Farris, 1977) (Fig. 2D) and a step matrix counting allele gains as twice as expensive as allele losses. In all cases, heuristic searches were conducted in PAUP* 4.0b10 (Swofford, 2003) with default parameters but MaxTrees set to 1,000 and ten addition sequence replicates. Branch support was inferred with 200 bootstrap replicates. Bayesian inference was conducted in MrBayes 3.2.3 (Ronquist et al., 2012) for data summarized at the species level and using the restriction site model, coding parameter set to “all”, two runs with two chains for 1,000,000 generations at temperature = 0.2, and sampling every 500 generations. Convergence of runs was verified from MrBayes' diagnostic values.

SNAPP species trees were inferred in BEAST 2.1.3 (Bouckaert et al., 2014) using default priors. At least ten million generations were run sampling every 1,000 generations. We examined trace shape and effective sample size (ESS) with Tracer 1.6 (Rambaut et al., 2013) and terminated runs either when the trace had stabilised and ESS values for all parameters were satisfactory or when they showed no sign of improvement even after 15 million generations. Summary trees were produced with TreeAnnotator 2.1.2.

In all successful Bayesian analyses, burn-in was set to exclude samples saved before stationarity was achieved. For analyses of simulated data that failed to achieve stationarity, we discarded the first 50% of

saved trees and summarised the remainder to examine how close the analysis came to inferring the underlying tree.

SVDquartet analysis was conducted in PAUP 4a149 (http://people.sc.fsu.edu/~dswofford/paup_test/) on data scored at the level of the individual, with species set as taxon partitions. We used default settings except for the species tree option and running 200 bootstrap replicates.

2.4. Rooting

Coalescent analyses and parsimony analyses using an asymmetric step-matrix produce rooted trees. In other analyses, trees for simulated data were rooted on the known outgroup, and the phylogeny of Australian *Pelargonium* was, where possible, outgroup-rooted on *P. havlasae* Domin and *P. littorale* Hügel following the results of Nicotra et al. (2016). The *Craspedia* dataset lacks an outgroup because no high quality sample of *C. haplorrhiza* was available, and consequently we focused on comparing the topologies of unrooted trees.

2.5. Criteria for tree evaluation

In the case of simulated data, agreement with the phylogeny on which the data were simulated was the only criterion for tree evaluation. In the case of *Pelargonium*, we were interested in agreement of results from parsimony analysis with the results from SNAPP. Pairwise K tree scores (Soria-Carrasco et al., 2007) were calculated from unrooted trees to compare topologies and branch length distributions. Although branch lengths are not directly comparable between parsimony and coalescent trees, a branch length distribution vastly different from the true (simulated) one would be undesirable e.g. for analyses of phylogenetic diversity (Faith, 1992).

In addition to comparing results across analyses, the *Craspedia* dataset was used to test the feasibility of using presence-absence scoring of SNP data. Accordingly, congruence of results from this scoring approach with those from other scoring approaches was the criterion. We will use the term clan (Wilkinson et al., 2007) to refer to whole branches of unrooted trees.

3. Results

3.1. Simulated data

SNAPP correctly inferred the balanced tree from simulated data, although with short terminal branch lengths (Fig. 1B). It also inferred the correct topology with mixed characteristics, but likewise with short terminal branches, and support for deeper relationships was low (Fig. 1N). After 15 million generations, SNAPP inferred the relationships between four species on the unbalanced tree but was unable to resolve the rest of the topology (Fig. 1H). The trace was still unstable and showed no sign of improvement. SVD quartet analysis in PAUP correctly inferred all three topologies, but branch length distribution diverged strongly from the true trees (Fig. 1C, I, O).

Bayesian summary trees and parsimony trees from summarised data were topologically identical with the true trees (Fig. 1D-F, J-L, P-R). Wagner and Dollo analyses of data scored allele-wise resulted in identical trees because the SNP simulation assumed only one mutation for every locus (not shown; these and all other trees are available in Supplementary Data S2). The step-matrix approach inferred the true root position for the unbalanced and mixed but not for the balanced tree (Fig. 1F, L, R).

3.2. Missing data

In our analyses of 500 simulated and pooled SNPs with varying amounts of missing data, performance of SNAPP differed depending on the amount of missing data. It quickly found the correct tree for 5% missing data, but all except three branches were extremely short (Fig. 3A), and even after fifteen million generations ESS for the topology posterior and likelihood were still <200, and the trace was still unstable. With 25% missing data, the correct topology was inferred but with the wrong root (Fig. 3G). After 10 M generations ESS was above 700 for all parameters, and the trace was stable. In the case of 75% missing data, SNAPP was unable to resolve the tree topology after 10 M generations, with posterior probability (PP) of 0.1 or less for all relationships (Fig. 3M). Nonetheless the trace appeared very stable and ESS was

indicated to be above 7,000 for every parameter. SVD quartets failed to reconstruct the true topology in all cases (Fig. 3B, H, N).

Bayesian and parsimony analyses of summarised data inferred the true tree in all cases except one (Fig. 3C-F, I-L, O-R): For 75% missing data, the step-matrix approach inferred a tree showing as two two species clades on a grade what should have been a clade of four species (Fig. 3R). However, it consistently succeeded in inferring the correct root position.

3.3. Pelargonium

SVD quartet analysis of the *Pelargonium* dataset retrieved a topology (Fig. 4B) that was largely compatible with the strongly supported relationships in the Bayesian summary tree produced by SNAPP (Fig. 4A) except for the separation of *P. havlasae* and *P. littorale*.

Bayesian analysis of data summarised at the species level supported the species pairs *P. havlasae* / *P. littorale* and *P. rodneyanum* Lindl. / *P. striatellum* but not the clade of the remaining four species (Fig. 4C). Locus-wise scoring and Wagner parsimony analysis resulted in four equally parsimonious trees with poor branch support (Fig. 4D). Except for the sister species relationship of *P. havlasae* and *P. littorale*, the trees showed little congruence with the results of SNAPP. The single most parsimonious tree for allele-wise scoring analysed with Wagner parsimony showed a higher degree of congruence with the SNAPP results in that *P. australe* Willd. and *P. drummondii* Turcz. as well as *P. helmsii* Carolin and *P. inodorum* Willd. were retrieved as sister species (not shown), but those relationships had only weak support from SNAPP. In addition, the sister group relationship of the two remaining species, which was strongly supported by SNAPP, was not inferred by the parsimony analysis. Finally, the single most parsimonious trees resulting from the Dollo (Fig. 4E) and step-matrix (Fig. 4F) analyses of data scored allele-wise were consistent with the SNAPP tree; relationships inside the clade of *P. australe*, *P. drummondii*, *P. helmsii* and *P. inodorum* were resolved differently, but there was little support for any specific topology in SNAPP. However, in contrast to SNAPP the step-matrix approach rooted the phylogeny between this clade and the remaining four species.

3.4. K tree scores

The best fit to the tree on which SNPs had been simulated was achieved by Bayesian inference from data scored allele-wise and Wagner parsimony analysis of data scored locus-wise (Table 1). SNAPP and SVD quartet analysis resulted in considerably poorer fit, with the step-matrix approach producing intermediate scores. Wagner analysis of data scored locus-wise also produced the tree with the best branch length fit to the published SNAPP phylogeny of Australian *Pelargonium*, although the two trees disagreed on well supported relationships.

3.5. Craspedia

The SVD quartet phylogeny of *Craspedia* (Fig. 5A) produced an outlier topology compared to all other analyses. However, bootstrap support (BS) was insignificant throughout. While the phylogenetic positions of the New Zealand lineage and of mainland alpine *C. leucantha* F.Muell. were distinctly unstable across analyses, the following three clans were resolved in the majority of the cases (Fig. 5C-D):

First, all analyses showed a large group of subalpine to alpine mainland species with varying leaf indumentum: *C. adenophora* K.L.McDougall & N.G.Walsh, *C. alba* J.Everett & Joy Thomps., *C. aurantia* J.Everett & Joy Thomps., *C. costiniana* J.Everett & Joy Thomps., *C. crocata* J.Everett & Joy Thomps., *C. jamesii*, *C. lamicola* J.Everett & Joy Thomps., and *C. maxgrayi* J.Everett & Joy Thomps. BS for this clan ranged from 77 to 100, and it was always supported by a PP of 1.

Second, all analyses scoring data allele-wise showed a small clan of one widespread subalpine and two Tasmanian endemic highland species, all characterised by narrow leaves with woolly indumentum: *C. glabrata* (Hook.f.) Rozefelds, *C. gracilis* Hook.f. and *C. macrocephala* Hook.; it was supported by a BS values of 88 to 98 and a PP of 1. Analyses scoring data locus-wise and Wagner parsimony analysis of data scored as present or absent resulted in the inclusion of *C. leucantha* in this clan (BS < 50 to 68, but PP 1). Other analyses scoring data as present or absent did not resolve this clan but showed its members as forming a grade between the previous and the next clan.

Third, Dollo parsimony and step-matrix analysis of data scored allele-wise produced a clan of the

lowland and coastal species *C. canens* J.Everett & Doust, *C. cynurica* Rozefelds & A.M.Buchanan, *C. glauca* (Labill.) Spreng., *C. paludicola* J.Everett & Doust, *C. rosulata* Rozefelds & A.M.Buchanan, *C. variabilis* J.Everett & Doust, and *C. preminghana* Rozefelds (BS 68 to 80). Most other scoring approaches and analyses resolved the same clan to include the New Zealand lineage (BS <50 to 100, PP 1). Analyses of data scored locus-wise and Wagner parsimony analysis of data scored as present or absent failed to resolve this clan either inclusive or exclusive of the New Zealand lineage, but there was little support for the relevant branches (BS <50 to 66, PP <0.95) with the exception of a single branch for locus-wise data under the Restriction Site Model (PP 0.99 for grouping the second clan inside the third).

All parsimony analyses retrieved a single most parsimonious tree except locus-wise scoring with Wagner parsimony, in which case four equally parsimonious trees were found.

4. Discussion

4.1. Choice of scoring approach

In the present study we examined three different ways of summarising SNP data for species-level consensus data. The first, which we called locus-wise, is comparable to producing a consensus sequence for all samples of a species or population (Fig. 2B). For analysis, multi-state characters are most appropriately treated as polymorphic, but Wagner analysis counts only gains and not losses of an allele. For simulated data, this approach consistently inferred the correct phylogeny, and often with the best fit to the true branch length distribution. In the case of *Pelargonium*, however, it did not retrieve one of the clades strongly supported by SNAPP (Nicotra et al., 2016), and for *Craspedia* locus-wise scoring resulted in poorly resolved and poorly supported relationships.

Our second approach, here called allele-wise, scored each allele as a separate character, with the allele either present (1) or absent (0) in a species (Fig. 2C, D). It follows the same reasoning as that behind a parsimony species tree under the criterion of minimising duplications and losses in a gene family (Maddison, 1997) in that it allows the number of allele gains and losses across the phylogeny (Fig. 2A) to be minimised. Results for simulated data were identical to those of the locus-wise approach, but for empirical datasets,

allele-wise scoring produced topologies that were more strongly supported and more congruent with SNAPP (Nicotra et al., 2016). In MrBayes, it produced the most accurate branch length distributions. We conclude that locus-wise scoring may not make full use of the available information, and that allele-wise scoring is more useful. It also allows the use of more different parsimony analyses than locus-wise scoring.

Finally, for the *Craspedia* dataset we also scored the SNPs themselves as either present, if there were data, or absent, if there were none. This approach treats the SNP table in the same way as restriction fragment based data such as Amplified Fragment Length Polymorphism (Vos et al., 1995), which would make sense if the study species were sufficiently distantly related that mere sharing of amplification success and successful establishment of homology for sequence reads were already indicative of relatedness.

The observation that parsimony trees resulting from presence/absence scoring showed a great degree of topological similarity to those resulting from other scoring approaches (Fig. 5C, D) confirmed our suspicion that large phylogenetic distance between study species contributed to the amount of missing data, as had been observed in other groups (Cariou et al., 2013). At the same time, it suggested that there might be some utility of presence/absence scoring in at least some SNP datasets, i.e. in those where species turned out to be more distantly related than would have been ideal for SNP calling.

However, there is the caveat that uneven sampling or uneven amplification success may have an impact on the results if the amount of missing data in some species is partially due to low sampling or if some species consistently show poorer PCR performance (Davey et al., 2011), perhaps because of sample quality or problematic secondary chemistry. In the case of our *Craspedia* phylogeny, it seems likely that the grouping of the New Zealand lineage with Australian lowland species was influenced by low reaction success in the former and lower sampling in the latter, both leading to higher amounts of missing data. Another caveat is that the tree from Bayesian inference was much more different from those using the SNP data than in the case of parsimony analysis.

We conclude that scoring SNP data as present or absent may have potential as an alternative scoring strategy at the genus level, but further study using more empirical datasets is required. The approach may be misleading at the sample level if amplification success or DNA amounts pooled for sequencing are too uneven. If using summary data across species, as in the present study, even sampling across species will be another concern.

4.2. Choice of parsimony analysis

In our simulations there was no significant topological difference between analyses of data scored allele-wise using Wagner parsimony (Kluge and Farris, 1969), Dollo parsimony (Farris, 1977) and a step-matrix penalising gains over losses, presumably because there was no homoplasy. The only exception occurred with 75% missing data, where the step-matrix approach inferred the wrong tree. However, marked differences were evident in the less unambiguous empirical datasets. For both *Pelargonium* and *Craspedia*, Dollo parsimony and the step-matrix approach produced phylogenies that were better supported by independent analysis or data sources, respectively, than Wagner parsimony. In the first case, they were more congruent with the tree derived from SNAPP, and in the second, chloroplast phylogeny (Ford et al., 2007) and morphological characters (Schmidt-Lebuhn and Milner, 2013) supported the topology to a higher degree than that of the Wagner trees. We conclude tentatively that analyses penalising allele gains over allele losses may be more appropriate for empirical datasets. Other researchers have previously argued for using a similar approach for restriction site data (Debry and Slade, 1985), and even for counting only gains (in that case gene duplications) in the context of gene tree parsimony analyses (Page and Charleston, 1997; Sanderson and McMahon, 2007).

4.3. Rooting and the utility of parsimony and likelihood species trees from SNPs

The main disadvantage of a Wagner or Dollo parsimony analyses is that the resulting trees are generally not ultrametric, representing character changes instead of coalescent units as in analyses using the coalescent model. This makes them inappropriate for analyses requiring ultrametric trees (e.g. studies of rate shifts). Perhaps most importantly, the lack of 'clock rooting' makes it necessary to use either outgroup rooting (Maddison et al., 1984), in which case it has to remain unclear what fraction of the internode connecting outgroup and ingroup belongs to either, midpoint rooting (Farris, 1972), or asymmetric step-matrices. The latter approach, however, inferred the wrong root position even for one of our homoplasy-free simulated datasets where the true root was known, and is known to be even more problematic in empirical situations

(Huelsenbeck et al., 2002). It can be assumed that midpoint rooting is most feasible if all species in the dataset are sampled to the same degree, and if there is little missing data. All *Craspedia* trees were considerably less ultrametric than the simulated ones or those of *Pelargonium*, presumably reflecting less even sampling across the species. Interestingly, midpoint rooting of the Dollo parsimony trees from simulated SNPs found the true root position when only 5% and 25% of the data were removed but was mislead by 75% missing data (not shown).

On the other hand, using an appropriate parsimony or likelihood analysis has some advantages over alternative approaches. Datasets SNAPP will analyse over weeks even on a high performance computing cluster (Yoder et al., 2013) can be analysed in minutes using SVD quartets or a combination of the species-level consensus data and parsimony methods tested in this study. Our results also indicate that they are more robust when faced with some topologies, as in our simulation on an unbalanced tree, and with appreciable amounts of missing data.

Considering especially the absence of incongruence for any well supported branches between the SNAPP analysis and the Dollo and step-matrix analyses of the *Pelargonium* dataset (Fig. 4A, D), an appropriately chosen parsimony analysis may be attractive even to researchers who otherwise prefer statistical methods, at the very least for quick data exploration. Another advantage of the data scoring and phylogenetic approaches explored in this study is their considerably higher tolerance for missing data, allowing the analysis of datasets that SNAPP would not be able to process at all.

Acknowledgements

We thank Justin Borevitz and Jason Bragg (Australian National University) for collaboration and helpful discussions, Ilse Breitwieser (Landcare Research), Miguel de Salas (Tasmanian Herbarium) and James Wood (Royal Tasmanian Botanic Gardens) for making available *Craspedia* samples, Peter Campbell (CSIRO) and David Bryant (University of Otago) for help with SNAPP analyses, Francisco Encinas Viso (CSIRO) for advice on SNP simulation, the Biomolecular Resource Facility of the Australian National University for sequencing services, and Karen Meusemann and an anonymous referee for comments on a previous version of this manuscript. This study was partly supported by a Centre of Biodiversity Analysis

Ignition Grant to A.N.S.-L. and Justin Borevitz in 2013/14.

References

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., Lander, E.S., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407, 513–516. doi:10.1038/35035083

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLOS ONE 3, e3376. doi:10.1371/journal.pone.0003376

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. Beast 2: A software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10, e1003537. doi:10.1371/journal.pcbi.1003537

Breitwieser, I., Ford, K.A., Smissen, R.D., 2010. A test of reproductive isolation among three sympatric putative species of *Craspedia* (Asteraceae: Gnaphalieae) at Mt Arthur in New Zealand. N. Z. J. Bot. 48, 75–81.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., RoyChoudhury, A., 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29, 1917–1932. doi:10.1093/molbev/mss086

Cariou, M., Duret, L., Charlat, S., 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecol. Evol. 3, 846–852. doi:10.1002/ece3.512

Chaudhary, R., Bansal, M.S., Wehe, A., Fernández-Baca, D., Eulenstein, O., 2010. iGTP: A software package for large-scale gene tree parsimony analysis. BMC Bioinformatics 11, 574. doi:10.1186/1471-2105-11-574

Chifman, J., Kubatko, L., 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics btu530. doi:10.1093/bioinformatics/btu530

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12,

499–510. doi:10.1038/nrg3012

Debry, R.W., Slade, N.A., 1985. Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. Syst. Biol. 34, 21–34. doi:10.1093/sysbio/34.1.21

Doyon, J.-P., Scornavacca, C., Gorbunov, K.Y., Szöllősi, G.J., Ranwez, V., Berry, V., 2010. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers, in: Tannier, E. (Ed.), Comparative Genomics, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 93–108.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A robust, simple Genotyping-By-Sequencing (GBS) approach for high diversity species. PLoS ONE 6, e19379. doi:10.1371/journal.pone.0019379

Faith, D.P., 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61, 1–10. doi:10.1016/0006-3207(92)91201-3

Farris, J.S., 1977. Phylogenetic analysis under Dollo's Law. Syst. Biol. 26, 77–88. doi:10.1093/sysbio/26.1.77

Farris, J.S., 1972. Estimating phylogenetic trees from distance matrices. Am. Nat. 106, 645–668.

Ford, K.A., Ward, J.M., Smissen, R.D., Wagstaff, S.J., Breitwieser, I., 2007. Phylogeny and biogeography of *Craspedia* (Asteraceae: Gnaphalieae) based on ITS, ETS and psbA-trnH sequence data. Taxon 56, 783–794.

Freudenstein, J.V., Pickett, K.M., Simmons, M.P., Wenzel, J.W., 2003. From basepairs to birdsongs: phylogenetic data in the age of genomics. Cladistics 19, 333–347. doi:10.1111/j.1096-0031.2003.tb00377.x

Gadagkar, S.R., Rosenberg, M.S., Kumar, S., 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. J. Exp. Zoolog. B Mol. Dev. Evol. 304B, 64–74. doi:10.1002/jez.b.21026

Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580. doi:10.1093/molbev/msp274

Huelsenbeck, J.P., Bollback, J.P., Levine, A.M., 2002. Inferring the root of a phylogenetic tree. Syst. Biol. 51, 32–43. doi:10.1080/106351502753475862

Kluge, A.G., Farris, J.S., 1969. Quantitative phyletics and the evolution of anurans. Syst. Biol. 18, 1–32. doi:10.1093/sysbio/18.1.1

Knowles, L.L., 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. Syst. Biol. 58, 463–467. doi:10.1093/sysbio/syp061

Lambert, S.M., Geneva, A.J., Luke Mahler, D., Glor, R.E., 2013. Using genomic data to revisit an early example of reproductive character displacement in Haitian *Anolis* lizards. Mol. Ecol. 22, 3981–3995. doi:10.1111/mec.12292

Laval, G., Excoffier, L., 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20, 2485–2487. doi:10.1093/bioinformatics/bth264

Lemmon, E.M., Lemmon, A.R., 2013. High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44, 99–121. doi:10.1146/annurev-ecolsys-110512-135822

Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24, 2542–2543. doi:10.1093/bioinformatics/btn484

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S., Costich, D.E., 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. PLoS Genet 9, e1003215. doi:10.1371/journal.pgen.1003215

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536.

Maddison, W.P., Donoghue, M.J., Maddison, D.R., 1984. Outgroup analysis and parsimony. Syst. Biol. 33, 83–103. doi:10.1093/sysbio/33.1.83

Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.

Maddison, W.P., Maddison, D.R., 2011. Mesquite: a modular system for evolutionary analysis.

Maio, N.D., Schrempf, D., Kosiol, C., 2015. PoMo: an allele frequency-based approach for species tree estimation. Syst. Biol. 64, 1018–1031. doi:10.1093/sysbio/syv048

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol., Morris Goodman Memorial Symposium 66, 526–538. doi:10.1016/j.ympev.2011.12.007

McVay, J.D., Carstens, B.C., 2013. Phylogenetic model choice: justifying a species tree or concatenation analysis. J. Phylogenetics Evol. Biol. doi:10.4172/2329-9002.1000114

Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31, i44–i52. doi:10.1093/bioinformatics/btv234

Misof, B., Meyer, B., von Reumont, B.M., Kück, P., Misof, K., Meusemann, K., 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinformatics 14, 348. doi:10.1186/1471-2105-14-348

Nicotra, A.B., Chong, C., Bragg, J.G., Chong, R.O., Aitken, N., Chuah, A., Lepschi, B.J., Borevitz, J.O., 2016. Population and phylogenomic decomposition via Genotyping-By-Sequencing in Australian *Pelargonium*. Mol. Ecol. 25, 2000–2014. doi:10.1111/mec.13584

Page, R.D.M., Charleston, M.A., 1997. Reconciled trees and incongruent gene and species trees, in: Mirkin, B.G. (Ed.), Mathematical Hierarchies and Biology: DIMACS Workshop, November 13-15, 1996, DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, Rhode Island, USA, pp. 57–70.

Python Software Foundation, 2016. Python Language Reference.

Raj, A., Stephens, M., Pritchard, J.K., 2014. Variational Inference of Population Structure in Large SNP Datasets. Genetics genetics.114.164350. doi:10.1534/genetics.114.164350

Rambaut, A., Suchard, M., Drummond, A.J., 2013. Tracer - MCMC Trace Analysis Tool.

Rheindt, F.E., Fujita, M.K., Wilton, P.R., Edwards, S.V., 2014. Introgression and phenotypic assimilation in *Zimmerius* flycatchers (Tyrannidae): Population genetic and phylogenetic inferences from genome-wide snps. Syst. Biol. 63, 134–152. doi:10.1093/sysbio/syt070

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.

RoyChoudhury, A., Felsenstein, J., Thompson, E.A., 2008. A two-stage pruning algorithm for likelihood computation for a population tree. Genetics 180, 1095–1105. doi:10.1534/genetics.107.085753

Rozefelds, A.C., 2002. A new species and new combination in *Craspedia* (Asteraceae) from Tasmania. Telopea 9, 813–819.

Rozefelds, A.C., Buchanan, A.M., Ford, K.A., 2011. New species of *Craspedia* (Asteraceae: Gnaphalieae) from Tasmania and determination of the identity of *C. macrocephala* Hook. Kanunnah 4, 93–116.

Sanderson, M.J., McMahon, M.M., 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol. Biol. 7, S3. doi:10.1186/1471-2148-7-S1-S3

Schmidt-Lebuhn, A.N., 2013. Reciprocal monophyly of *Craspedia* and *Pycnosorus* (Asteraceae, Gnaphalieae) and the problems of using ribosomal DNA at the lowest taxonomic levels. Aust. Syst. Bot. 26, 233–237.

Schmidt-Lebuhn, A.N., Milner, K.V., 2013. A quantitative study of morphology in Australian *Craspedia* (Asteraceae, Gnaphalieae). Aust. Syst. Bot. 26, 238–254.

Shaw, T.I., Ruan, Z., Glenn, T.C., Liu, L., 2013. STRAW: Species TRee Analysis Web server. Nucleic Acids Res. doi:10.1093/nar/gkt377

Soria-Carrasco, V., Talavera, G., Igea, J., Castresana, J., 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23, 2954–2956. doi:10.1093/bioinformatics/btm466

Swofford, D.L., 2003. PAUP*: phylogenetic analysis using parsimony, version 4.0 b10. Sinauer Associates, Sunderland.

Szöllősi, G.J., Tannier, E., Daubin, V., Boussau, B., 2015. The inference of gene trees with species trees. Syst. Biol. 64, e42–e62. doi:10.1093/sysbio/syu048

Tonini, J., Moore, A., Stern, D., Shcheglovitova, M., Ortí, G., 2015. Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. PLOS Curr. Tree Life. doi:10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. van de, Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., Zabeau, M., 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 23, 4407–4414. doi:10.1093/nar/23.21.4407

Wilkinson, M., McInerney, J.O., Hirt, R.P., Foster, P.G., Embley, T.M., 2007. Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends Ecol. Evol. 22, 114–115. doi:10.1016/j.tree.2007.01.002

Yoder, J.B., Briskine, R., Mudge, J., Farmer, A., Paape, T., Steele, K., Weiblen, G.D., Bharti, A.K., Zhou, P.,

May, G.D., Young, N.D., Tiffin, P., 2013. Phylogenetic signal variation in the genomes of *Medicago* (Fabaceae). Syst. Biol. syt009. doi:10.1093/sysbio/syt009

Figures

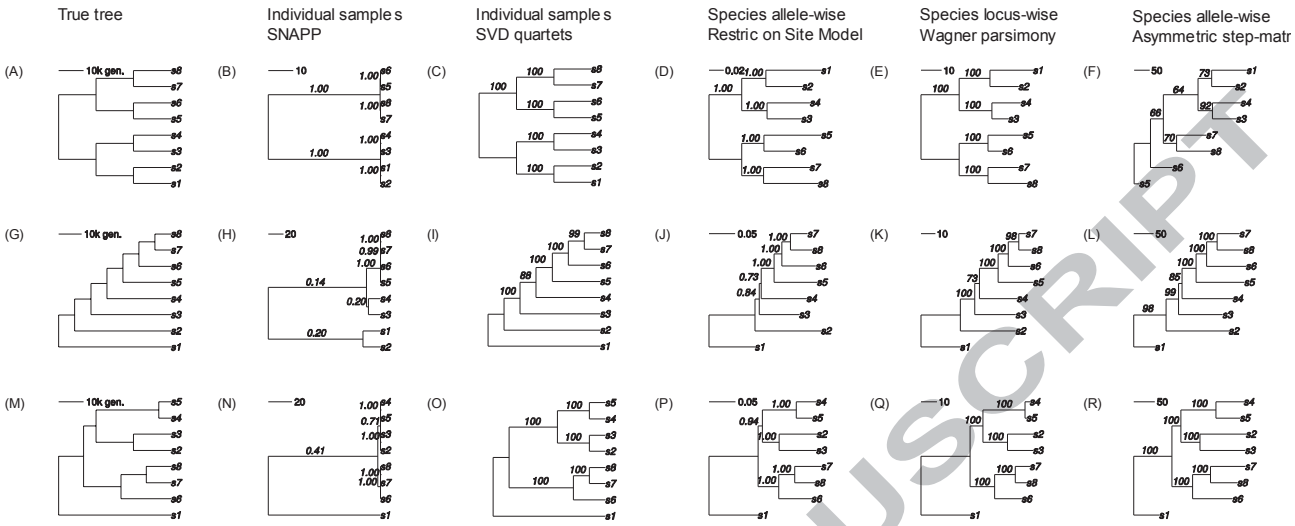
Fig. 1. Results of phylogenetic analyses of 500 SNPs simulated for eight species at three samples per species. (A), (G) and (M) true trees, (B), (H) and (N) results of Bayesian coalescent analyses using SNAPP, (C), (I) and (O) SVD quartet analyses, (D), (J) and (P) Bayesian analysis in MrBayes of SNPs summarized allele-wise at the species level, (E), (K) and (Q) parsimony analyses treating SNPs scored locus-wise as Wagner characters, and (F), (L) and (R) results of parsimony analyses with SNPs scored allele-wise and using an asymmetric step-matrix. SVD quartet and Wagner parsimony trees were rooted on the known outgroups. Numbers above branches of Bayesian summary trees indicate Posterior Probability scores, numbers above branches of other trees indicate Bootstrap support if above 50%.

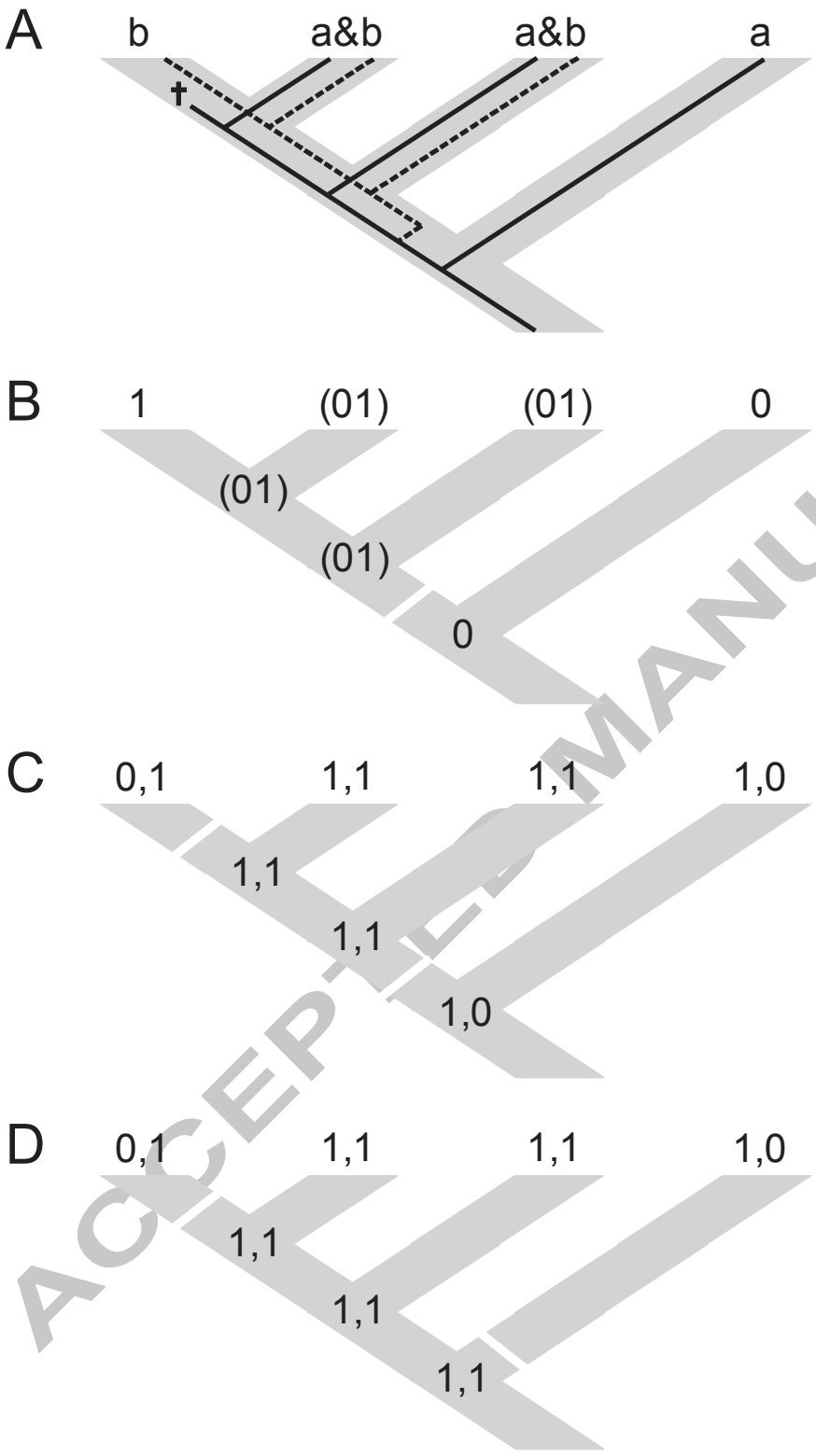
Fig. 2. (A) Hypothetical evolution of one SNP locus in a species phylogeny. After divergence of the first species, the second allele B arises through mutation and is inherited by all other species while one of them loses the original allele A. (B) Locus-wise scoring and reconstruction of ancestral states with Wagner characters and multiple states treated as polymorphic. (C) Allele-wise scoring and reconstruction of ancestral states under Wagner parsimony. (D) Allele-wise scoring and reconstruction of ancestral states under Dollo parsimony. White ticks indicate character changes as inferred by PAUP. Note that in (B) the gain of an allele is counted as a change, but a loss is not.

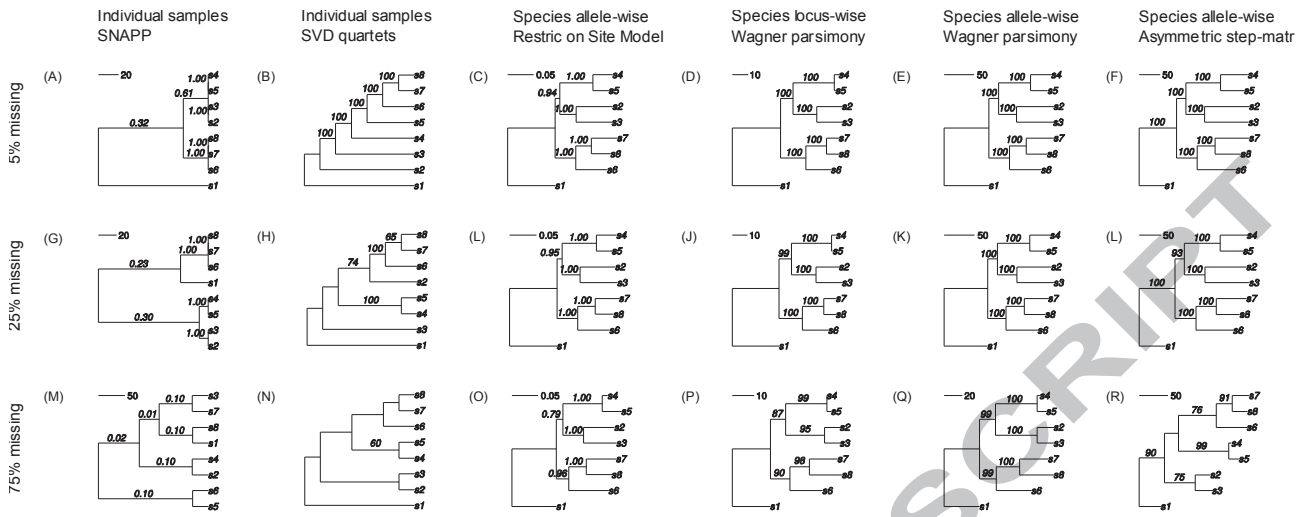
Fig. 3. Results of phylogenetic analysis of 500 simulated SNPs with 5% (A-E), 25% (F-J) and 75% (K-O) data missing at the level of the genotyped individual. (A), (G) and (M) coalescent analysis of individual sample data in SNAPP, (B), (H) and (N) SVD quartet analysis of individual samples, (C), (I) and (O) analysis in MrBayes of data summarized allele-wise at the species level, (D), (J), and (P) Wagner parsimony analysis of data scored locus-wise, (E), (K) and (Q) Wagner parsimony analysis of data scored allele-wise, and (F), (L) and (R) parsimony analysis of data scored allele-wise using an asymmetric step-matrix. Numbers above branches of Bayesian summary trees indicate Posterior Probability scores, numbers above branches of other trees indicate Bootstrap support if above 50%. SVD quartet trees and Wagner parsimony trees were rooted on s1. The true tree is in all cases that of Fig. 1m.

Fig. 4. Results of phylogenetic analyses of a SNP dataset for Australian *Pelargonium* (Nicotra et al., 2016). (A) SNAPP analysis as presented by Nicotra et al., (B) SVD quartet analysis of individual samples, (C) Bayesian analysis of data summarised allele-wise at the species level, (D) one of four equally parsimonious trees for SNP data scored locus-wise and treated as Wagner characters, with dashed lines indicating branches collapsing in the strict consensus tree, (E) single most parsimonious tree for data scored allele-wise and treated as Dollo characters, and (F) single most parsimonious tree for data scored allele-wise under an asymmetric step-matrix. Parsimony trees were rooted on *P. littorale* and *P. havlasae* if possible following the results from SNAPP. Numbers above branches in (A) and (C) indicate Posterior Probability scores, otherwise Bootstrap support if above 50%.

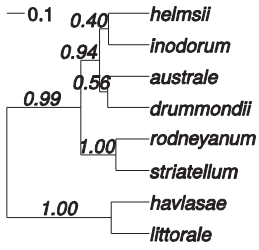
Fig. 5. Results of phylogenetic analyses of SNP data for 240 specimens of *Craspedia*. (A) SVD quartet analysis of individual samples, (B) Bayesian analysis of data summarised allele-wise at the species level, (C) single most parsimonious tree for data scored allele-wise and treated as Dollo characters, and (D) single most parsimonious tree for data scored as absent or present and treated as Dollo characters. Numbers above branches indicate Bootstrap support above 50 or Bayesian Posterior Probability above 95%.



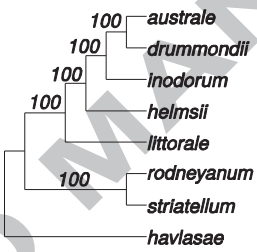




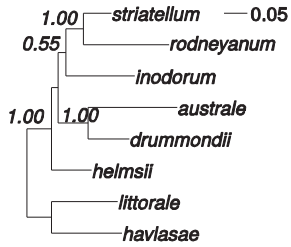
Individual samples SNAPP



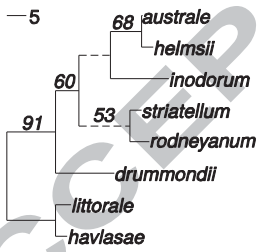
Individual samples SVD quartets



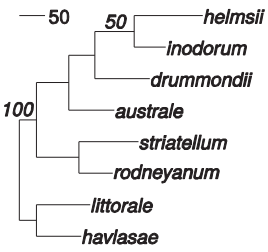
Species allele-wise Restric on Site Model



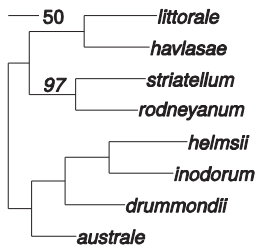
Species locus-wise Wagner parsimony



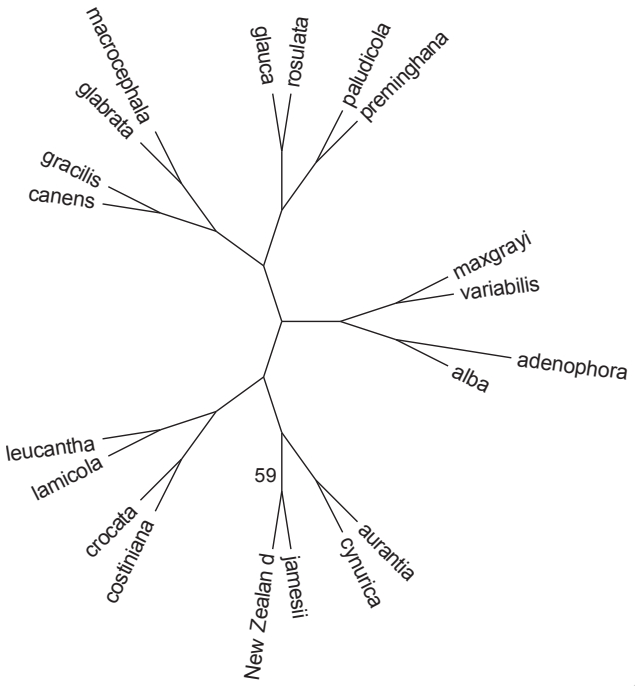
Species allele-wise Dollo parsimony



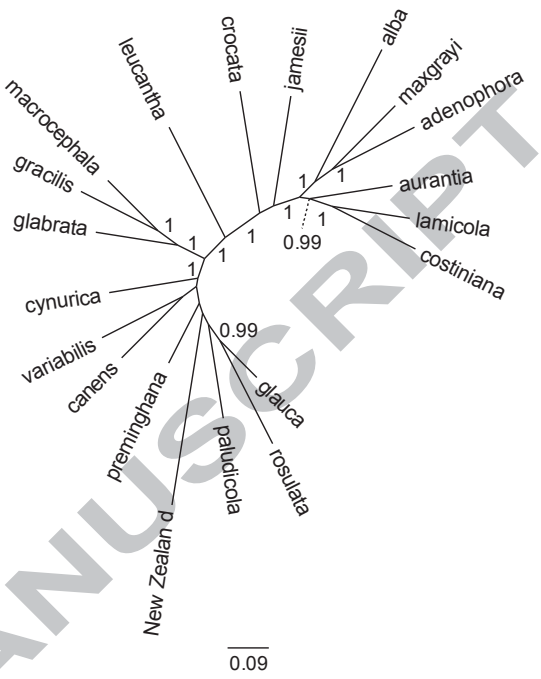
Species allele-wise Asymmetric step-matri



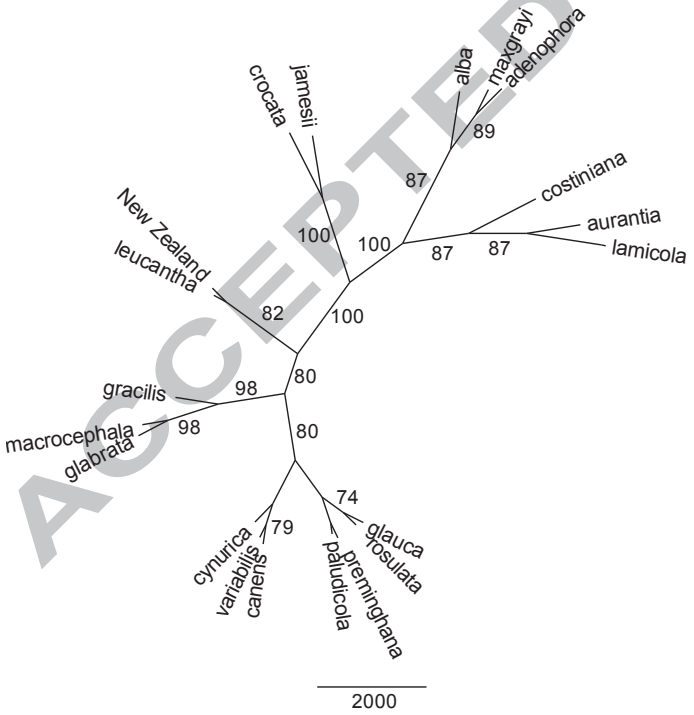
(A) Individual samples, SVD quartets



(B) Species allele-wise, Restriction Site Model



(C) Species allele-wise, Dollo parsimony



(D) Species present / absent, Doll o parsimony

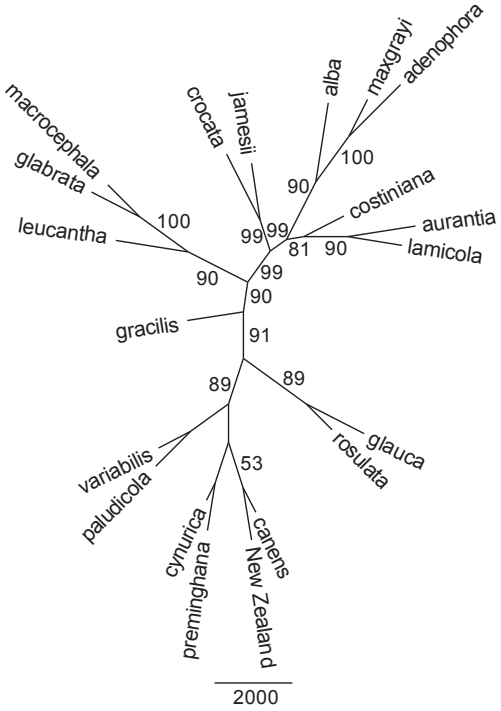


Table 1

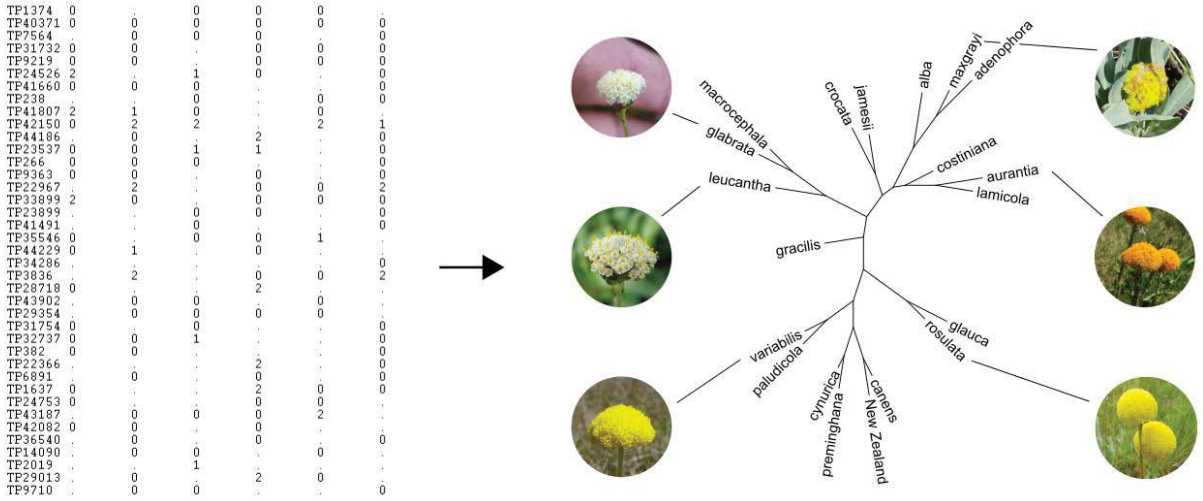
Summary of the results. Numbers in brackets are the K tree distances between the result and a reference tree, with lower numbers indicating a better fit of topology and branch length distribution. For simulated data, the reference is the tree on which the SNPs were simulated. For *Pelargonium*, the reference is the phylogeny produced by SNAPP. Where multiple equally parsimonious trees were recovered from search, only one was chosen to calculate K tree distances.

Dataset	**Purpose**	**Bayesian**			**Likelihood**	**Parsimony**			
		Individuals SNAPP	**Locus-wise MrBayes**	**Allele-wise MrBayes**	**Individuals SVD quartets**	**Locus-wise Wagner**	**Allele-wise Wagner**	**Allele-wise Dollo**	**Allele-wise step-matrix**
Simulated on balanced tree	Is correct topology inferred?	Yes, but odd branch lengths (51.90)	Yes (40.46)	Yes (17.69)	Yes	Yes (13.93)	Yes (17.66)	Yes (17.66)	Yes, but wrong rooting (18.89)
Simulated on unbalanced tree	Is correct topology inferred?	Failed to converge	Yes (64.93)	Yes (17.47)	Yes	Yes (11.74)	Yes (18.67)	Yes (18.54)	Yes (30.62)
Simulated on mixed tree	Is correct topology inferred?	Yes, but odd branch lengths (47.63)	Yes (42.56)	Yes (15.20)	Yes	Yes (8.36)	Yes (16.98)	Yes (16.71)	Yes (30.75)
Simulated on mixed tree, 5% missing	Is correct topology inferred?	Failed to converge	Yes (42.64)	Yes (15.06)	No	Yes (8.71)	Yes (16.74)	Yes (16.55)	Yes (30.56)
Simulated on mixed tree, 25% missing	Is correct topology inferred?	Yes, but wrong rooting (70.93)	Yes (42.78)	Yes (14.29)	No	Yes (8.59)	Yes (15.50)	Yes (15.33)	Yes (29.06)
Simulated on mixed tree, 75% missing	Is correct topology inferred?	Unresolved (71.74)	Yes (40.77)	Yes (15.26)	No	Yes (20.73)	Yes (23.14)	Yes (23.28)	No, but correct rooting (30.62)
Empirical *Pelargonium*	Are results consistent with relationships strongly supported	(Supplied by Nicotra et al.)	Yes[1] (0.82)	Yes[1] (0.85)	No	No (0.73)	No (0.86)	Yes[1] (0.87)	Yes[1], but different rooting (0.82)

	by SNAPP?								
Empirical *Craspedia* scored for present or absent data	Are results consistent with those from scoring for SNP value?	Computationally not feasible and/or crashing	No	No	Not applicable	No	Not applicable	Mostly yes but placement of New Zealand lineage differs	Mostly yes but placement of New Zealand lineage differs

[1]) Except for relationships that did not have significant support in the reference analysis and can thus be considered ambiguous.

Graphical abstract



Highlights

- We compare three approaches to summarising SNP data and inferring species trees.
- Available coalescent approaches struggle with large amounts of missing data.
- Bayesian inference and Dollo parsimony on allele data approximate SNAPP results.
- Scoring SNP data as present/absent is a potential alternative when data are patchy.