Constructing Phylogenies in the Presence Of Intra-Individual Site Polymorphisms (2ISPs) with a Focus on the Nuclear Ribosomal Cistron

ALASTAIR J. POTTS1,*, TERRY A. HEDDERSON1, and GUIDO W. GRIMM2

1Bolus Herbarium, Department of Biological Sciences, University of Cape Town, Cape Town, Western Cape, 7700, South Africa; and 2Department of Palaeobiology, Swedish Museum of Natural History, P.O. Box 50007, 104 05 Stockholm, Sweden

*Correspondence to be sent to: Alastair J. Potts, Botany Department, Nelson Mandela Metropolitan University, PO Box 77000, Port Elizabeth, 6031, South Africa; E-mail: potts.a@gmail.com.

Received 31 May 2013; revised returned 19 July 2013; accepted 2 August 2013

Abstract—Nuclear DNA is widely used to estimate phylogenetic and phylogeographic relationships. Nuclear gene variants may be present in an individual's genome, and these result in Intra-Individual Site Polymorphisms (2ISP; pronounced "twisp") in direct-PCR or individual-consensus sequences based on a sample of clones or fragment sequences from next generation sequencing (NGS). 2ISPs can occur fairly often, especially within, but not restricted to, high-copy-number regions such as the widely used internal transcribed spacers of the nuclear ribosomal cistron. Dealing with 2ISPs has been problematic as phylogeny reconstruction optimality criteria generally do not take account of this variation. Here we test whether an approach that treats 2ISPs as additional (termed "informative"), rather than ambiguous, characters offers improved support in three common criteria used for phylogenetic inference: Minimum Evolution (via Neighbour Joining), Maximum Parsimony, and Maximum Likelihood. We demonstrate significant improvements using the 2ISP-informative treatment with simulated, real-world, and case-study data sets. We envisage that this 2ISP-informative approach will greatly aid phylogenetic inference using any nuclear DNA regions that contain polymorphisms within individuals (including consensus sequences generated from NGS), especially at the intragenic or intraspecific level. [Intra-individual site polymorphism; nuclear DNA; phylogeny reconstruction; phylogenetic networks.]

For most organisms, biparentally inherited nuclear DNA regions offer crucial contrasts to the generally uniparentally inherited chloroplast or mitochondrial genomes. In plants, nuclear regions are widely used for phylogenetic inference and phylogeography (Chiang and Schaal 1999; Feliner et al. 2004; Rosselló et al. 2003; Rosselló et al. 2007; King and Roalson 2008). However, as a result of several processes, including incomplete concerted evolution, recombination, introgression and hybridization, and autopolyplodyization, multicopy nuclear regions often contain intra-individual polymorphisms (Bailey et al. 1995; Whittall et al. 2000; Rosselló et al. 2007; Álvarez and Wendel 2003). Such polymorphisms may provide additional signal for phylogenetic inference, but most algorithms and software treat polymorphic sites within sequences as ambiguous or missing characters. This usually hampers phylogeny reconstruction.

Sites showing multiple peaks within an individual are common in direct-PCR sequences (Baldwin et al. 1995; Álvarez and Wendel 2003). Such sites have been termed superimposed nucleotide additive patterns (SNAP; Whittall et al. 2000) or additive polymorphic sites (APS; Aguilar and Feliner 2003). Both of these terms are conditional on sampling, which is often incomplete, and neither encompasses the full range of polymorphism that can occur at a site. The definition of SNAP is restricted to parsimony-informative nucleotide polymorphisms, excluding indels although these are detectable and often informative (e.g., Baldwin et al. 1995; Whittall et al. 2000; Rosselló et al. 2007; Saurabh et al. 2012). The APS definition refers specifically to polymorphisms where the two bases are found separately in other individuals, ignoring those where only one base is found elsewhere in the data set. Here we define the Intra-Individual Site Polymorphism (2ISP; pronounced "twisp") as any polymorphic site (including indels) within an individual. In general, 2ISPs can be identified by direct-PCR or by generating consensus sequences from cloning or next generation sequencing, and can reflect intragenomic (inter-loci) or, in the case of nuclear-encoded ribosomal DNA, inter- and intra-array heterogeneity within an individual (elaborated below).

Most phylogenetic software (BEAST, Drummond and Rambaut 2007; distance module in MESQUITE, Maddison and Maddison 2010; MrBayes, Ronquist, and Huelsenbeck 2003; PAUP*, Swofford 2002) treats IUPAC codes of DNA base combinations (e.g., "Y" representing "C" and "T") as ambiguous characters or missing data. Depending on the optimality criterion and settings, IUPAC codes are, by default, either averaged (e.g., PAUP*) or ignored (e.g., BEAST), resulting in a lowered ability to discriminate among topologies or reduced branch support. In contrast, RAxML (Stamatakis 2006) treats all IUPAC codes as polymorphisms because the probability of substituting an "A" by "Y" equals that of "A" by "C" and/or "T". Nonetheless, this can still flatten the likelihood surface, making it more difficult to determine the best-known tree and reducing support values (Felsenstein 2004). Thus, polymorphism-aware treatments of IUPAC codes may have profound implications for
phylogeny reconstruction, especially in data sets where intra-individual variability exceeds intraspecific or intrageneric variability.

The internal transcribed spacers (ITS-1 and ITS-2) of the nuclear-encoded ribosomal 18S-5.8S-25S rDNA citron (35S rDNA) offer an excellent example of the problems associated with multicopy regions and intra-individual variation. These are the most widely used nuclear regions for plant phylogeny construction (see online Appendix S1, available from Dryad data repository; doi:10.5061/dryad.21j2, for a summary of GenBank accessions for commonly used plant DNA regions), but often contain 2ISPs (Alvarez and Wendel 2003; Feliner and Rossello 2007). The 35S rDNA citrons form a multigene family arranged in tandem arrays that are confined to one or several chromosomal loci, the nucleolus organizer region(s) (NOR; reviewed in Volkov et al. 1999), each comprising hundreds to thousands of copies (Rogers and Bendich 1987; Hemleben et al. 1988).

All rDNA copies within an individual were long thought to be virtually identical due to concerted evolution (reviewed in Nei and Rooney 2005), which permanently homogenizes gene duplicates via a combination of unequal crossing over and gene conversion (Zimmer et al. 1980; Elder and Turner 1995). However, concerted evolution in 35S rDNA is rendered inefficient by high copy numbers (leading to visible heterogeneity; Elder and Turner 1995; reviewed in Volkov et al. 2007) and multiple NORs (giving rise to inter-array heterogeneity; Schlötterer and Tautz 1994; Copenhaver and Pikaard 1996). Thus, an individual may contain multiple ITS variants. Analogous situations occur with the external transcribed spacers (ETS) of the 35S rDNA and the 5S intergenic spacers (5S-IGS). There are several mechanisms that can give rise to 2ISPs (Arneheim et al. 1980; reviewed in Volkov et al. 1999; Pikaard 2001): 1) Incomplete concerted evolution, where rates of mutation and recombination outstrip the homogenization process, 2) duplication and translocation of the NOR, producing nonorthologous (paralogous) NORs, 3) autopolyplody, 4) hybridization, often connected with allopolyploidy in plants, resulting in homoeology, 5) deficiencies in gene repair mechanisms (retention of pseudogene copies within an array), and 6) intragenomic competition between NORs, with potential loss of functionality and elimination. Depending on the source (e.g., incomplete concerted evolution through generations versus hybridization) 2ISPs may provide additional information for phylogenetic inference.

Whilst 2ISPs may contain valuable genealogical information (e.g., Fama et al. 2000; Feliner et al. 2004), the only means by which it has been utilized is via distance-based phylogenetic trees or networks (Joly and Bruneau 2006; Göker and Grimm 2008) or complex recoding (Göker et al. 2007a). Here we introduce a straightforward approach that, under parsimony or distance criteria, uses a simple step matrix to define additional steps for shifts between monomorphic and polymorphic sites (Fig. 1). Under maximum likelihood, monomorphic and polymorphic sites are treated as unique characters. This may seem overly simplistic as the exact mechanisms producing 2ISPs are largely unexplored (Elder and Turner 1995). Also, the degree of homogeneity within individuals and populations, that is, the fixation of the mutation in the gene pool, may be influenced by intra- and interarray competition (particularly in polyploids; Chen and Pikaard 1997; Volkov et al. 1999; Komarova et al. 2004) and historical demography (de Sousa Queiroz et al. 2011). With many possible conflicting processes of unknown relative importance we suggest that, as a starting point, 2ISPs and the gain or loss of within-individual variation be treated as simple substitution events, following logic similar to that used in polymorphism parsimony (Inger 1967; Farris 1978; Felsenstein 1979), which treats the retention of polymorphism as a step (illustrated in Felsenstein 2004, p. 76).

We demonstrate that treating 2ISPs as “informative” characters indeed provides additional information for inferring evolutionary relationships. We use simulations to determine the phylogenetic outcome when the source of 2ISPs is via inheritance (2ISPs inherited from polymorphic parent/s). We then investigate phylogenetic outcomes using real-world data sets under the 2ISP-informative versus the traditional 2ISP-ambiguous treatments. Lastly, we present three detailed case studies to establish the effects of hybrids and to highlight the limitations and advantages of our approach.

MATERIALS AND METHODS

Given the wide range of software and analyses used in this study, we provide a relatively brief overview of the methods below and a detailed description of methods in the online Appendix S2 (available from Dryad data repository; doi:10.5061/dryad.21j2/2).
TREATING 2ISPs AS INFORMATIVE

For parsimony and distance-based reconstruction we treat the transitions from a monomorphic state to another via a polymorphic state as ordered substitution events (i.e., C→Y→T) accommodated in a step matrix with 15 states (or more if indel coding is included; Fig. 1). For distance algorithms, the step matrix is used to calculate a modified uncorrected p-distance incorporating polymorphisms (hereafter termed polymorphism p-distance; reasons for not using model-based distances are given in online Appendix S2, available from Dryad data repository; doi:10.5061/dryad.21jj2).

Our parsimony and distance-based approaches are naïve in the sense that probabilities of fixation or loss of a 2ISP are likely to differ. However, this can be accommodated in a maximum likelihood (ML) framework. Lacking a general transition model that allows simultaneous estimation of substitution and homogenization probabilities, we utilize an ad hoc 2ISP-informative implementation in the ML approach by simply treating each IUPAC code (i.e., monomorphic and polymorphic bases) as a unique character using the multistate analysis implementation of categorical data in RAxML. Transition rates between characters are thus based on the signal in the underlying data and essentially free from further constraints in contrast to the parsimony and distance-based treatments.

SIMULATIONS: INHERITED VARIATION

To simulate inherited polymorphism induced by two sets of independently evolving sequences, a random tree was generated, two DNA data sets (A1 and A2) were simulated onto this tree (across a range of substitution rates), and the tip sequences from these two data sets were merged to form a polymorphic, combined variant, data set (A1&2). This process was repeated 200 times for replication and is further elaborated in the online Appendix S2, available from Dryad data repository; doi:10.5061/dryad.21jj2. The A1 and A2 DNA data sets can be considered independent histories of two codominant variants present in the gene pool of a lineage. In the case of ITS (or ETS or 5S-IGS data) they could either represent intra-array (paralogy in a broad sense), allelic (between homologous NORs), homoeologous (between orthologous NORs/5S loci) or paralogous s.s. variation (between NORs/5S loci originated by duplication and translocation). The A1&2 data sets represent the substitution and inheritance of polymorphisms along branches and branching events. This is a highly simplified model of multicopy evolution and inheritance as it does not include the effects of concerted or reticulate evolution.

Phylogenetic trees were inferred using the neighbor-joining algorithm (NJ) implemented in PAUP* version 4b10 and the APE library version 2.71 (Paradis et al. 2004)—with polymorphism p distances calculated using the PHANGORN library version 1.99 (Schliep 2011)—in R for the 2ISP-ambiguous (NJ-A) and 2ISP-informative (NJ-I) treatments, respectively, to find topologies that fulfill the minimum evolution (ME) criterion. Maximum parsimony (MP) was implemented in PAUP*. 2ISPs were treated as either ambiguities (MP-A), the default treatment, or as informative (MP-I) using the cost matrix shown in Figure 1. RAxML 7.2.6 was used to compute trees and perform bootstrap analyses under ML (Stamatakis 2006). Whilst a standard RAxML analysis includes polymorphic bases, this can still lead to flattening of the likelihood surface, making it more difficult to determine the best-known tree and reducing support values. Thus, the standard analysis was treated as the analogue to the 2ISP-ambiguous treatment under NJ and MP (ML-A). RAxML includes a multistate analysis for any kind of categorical data (-m MULTIGAMMA -k GTR); this would treat each IUPAC code as a unique character, thereby estimating the rate of transitions between characters (i.e., estimating the rate of transitions between states instead of the single steps shown in Fig. 1). This was considered the 2ISP-informative treatment (ML-I). Under all three optimality criteria (ME, MP, ML), branch support was assessed using nonparametric bootstrapping (Felsenstein 1985).

Bootstrap bipartition support was used to evaluate the phylogenetic outcomes between the 2ISP-ambiguous and 2ISP-informative treatments. Bipartitions were paired between the ambiguous and informative treatments for each method per data set. In order to avoid the majority of bipartition pairs with only low support values, which are not of direct interest, a threshold of 70% was used to eliminate any bipartition pairs where both support values were lower than the threshold. The number of bipartitions supported above the threshold for each simulated data set (A1, A2, and A1&2) was calculated and converted to a percentage of the supportable bipartitions present in the true tree (number of tips minus two in all cases, therefore 18). The nonparametric Mann-Whitney signed rank test (Hollander and Wolfe 1973, p. 68–75) was used (as the distributions of the bootstrap bipartition results were often severely skewed) to determine whether there were overall significant differences in the percentage of supported bipartitions between data sets (A1, A2, and A1&2) within each algorithm and treatment combination (e.g., NJ-A). The test was two-sided and significance determined at α < 0.05. To investigate the occurrence of false positives (i.e., supported bipartitions that are not found on the true tree), the bipartitions supported above the 70% threshold were also compared with the tree used to simulate the data. The false positive rate was calculated as the proportion of incorrectly supported bipartitions to the sum of all supported bipartitions.

Published Data Sets

Performances of the 2ISP-informative and 2ISP-ambiguous treatments were compared using 21 previously published DNA alignments and a novel data
set generated for this study. The data sets include 2ISPs found in direct-sequenced PCR products as well as strict individual-consensus sequences based on a collection of clones. As 2ISPs occur on many other nuclear gene regions, data sets were not limited to ITS. Aligned sequences were either obtained directly from the authors or downloaded from TreeBase (www.treebase.org). If indels were coded as binary characters in the data set, then these codings were kept for subsequent analyses. Three data sets were also treated as case studies to explore the advantages and limitations of the 2ISP-informative approach, and are discussed in the next section. All data sets were analyzed using NJ, MP, and ML using the same software employed for the simulation studies (see online Appendix S2, available from Dryad data repository; doi:10.5061/dryad.21i2j) for set-up details and processing of results.

We compared the difference of support values between the 2ISP-ambiguous versus 2ISP-informative treatments (Δ support) for each algorithm against the information content of standard DNA and 2ISP characters within the data sets. The Δ support values were calculated as the difference in (equivalent) bipartition bootstrap support between the 2ISP-informative and 2ISP-ambiguous treatments where at least one of the pairs received support above a low, moderate, or high bootstrap support threshold (>50%, >70%, or >90%, respectively). The thresholds were used to avoid calculating Δ support values for bipartition pairs that only received weak support (<50%); such pairs often dominated the pairwise bipartition comparisons. The information content within each data set was characterized using a parsimony-informative sites index (P-index). The P-index was calculated as follows:

\[
P_{\text{2ISP}} - P_{\text{std}}
\]

\[
P_{\text{2ISP}} + P_{\text{std}}
\]

where \( P_{\text{2ISP}} \) and \( P_{\text{std}} \) are the number of parsimony-informative sites for 2ISP and standard DNA characters, respectively. The P-index ranges from −1 to 1, where all parsimony-informative sites are either exclusively standard or 2ISP DNA characters, respectively. To detect potentially incompatible or ambiguous signals in the data sets, such as those caused by hybridization or allopolyploidization, we inferred neighbor-net splits graphs (Bryant and Moulton 2004) and consensus networks (Holland and Moulton 2003) based on the bootstrap samples (“bootstrap networks,” e.g., Grimm et al. 2006) using SPLITSTREE version 4.8 (Huson and Bryant 2006) for a subset of published studies and the novel Nymania capensis data set generated in this study.

**Case studies: Acer, Hieracium, and Nymania**

We used data from three angiosperm groups, Acer sect. Acer (Sapindaceae; Grimm et al. 2007b; Göker and Grimm 2008), Hieracium L. s.l. (Asteraceae; Fehrer et al. 2009), and Nymania capensis (Chinese Lantern; Meliaceae; this study), that exhibited intra- and interindividual 2ISP variability. The Hieracium and Nymania data sets were generated from direct-PCR sequencing and 2ISPs were coded if they occurred in both reading directions and any secondary peak was greater than 30% the height of the primary peak in the trace file. In contrast, the Acer data set comprised strict individual-consensus sequences of ITS clones. The Hieracium and Acer data sets contain numerous putative hybrids (identified by 2ISPs and clone sequences, respectively) and were used to explore the effects of hybrids on phylogenetic support. The Hieracium data set contained 60 sequences from the S′ external transcribed spacer (ETS) of the 35S rDNA. The putative hybrids in the Hieracium data set were between two genetically and geographically divergent clades (Fehrer et al. 2009). The Acer data set contained ITS consensus sequences from 27 individuals, including five that showed evidence of reticulation/lineage crossing with signals that were analogous to F1 hybrids (Grimm et al. 2007b). The Nymania capensis data set comprised 30 individuals sampled across three primary drainage basins in the Albany Subtropical Thicket biome, which spans the Western and Eastern Cape Provinces of South Africa. Ten individuals were sampled per drainage basin. Two additional individuals from the disjunct northern distribution of the species were used as outgroups (BOL48535 and BOL60966). Full collection details for these individuals and details regarding DNA extraction, PCR, cloning, and sequence assembly are given in online Appendix S2 (available from Dryad data repository; doi:10.5061/dryad.21i2j). In brief, proofreading polymerase was used for PCR amplification of ITS.

**RESULTS**

**Simulations: Inherited Variation**

The A1, A2, and A1&2 simulated data sets obtained across different substitution rates are described in Table 1. No significant differences in percentage of true bipartitions supported above the selected threshold (70%) were detected between the A1 and A2 data sets, irrespective of 2ISP-treatment or substitution rate (Table 1). In contrast, in nearly all cases, the 2ISP-ambiguous approach recovers significantly (P < 0.05) fewer bootstrap-supported bipartitions when comparing A1&2 with A1 or A2, whereas the 2ISP-informative approach recovers significantly more. The only exception is under ML at the highest substitution rate where no differences could be demonstrated. The false positive rate (supported bipartitions that are not true bipartitions supported above the selected threshold) remained below 5% across all methods, treatments, and data sets (Table 1).

**Published Data Sets**

The published data sets represent real-world situations where multiple, sometimes conflicting,
Table 1. Summary statistics of data sets, overall bootstrap support and incorrectly supported bipartitions (false positives) for the inherited polymorphism simulations across a range of substitution rates.

<table>
<thead>
<tr>
<th>Substitution rate (substitutions per site)</th>
<th>0.010</th>
<th>0.025</th>
<th>0.050</th>
<th>0.100</th>
<th>0.200</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset statistics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PI-2ISP</td>
<td>24 [4]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Percentage of NJ-A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bootstrap-supported</td>
<td>24 (12-29)*</td>
<td>24 (18-31)*</td>
<td>0 (0-0)*</td>
<td>60 (71-88)*</td>
<td>59 (53-71)*</td>
</tr>
<tr>
<td>MP-A</td>
<td>18 (6-24)*</td>
<td>18 (12-28)*</td>
<td>0 (0-0)*</td>
<td>71 (65-76)*</td>
<td>71 (65-76)*</td>
</tr>
<tr>
<td>MP-I</td>
<td>18 (6-24)*</td>
<td>18 (12-28)*</td>
<td>0 (0-0)*</td>
<td>71 (65-76)*</td>
<td>71 (65-76)*</td>
</tr>
<tr>
<td>ML-A</td>
<td>47 (35-53)*</td>
<td>47 (35-53)*</td>
<td>0 (0-0)*</td>
<td>85 (82-94)*</td>
<td>85 (82-94)*</td>
</tr>
<tr>
<td>ML-I</td>
<td>47 (35-53)*</td>
<td>47 (35-53)*</td>
<td>0 (0-0)*</td>
<td>85 (82-94)*</td>
<td>85 (82-94)*</td>
</tr>
<tr>
<td><strong>False positives</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NJ-A</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>MP-A</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ML-A</td>
<td>0.09</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ML-I</td>
<td>0.09</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note: The GTR+ model was used for all ML analyses.

aVS-STD: standard variable sites; VS-2ISP: intra-individual site polymorphism variable sites; PI-STD: standard parsimony-informative sites; PI-2ISP: intra-individual site polymorphism parsimony-informative sites.

bThe median percentage (25th and 75th percentiles) of supported bipartitions (i.e., bipartitions with above a bootstrap support threshold of 70%) of the total number of supportable bipartitions (based on tree bifurcations related to the number of tips) across the simulated data sets for each method and 2ISP-treatment. Statistical tests used nonparametric Mann-Whitney signed rank test to compare the results from the A1, A2, and A1&2 data sets within the different method and 2ISP-treatment combinations.


cStatistical tests used nonparametric Mann-Whitney signed rank test to compare the results from the A1, A2, and A1&2 data sets within the different method and 2ISP-treatment combinations.

dFalse positives estimates are the proportion of supported bootstrap bipartitions that are not found on the tree used to simulate the data to all supported bipartitions (only bipartitions above the 70% support threshold are considered).

eFalse positives estimates are the proportion of supported bootstrap bipartitions that are not found on the tree used to simulate the data to all supported bipartitions (only bipartitions above the 70% support threshold are considered).

fDissimilar superscript daggers denote significant differences in the percentage of bootstrap-supported bipartitions between data sets within each method and substitution rate at α < 0.05.
processes lead to the presence of 2ISPs. The data sets were selected to represent a range of P-index values (i.e., 2ISPs rarely to commonly present as parsimony-informative characters; Table 2). The majority of the data sets were from the ITS region, but other nuclear regions are also included. Unsurprisingly, as the P-index increased, so did the relative number of samples that contained 2ISPs.

The magnitude and significance of Δ support (the difference between pairwise bipartitions inferred from the 2ISP-informative and 2ISP-ambiguous treatments) varied across thresholds, methods, and P-index values (Fig. 2). However, many data sets exhibit a significant increase in Δ support (better performance of the informative treatment) across the bootstrap threshold criteria (Fig. 2). In no case did the ambiguous treatment outperform the informative treatment (i.e., no Δ support tests were significantly less than zero, all $p > 0.05$), irrespective of method or threshold. The informative treatment tends to significantly increase Δ support as the P-index increases, although this did not apply to all high (greater than 0) P-index data sets.

The bipartition support summary shown in Figure 2 masks considerable data set-specific nuances (e.g., inflicted by hybrid samples). Thus, we also provide neighbour-net splits graphs (based on polymorphism $p$-distances) annotated with bipartition support from bootstrap networks for all methods and treatments from 11 data sets in online Appendix S1 (see Dryad data repository; doi:10.5061/dryad.21jj2). These data sets were selected to represent the full range of P-index values. Short synopses are provided for each data set in the online figure legends.

**Case Study 1: Hybrid taxa (clone-consensus data set including taxa of hybrid origin)**

The *Acer* data set contained 59 parsimony-informative sites comprising 36 standard DNA characters and 20 2ISP characters, with substantial overlap of sites with both standard and 2ISP characters in the alignment (Table 2). This data set comprises various distinct, mostly species-specific, ITS variants and includes a number of genetically, and partly morphologically, identified hybrids. The neighbour-net splits graph based on polymorphism $p$-distances (Fig. 3a) is very similar to the splits graph based on PBC host-associate distances (Göker and Grimm 2008, figure 9); as in the latter study, the (putative) hybrid specimens represented by clones diagnostic for two different parent lineages (*A. semprevirens* × *A. ikarianum, A. semprevirens* × *A. opalus, *A. heldreichii* × *A. pseudoplatanus; Grimm et al. 2007b) form pronounced box-like structures. Their hybrid status is corroborated by the fact that the best-supported phylogenetic splits associate them with their respective parent lineages, but (4) support is never $\geq$ 50 (Figs. 3a, S3.5). There are a few other noteworthy patterns that should be highlighted: 1) All deeply divergent lineages (long edge lengths) received moderate ($>70\%$) to high ($>90\%$) support for all -A and -I treatments, 2) in one group (*A. ibericum*–*A. hyrcanum*; considered conspecific by some authors), all -I treatments recover moderate to high support, whereas only NJ supports this split under -A treatment, and 3) two hybrid samples between *A. monspessulanum* and *A. opalus* are found to have moderate to high bootstrap support with *A. monspessulanum* for MP-A and ML-A, respectively; similar support is not observed in any of the -I treatments. This is an example of the ambiguous approach being misled by its treatment of 2ISPs to support a clade that contains hybrids with a distant lineage.

**Case study 2: Hybrid swarm (direct-PCR data set including taxa of hybrid origin)**

The *Hieracium* data set contained 91 parsimony-informative sites, 28 of which were standard DNA characters and 63 2ISP characters, with a large degree of overlap of standard and 2ISP parsimony-informative sites in the alignment (Table 2). Fehrer et al. (2009) identified 29 samples in the *Hieracium* data set as putative hybrids based on 2ISP patterns. Seventeen of these samples were centrally located in the neighbour-net splits graph (Fig. 3b) and represent a hybrid swarm as they contained 2ISP patterns shared between two genetically and largely geographically divergent clades (Fehrer et al. 2009). Two putative hybrids represent intraspecific crossing of subclades of the “eastern clade”. We highlight the following observations: 1) The “western” and “eastern” clades are not well supported by all the -A treatments (ML-A and MP-A $\leq 33\%$, NJ-A $= 79\%$/87\%), whereas all -I treatments have moderate to high support for these clades (all $\geq 80\%$), 2) no moderate to high support was found for -A treatments that was not also supported by -I treatments, 3) often the -I treatments provided greater support within the “western” and “eastern” clades than the corresponding -A treatments (e.g., *H. umbellatum* subclade), and 4) the -I treatment provides greater support for clades within the hybrid swarm.

**Case study 3: Inherited polymorphisms (Direct-PCR intraspecific data set)**

The *Nymania* data set contained 33 parsimony-informative sites; 12 of these included standard DNA characters whilst all 33 had 2ISP characters for at least some individuals (Table 2; Appendix A.1; see Dryad data repository; doi:10.5061/dryad.21jj2 for the full data set alignment; Genbank Accessions: KF443002-KF443033). Three 2ISP indels were observed, which were coded as simple substitutions for NJ and MP analyses or as additional binary states for ML analyses. Each 2ISP indel was confined to samples from a specific drainage basin. No substitutions were observed in four conserved motifs nor in the whole 5.8S region in the direct sequences,
## Table 2. Summary information for published data sets.

<table>
<thead>
<tr>
<th>P-index</th>
<th>N&lt;sup&gt;b&lt;/sup&gt;</th>
<th>chars</th>
<th>STD</th>
<th>ALL</th>
<th>2ISP</th>
<th>N&lt;sub&gt;2ISP&lt;/sub&gt;</th>
<th>Mean</th>
<th>Median</th>
<th>Min.</th>
<th>Max.</th>
<th>Taxa</th>
<th>Region</th>
<th>Source&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.97</td>
<td>27</td>
<td>659</td>
<td>305</td>
<td>308</td>
<td>5</td>
<td>6</td>
<td>4.5</td>
<td>0</td>
<td>0</td>
<td>74</td>
<td>Amaryllidaceae</td>
<td>ITS</td>
<td>1 and 2</td>
<td>(Meerow et al. 2006)</td>
</tr>
<tr>
<td>-0.94</td>
<td>25</td>
<td>627</td>
<td>103</td>
<td>104</td>
<td>3</td>
<td>9</td>
<td>3.0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>Erpetonopsis</td>
<td>ITS</td>
<td>1</td>
<td>(Allen et al. 2003)</td>
</tr>
<tr>
<td>-0.93</td>
<td>21</td>
<td>634</td>
<td>136</td>
<td>138</td>
<td>5</td>
<td>15</td>
<td>2.8</td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>Sambucus</td>
<td>ITS</td>
<td>1</td>
<td>(Eriksson and Donoghue 1997)</td>
</tr>
<tr>
<td>-0.69</td>
<td>58</td>
<td>1046</td>
<td>83</td>
<td>85</td>
<td>15</td>
<td>25</td>
<td>1.6</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>Pragia</td>
<td>PIRNC</td>
<td>1</td>
<td>(Jalaliy and Sytsma 2000)</td>
</tr>
<tr>
<td>-0.68</td>
<td>14</td>
<td>718</td>
<td>69</td>
<td>81</td>
<td>13</td>
<td>7</td>
<td>3.9</td>
<td>1</td>
<td>0</td>
<td>17</td>
<td>Krugia</td>
<td>ITS</td>
<td>1</td>
<td>(Kim and Jansen 1994)</td>
</tr>
<tr>
<td>-0.57</td>
<td>21</td>
<td>897</td>
<td>98</td>
<td>114</td>
<td>27</td>
<td>30</td>
<td>5.4</td>
<td>5</td>
<td>0</td>
<td>18</td>
<td>Tugilia</td>
<td>ITS</td>
<td>1</td>
<td>(Moore et al. 2006)</td>
</tr>
<tr>
<td>-0.42</td>
<td>31</td>
<td>213</td>
<td>68</td>
<td>90</td>
<td>26</td>
<td>26</td>
<td>4.0</td>
<td>3</td>
<td>0</td>
<td>14</td>
<td>Machoerantheminae</td>
<td>5S-IGS</td>
<td>2</td>
<td>(Morgan et al. 2009)</td>
</tr>
<tr>
<td>-0.27</td>
<td>37</td>
<td>676</td>
<td>151</td>
<td>193</td>
<td>86</td>
<td>29</td>
<td>12.2</td>
<td>2</td>
<td>0</td>
<td>53</td>
<td>Tugilia</td>
<td>WAXY</td>
<td>1</td>
<td>(Moore et al. 2006)</td>
</tr>
<tr>
<td>0.24</td>
<td>127</td>
<td>603</td>
<td>54</td>
<td>66</td>
<td>50</td>
<td>1</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>Ranunculus</td>
<td>ITS</td>
<td>2</td>
<td>(Joly et al. 2009)</td>
</tr>
<tr>
<td>0.04</td>
<td>37</td>
<td>1337</td>
<td>23</td>
<td>26</td>
<td>25</td>
<td>24</td>
<td>5.3</td>
<td>1</td>
<td>0</td>
<td>22</td>
<td>Rosa</td>
<td>ITS</td>
<td>2</td>
<td>(Joly et al. 2009)</td>
</tr>
<tr>
<td>0.09</td>
<td>37</td>
<td>757</td>
<td>36</td>
<td>30</td>
<td>19</td>
<td>24</td>
<td>2.2</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>Rosa</td>
<td>GAPDH</td>
<td>2</td>
<td>(Joly et al. 2009)</td>
</tr>
<tr>
<td>0.16</td>
<td>37</td>
<td>392</td>
<td>36</td>
<td>59</td>
<td>50</td>
<td>35</td>
<td>5.0</td>
<td>3</td>
<td>0</td>
<td>33</td>
<td>Acer</td>
<td>Acet</td>
<td>ITS</td>
<td>(Goller and Grimm 2008)</td>
</tr>
<tr>
<td>0.26</td>
<td>67</td>
<td>353</td>
<td>81</td>
<td>160</td>
<td>139</td>
<td>53</td>
<td>11.9</td>
<td>3</td>
<td>0</td>
<td>72</td>
<td>Platanus</td>
<td>ITS</td>
<td>2</td>
<td>(Grimm and Denk 2003)</td>
</tr>
<tr>
<td>0.29</td>
<td>224</td>
<td>496</td>
<td>134</td>
<td>248</td>
<td>243</td>
<td>176</td>
<td>11.8</td>
<td>10</td>
<td>0</td>
<td>59</td>
<td>Quercus</td>
<td>ITS</td>
<td>2</td>
<td>(Denk and Grimm 2011)</td>
</tr>
<tr>
<td>0.30</td>
<td>37</td>
<td>634</td>
<td>15</td>
<td>29</td>
<td>28</td>
<td>28</td>
<td>3.0</td>
<td>2</td>
<td>0</td>
<td>12</td>
<td>Rosa</td>
<td>TPI</td>
<td>2</td>
<td>(Joly and Bruneau 2006)</td>
</tr>
<tr>
<td>0.37</td>
<td>236</td>
<td>410</td>
<td>112</td>
<td>263</td>
<td>245</td>
<td>207</td>
<td>7.4</td>
<td>6</td>
<td>0</td>
<td>35</td>
<td>Quercus</td>
<td>5S-IGS</td>
<td>2</td>
<td>(Grimm and Denk 2011)</td>
</tr>
<tr>
<td>0.40</td>
<td>67</td>
<td>385</td>
<td>26</td>
<td>73</td>
<td>60</td>
<td>57</td>
<td>4.9</td>
<td>3</td>
<td>0</td>
<td>22</td>
<td>Platanus</td>
<td>5S-IGS</td>
<td>2</td>
<td>(Grimm and Denk 2011)</td>
</tr>
<tr>
<td>0.42</td>
<td>11</td>
<td>703</td>
<td>7</td>
<td>19</td>
<td>17</td>
<td>11</td>
<td>9.45</td>
<td>8</td>
<td>3</td>
<td>17</td>
<td>Zelkova</td>
<td>ITS</td>
<td>2</td>
<td>(Goller and Grimm 2008)</td>
</tr>
<tr>
<td>0.47</td>
<td>30</td>
<td>638</td>
<td>12</td>
<td>33</td>
<td>33</td>
<td>30</td>
<td>5.7</td>
<td>6</td>
<td>1</td>
<td>11</td>
<td>Nymania</td>
<td>ITS</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>0.53</td>
<td>60</td>
<td>521</td>
<td>28</td>
<td>91</td>
<td>90</td>
<td>60</td>
<td>11.4</td>
<td>7</td>
<td>1</td>
<td>33</td>
<td>Heteracanthus</td>
<td>ETS</td>
<td>1</td>
<td>(Febrer et al. 2009)</td>
</tr>
<tr>
<td>0.61</td>
<td>25</td>
<td>623</td>
<td>8</td>
<td>33</td>
<td>33</td>
<td>25</td>
<td>7.8</td>
<td>9</td>
<td>0</td>
<td>29</td>
<td>Lathyrus</td>
<td>ITS</td>
<td>2</td>
<td>(Schlee et al. 2011)</td>
</tr>
<tr>
<td>0.79</td>
<td>28</td>
<td>773</td>
<td>10</td>
<td>87</td>
<td>86</td>
<td>27</td>
<td>16.0</td>
<td>14</td>
<td>0</td>
<td>38</td>
<td>Fagus</td>
<td>ITS</td>
<td>2</td>
<td>(Goller and Grimm 2008)</td>
</tr>
</tbody>
</table>

Note: All summary statistics are calculated with outgroups removed.

<sup>a</sup>Calculated as (PI<sub>2ISP</sub> − PI<sub>STD</sub>)/(PI<sub>2ISP</sub> + PI<sub>STD</sub>). The P-index ranges from −1 to 1, where all parsimony informative sites are only standard or only 2ISP DNA characters, respectively.

<sup>b</sup>N: the number of individuals; Align. Chars.: the numbers of characters in the alignment; PI: the number of parsimony informative sites for standard characters (STD), all characters (ALL) and intra-individual site polymorphisms (2ISP) characters; N<sub>2ISP</sub>: the number of samples that contain 2ISPs, and the mean, median, minimum and maximum number of 2ISPs per sequence within each data set.

<sup>c</sup>1: direct-PCR sequences; 2: strict individual-consensus sequences of clones.
Figure 2. The P-index compared with Δ support from real world data sets. Δ support is the change in the bootstrap support of paired bipartitions between the 2ISPinformative and 2ISP-ambiguous treatments. This is calculated for paired bipartitions where support for at least one of the bipartitions is above a specified bootstrap support threshold (i.e., bipartitions with low support in both treatments are ignored; see online Appendix S2 for details). Three different bootstrap thresholds are considered: >50%, >70%, and >90%, respectively. Treatments of 2ISPs are compared within three different phylogenetic methods (NJ: Neighbour joining; MP: Maximum parsimony; ML: Maximum likelihood) and 2ISP-informative (-I) and 2ISP-ambiguous (-A) treatments. The median difference is shown (square symbol) with the 25th to 75th and 1st to 100th percentiles (thick grey lines and stippled grey lines, respectively). The significance of Δ support values for each data set was tested using a Wilcoxon signed rank test (except in instances where there were insufficient bipartition pairs: n < 5).

Therefore we considered the data set free of pseudogene ITS variants. In the neighbor-net splits graph, three distinct clusters were resolved, each corresponding directly to a drainage basin (Fig. 4a). The only exception is individual AJP0532 which was collected in the Gamtoos basin but is grouped with individuals from the Sundays basin; however, this collection locality was very close to the watershed boundary between the Gamtoos and the Sundays basins and, given its clustering within the Sundays, we subsequently treated it as a member of the Sundays population (this is supported by chloroplast sequence data, Potts et al. 2013a).
Figure 3. Neighbor-net splits graph for the (a) Acer and (b) Hieracium data sets based on polymorphism p-distances (P-index values of 0.16 and 0.53, respectively). Bootstrap support for the standard (-A) and informative (-I) treatment of intra-individual site polymorphism shown for three algorithms: Maximum Likelihood (ML), Maximum Parsimony (MP) and Neighbor Joining (NJ). All bipartitions in the bootstrap networks (not shown) with support ≥20% in more than one analysis are shown.
Phylogenetic inference using the informative approach on this 2ISP-rich data set generally supported the splits graph groupings described above (Fig. 4e–g). The samples separate into well-supported clades that correspond to drainage basins under NJ-I. In contrast only a single clade that includes a few samples from the Gouritz basin receives support under NJ-A. The MP-A strict consensus approaches a star tree, that is, is entirely unresolved, whereas the Gouritz, Gamtoos, and Sundays samples received high, moderate, and low support (>50%) as clades, respectively, in the MP-I tree. The ML-A tree supports only one clade (Gamtoos) whereas supported clades for the other two basins do not contain all of the samples from the respective basins. In contrast, the Gamtoos and Gouritz samples form well-supported clades in the ML-I tree, with greater intraclade resolution. The Sundays samples only receive low support from NJ-I and MP-I and form a basal grade in the ML-I tree. This lack of support and basal grading is due to a rooting issue. There is simply not enough signal in the data to support a Sundays clade under the conditions used by NJ-I and MP-I due to the placement of the outgroup-determined ingroup root, and ML-I nests the outgroup within the Sundays clade. When these analyses are performed without the distantly related outgroup samples, all -I treatments resolve the Sundays clade with moderate to high support (all ≥86 in Fig. 4a).

**Discussion**

Multicopy nuclear DNA regions are potentially of immense utility in reconstructing evolutionary history. However, reconstructing phylogenies from these regions is frequently frustrated by the occurrence of 2ISPs. Here we argue that this is largely due to the manner in which such sites are treated, with previous approaches generally considering 2ISPs as uncertainties, thereby diminishing or nullifying their contribution to phylogeny reconstruction. We show here that treating them as informative, either through use of simple stepmatrices under NJ and MP or by including them as alternative character states under ML, can substantively increase the phylogenetic signal that can be extracted from such sites. We have avoided complicating the performance of the various phylogenetic criteria under the 2ISP-informative treatment as this extends beyond the current aims of this article.

The effect of polymorphism on phylogeny reconstruction, irrespective of data source (e.g., morphology or DNA), is a long-standing problem (reviewed in Wiens 1999). Various approaches have been used to mitigate the decline in resolution and support when 2ISPs are present: 1) Removing putative hybrids (e.g., Whittall et al. 2000; Aguilar and Feliner 2003; Fehrer et al. 2009), 2) removing samples with two or more 2ISPs (e.g., Hanna et al. 2007), 3) removing sites that contain 2ISPs (e.g., Scherson et al. 2008), 4) treating all DNA bases (including 2ISPs) as unordered characters (e.g., Campbell et al. 1997; Fama et al. 2000), or 5) cloning samples containing 2ISPs and including clones rather than the original sequence in the phylogenetic analyses (suggested standard practice by Feliner and Rosselló 2007). In an attempt to address analytical issues created by 2ISPs, statistical haplotype phasing methods have even been used to infer variants within ITS sequences based on 2ISP patterns (e.g., Lorenz-Lemke et al. 2005), despite such methods having been developed for low-copy regions. Statistical haplotype inference assumes that the gene region in question is diploid and two alleles are present (Clark 1990; Stephens et al. 2001)—this assumption is violated with complex multigene families. All of these approaches can result in the loss of information either through direct removal or because the association between variants and samples is lost (but see the following examples that show how to preserve the 2ISP information represented in cloned variants Joly and Bruneau 2006; Grimm et al. 2007a; Göker and Grimm 2008).

Our simulations demonstrate clearly that reduced resolution and branch support in the presence of 2ISPs is largely due to their treatment as uncertain or missing characters. Treating them as informative enabled the incorporation of these additional data in a straightforward manner that significantly enhanced phylogenetic inference (Table 1). Performance of the 2ISP-informative treatment was also considerably, and often significantly, better than the standard 2ISP-ambiguous treatment for real-world data (Figs. 2, 4, online Appendix S3, available from Dryad data repository; doi:10.5061/dryad.2lj2j). The presence of putative hybrid samples does not result in the informative treatment supporting erroneous relationships (Figs. 3, S3.1–3.11). Thus, the substantial improvements seen when real-world data sets are analyzed under the 2ISP-informative treatment (Figs. 2–4, S3.1–3.11) demonstrate that 2ISP variation provides quickly accessible additional information on phylogenetic relationships among samples, presumably because it encapsulates considerable inherited variation. The 2ISP-informative approach is especially more powerful when data sets have a positive P-index (2ISPs outweigh standard parsimony-informative characters); however, even under negative P-index values an improvement in resolution and support is evident (Fig. 2). These results suggest that phylogenetic inference under the 2ISP-informative approach does not require the exact sources of intra- or interarray variation to be known. These include many different mechanisms, some of which maintain evolutionary signal (e.g., incomplete concerted evolution where mutation rates outstrip homogenization processes) and others that blur it (e.g., hybridization).

The 2ISP-informative treatment does not solve the general problem of conflicting signal from hybrids, or reticulation in general, in tree inference. However, this is not the aim of the approach. The placement of hybrids within a tree is (and should be) unpredictable—at best they group with one of the progenitors—and this affects
FIGURE 4. The *Nymania capensis* ITS data set analysed using network and tree reconstruction methods. a) Neighbour-net splits graph with intra-individual site polymorphisms (2ISPs) treated as informative characters. Bootstrap support values for a subset of important bipartitions for three methods (NJ: neighbor joining; MP: maximum parsimony; ML: maximum likelihood) are plotted for the 2ISP-ambiguous (-A) and 2ISP-informative (-I) treatments (see online Appendix S3 for a fully labelled figure). Support for the Sundays clade when the outgroup was excluded is shown (arrow) and is discussed in the text. b–g) NJ, MP, and ML trees and bootstrap support inferred when treating 2ISPs as ambiguous or informative characters. Dashed branches have been reduced by a factor of 10. Branch support is shown using a combination of line thickness and color.
bootstrap support. Hybrid samples have traditionally been removed because of the conflicting signal they carry (Vriesendorp and Bakker 2005). Fehrer et al. (2009) had to remove putative hybrid sequences from the *Hieracium* data set in order to resolve the two geographic clades with sufficient support. However, as highlighted by Vriesendorp and Bakker (2005, p. 598), “the common practice of leaving suspected [hybrid] taxa out of the analysis to avoid confounding effects on phylogenetic reconstruction will not stimulate further progress.” The incapacity of dichotomous trees to deal with reticulate signal can be overcome by using phylogenetic networks based on collections of trees or distance matrices (Vriesendorp and Bakker 2005; Huson and Bryant 2006; Lockhart 2006), but most implementations of these network methods do not treat 2ISPs as informative (e.g., NeighbourNet, Bryant and Moulton 2004—but see the “Average” option for calculated uncorrected p-distance in SPLITSTREE; statistical parsimony networks, Clement et al. 2000). 2ISPs can contribute to network reconstruction either by using the 2ISP-informative approach to differentiate between insufficient data versus conflicting signals as the cause of weak support or by basing distance matrices on polymorphism p-distances (e.g., Figs. 3, 4a, online Appendix S3, available from Dryad data repository; doi:10.5061/dryad.21j[2]). For example, the western and eastern clades identified by Fehrer et al. (2009) are not supported under MPA-M (Maddison 1997), which in combination may improve phylogeny support and may enhance the evolutionary signal detectable by direct sequencing (although rare variants can persist in the genome even through multiple speciation events, Mahelka and Keckey 2010). Therefore, the full (hybrid-included) data set (Fig. 3b); the recoding of polymorphisms to the most frequent base at a given site in a data set under NJ- A (the default treatment of 2ISPs in PAUP* for NJ analyses), however, has, somewhat fortuitously, led to support for these two clades. In contrast, the informative treatment shows moderate to high support for these clades (all ≥ 80) under all inference methods despite the presence of hybrids (Fig. 3b). This is true for many other bipartition comparisons observed in this and the *Acer* neighbour-net splits graph (Fig. 3a).

When 2ISPs are derived from population-level processes (or independent evolutionary histories in general) that maintain the underlying evolutionary signal, then the informative approach dramatically improves phylogeny support and may enhance the resolution of species or population trees (e.g., Table 1, Figs. 3 and 4). This is because the separate variants represent independent and unique gene histories (Maddison 1997), which in combination may improve estimation of the species or population trees (e.g., Göker and Grimm 2008, figure 5). The simplistic single variant simulations used here demonstrate that the combined A1&2 data sets usually resolve more “true” bipartitions, with better support, than the independent variant data sets (A1 and A2; Table 1) when analyzed using the 2ISP-informative treatment. Thus, 2ISPs may offer a consensus signal of the underlying coalescent gene trees of a number of variants.

An example of multiple variants representing population-level processes is seen in *Nymania capensis* (see Schloe et al. 2011 and Fig. S3.10 for another example). The 2ISPs here are unlikely to be due to interspecies hybridization (*Nymania* is a monotypic genus), but can instead be attributed to lack of concerted evolution, divergence of ITS copies, and inheritance of polymorphisms. The clusters found in the splits graph (Fig. 4a) and clades resolved in the NJ-I, MP-I, and ML-I trees (Figs. 4e–g) provide valuable phylogeographic information, and suggest that populations have been isolated in different drainage basins (likely driven by range contraction during the glacial periods in the Pleistocene, Potts et al. 2013b), as supported by chloroplast data (Potts et al. 2013a). The high proportion of 2ISPs in the *Nymania* data set would previously have precluded it from being used because many clades would be unsupported (e.g., NJ-A and ML-A; Fig. 4b,d) or so many most-parasimonious trees exist that the strict consensus tree is star-like (MP-A; Fig. 4c). Thus, where population-level heterogeneity exists, treating 2ISPs as informative provides both a means of dealing with this heterogeneity, allowing polymorphisms in hypothetical ancestral taxa, and an estimate of the “population” tree from summaries of different variants. Unfortunately, the process of concerted evolution can greatly reduce the variant heterogeneity and thus remove much of the evolutionary signal detectable by direct sequencing (although rare variants can persist in the genome even through multiple speciation events, Mahelka and Keckey 2010). Therefore, the full (hybrid-included) data set (Fig. 3b); the recoding of polymorphisms to the most frequent base at a given site in a data set under NJ- A (the default treatment of 2ISPs in PAUP* for NJ analyses), however, has, somewhat fortuitously, led to support for these two clades. In contrast, the informative treatment shows moderate to high support for these clades (all ≥ 80) under all inference methods despite the presence of hybrids (Fig. 3b).
signal from the unavoidable polymorphic base calls. Inference programs fail to extract crucial phylogenetic approaches have not been widely used because most will most likely generate 2ISPs. However, consensus analyses. Creating a consensus sequence of all copies allows for straightforward array-based phylogenetic arrays are combined into a single consensus sequence, based approach, where all copies within an array or sequences will always be an issue. Using a consensus-available data to a meaningful number of representative in fast phylogenetic inferences, the reduction of all will be required. Even with the expected advances that incorporates inter- and intra-array processes chromosomes and between individuals. Also, a model number of copies may vary between both parental other cases homology cannot be assumed and the first and last repeat of each array are strictly homologous Aligning such data may be impossible because only the first and last repeat of other sequenced arrays from the same individual or other individuals. In all other cases homology cannot be assumed and the number of copies may vary between both parental chromosomes and between individuals. Also, a model that incorporates inter- and intra-array processes will be required. Even with the expected advances in fast phylogenetic inferences, the reduction of all available data to a meaningful number of representative sequences will always be an issue. Using a consensus-based approach, where all copies within an array or arrays are combined into a single consensus sequence, allows for straightforward array-based phylogenetic analyses. Creating a consensus sequence of all copies will most likely generate 2ISPs. However, consensus approaches have not been widely used because most inference programs fail to extract crucial phylogenetic signal from the unavoidable polymorphic base calls. Such polymorphisms are the raison d'être of the 2ISP-informative approach.

**Conclusion**

In a phylogeographic study of *Vaccinium uliginosum*, Eidesen et al. (2007) state that “as a consequence [of intra-individual site polymorphisms], unless all ITS PCR products are cloned, any phylogenetic signal useful for inferring relatively recent phylogeographical patterns is effectively concealed by the polymorphisms caused by the two [or more] paralogous ITS repeats.” This frustration has been shared by many researchers. Here we have demonstrated that treating 2ISPs as informative characters can dramatically improve phylogenetic resolution. We envisage that this method should greatly aid phylogenetic inference at the intrageneric or intraspecific level, including phylogeographic studies.

**Supplementary Material**

Supplementary material, including data files, DNA matrices, and online-only appendices, can be found in the Dryad data repository at http://datadryad.org/, doi:10.5061/dryad.2j1j2. The online-only appendices contain the following: Online Appendix S1: Summary of the frequency of DNA regions sampled across all plants. Online Appendix S2: Further methods details. Online Appendix S3: Neighbor-net splits graphs with bootstrap bipartition support for method and treatment combinations for a subset of the published data sets used in this study.

**Funding**

This work was supported by the South African National Research Foundation (A.J.P., T.A.H.); the Claude Leon Foundation (A.J.F.); University of Cape Town (A.J.P., T.A.H.); and the Swedish Research Council (G.W.G.). Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the supporting organizations.

**Acknowledgments**

We greatly appreciate the help of Alexandros Stamatakis with RAxML and for commenting on an early draft of this manuscript. Also, Emmanuel Paradis and Klaus Schliep provided invaluable assistance with the APE and PHANGORN libraries in R. We thank two reviewers for valuable comments that greatly improved this study. All RAxML analyses were conducted on facilities kindly provided by the Centre for High Performance Computing, Cape Town.
### APPENDIX

Table A.1 Variable sites in direct-PCR ribosomal ITS sequences from *Nymania capensis* accessions

<table>
<thead>
<tr>
<th>Drainage Basin</th>
<th>Sample</th>
<th>ITS-1</th>
<th>ITS-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>nuc. pos</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Consensus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gamtoos</td>
<td>AJP0270</td>
<td>A C</td>
<td>r w</td>
</tr>
<tr>
<td></td>
<td>AJP0540</td>
<td>m M</td>
<td>r w y</td>
</tr>
<tr>
<td></td>
<td>AJP058</td>
<td>m M</td>
<td>r w y</td>
</tr>
<tr>
<td></td>
<td>AJP055</td>
<td>m M</td>
<td>r w y</td>
</tr>
<tr>
<td></td>
<td>AJP059</td>
<td>m M</td>
<td>r w y</td>
</tr>
<tr>
<td>Gouritz</td>
<td>AJP0030</td>
<td>W W</td>
<td>m w</td>
</tr>
<tr>
<td></td>
<td>AJP0071</td>
<td>W W</td>
<td>m w</td>
</tr>
<tr>
<td></td>
<td>AJP0074</td>
<td>W W</td>
<td>m w</td>
</tr>
<tr>
<td></td>
<td>AJP0118</td>
<td>W W</td>
<td>m w</td>
</tr>
<tr>
<td></td>
<td>AJP0243</td>
<td>W W</td>
<td>m w</td>
</tr>
<tr>
<td></td>
<td>AJP0774</td>
<td>W W</td>
<td>m w</td>
</tr>
<tr>
<td>Sundays</td>
<td>AJP0465</td>
<td>k k</td>
<td>m X</td>
</tr>
<tr>
<td></td>
<td>AJP0466</td>
<td>k k</td>
<td>m X</td>
</tr>
<tr>
<td></td>
<td>AJP0492</td>
<td>k m</td>
<td>m X</td>
</tr>
<tr>
<td></td>
<td>AJP0495</td>
<td>k m</td>
<td>m X</td>
</tr>
<tr>
<td></td>
<td>AJP0530</td>
<td>k m</td>
<td>m X</td>
</tr>
<tr>
<td></td>
<td>AJP0532</td>
<td>k m</td>
<td>m X</td>
</tr>
<tr>
<td></td>
<td>AJP0625</td>
<td>k m</td>
<td>m X</td>
</tr>
<tr>
<td></td>
<td>AJP0632</td>
<td>k m</td>
<td>m X</td>
</tr>
<tr>
<td></td>
<td>AJP0637</td>
<td>k m</td>
<td>m X</td>
</tr>
<tr>
<td></td>
<td>AJP0828</td>
<td>k m</td>
<td>m X</td>
</tr>
</tbody>
</table>

21SPs are coded using IUPAC nomenclature, with capital letters indicating complete overlap of bases and small letters indicating one base dominant over another at a site in the trace file. 21SPs involving a base and an indel are coded using “X”. All sequences are compared with the reference consensus sequence.

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary

Denk T., Grimm G.W. 2010. The oaks of western Eurasia: traditional

Fehrer J., Krak K., Chrtek J. 2009. Intra-individual polymorphism in

Farris J.S. 1978. Inferring phylogenetic trees from chromosome

Clement M., Posada D., Crandall K. 2000. TCS: a computer program to

Clark A.G. 1990. Inference of haplotypes from PCR-amplified samples


Lorenz-Lemke A.P., Muschner V.C., Bonatto S.L., Cervi A.C., Salzano F.M., Freitas L.B. 2005. Phylogenetic inferences concerning...


