Syst. Biol. 57(3):347–366, 2008
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802044037

Optimizing Automated AFLP Scoring Parameters to Improve Phylogenetic Resolution

Barbara R. Holland,[1,2] Andrew C. Clarke,[1,3] and Heidi M. Meudt[1,4]

[1]*Allan Wilson Centre for Molecular Ecology and Evolution,* [2]*Institute of Fundamental Sciences, and* [3]*Institute of Molecular BioSciences, Massey University, Private Bag 11222, Palmerston North 4442, New Zealand; E-mail: B.R.Holland@massey.ac.nz (B.R.H.); A.C.Clarke@massey.ac.nz (A.C.C.)*
[4]*Museum of New Zealand Te Papa Tongarewa, PO Box 467, Wellington 6140, New Zealand; HeidiM@tepapa.govt.nz (H.M.M.)*

Abstract.—The amplified fragment length polymorphism (AFLP) technique is an increasingly popular component of the phylogenetic toolbox, particularly for plant species. Technological advances in capillary electrophoresis now allow very precise estimates of DNA fragment mobility and amplitude, and current AFLP software allows greater control of data scoring and the production of the binary character matrix. However, for AFLP to become a useful modern tool for large data sets, improvements to automated scoring are required. We design a procedure that can be used to optimize AFLP scoring parameters to improve phylogenetic resolution and demonstrate it for two AFLP scoring programs (GeneMapper and GeneMarker). In general, we found that there was a trade-off between getting more characters of lower quality and fewer characters of high quality. Conservative settings that gave the least error did not give the best phylogenetic resolution, as too many useful characters were discarded. For example, in GeneMapper, we found that bin width was a crucial parameter, and that although reducing bin width from 1.0 to 0.5 base pairs increased the error rate, it nevertheless improved resolution due to the increased number of informative characters. For our 30-taxon data sets, moving from default to optimized parameter settings gave between 3 and 11 extra internal edges with >50% bootstrap support, in the best case increasing the number of resolved edges from 14 to 25 out of a possible 27. Nevertheless, improvements to current AFLP software packages are needed to (1) make use of replicate profiles to calibrate the data and perform error calculations and (2) perform tests to optimize scoring parameters in a rigorous and automated way. This is true not only when AFLP data are used for phylogenetics, but also for other applications, including linkage mapping and population genetics. [AFLP; amplified fragment length polymorphism; automated scoring; error rates; phylogenetic resolution; phylogeny; scoring parameters.]

Amplified fragment length polymorphism (AFLP) DNA fingerprinting (Vos et al., 1995) is a firmly established molecular marker technique for evolutionary, genetic, and ecological studies of plants, animals, and microorganisms (Mueller and Wolfenbarger, 1999; Meudt and Clarke, 2007). AFLP has a number of broad applications, ranging from linkage mapping to analyses using population-based and phylogenetic methods. Of particular interest in this study is the use of AFLP to generate data for phylogenetic studies. Although some researchers have suggested that AFLP data are inappropriate for phylogenetic applications (Hollingsworth and Ennos, 2004; Kosman and Leonard, 2005), several empirical studies have revealed tree-like properties in AFLP data sets, and AFLP data are increasingly being used to estimate phylogenies, including for very shallow radiations (e.g., Marhold et al., 2004; Sullivan et al., 2004; Koopman, 2005; Mendelson and Shaw, 2005; Spooner et al., 2005a; Albach, 2007; Kilian et al., 2007). Meudt and Clarke (2007) reviewed several conditions in which the AFLP technique can be ideal. For accurate phylogeny estimation, these conditions include high genomic heterogeneity (i.e., when it is necessary to analyze many loci to ascertain an accurate measure of genomic diversity), low genetic variability (generally intraspecific comparisons of, for example, crop species, and intrageneric comparisons such as rapid species radiations), and studies of polyploids where it is very difficult to use single-locus nuclear sequencing markers because of problems distinguishing the many alleles that may be present at each locus. Because AFLPs are predominantly nuclear markers that are widely distributed throughout the genome, they are particularly powerful for studying the phylogeny of organisms such as plants for which other nuclear and organellar markers are often lacking, insufficiently variable, or even inappropriate (Després et al., 2003; Pelser et al., 2003; Marhold et al., 2004; Bensch and Åkesson, 2005; Tremetsberger et al., 2006; Pellmyr et al., 2007). AFLPs can also complement other marker systems (such as DNA sequencing markers) in a phylogenetic study by, for example, providing resolution in different parts of the tree (Després et al., 2003; Pelser et al., 2003; Marhold et al., 2004; Koopman, 2005; Spooner et al., 2005b). AFLPs may not be able to provide an accurate estimate of the species phylogeny when genetic divergence is too high (although precisely where this cut-off is has been the subject of debate; see Meudt and Clarke (2007) and references therein), or when frequent hybridization and reticulation have obscured the phylogenetic signal throughout the phylogeny, but these shortcomings are not unique to AFLP data.

The AFLP technique is usually applied to sets of organisms with low genetic divergence (Bonin et al., 2007; Meudt and Clarke, 2007) and, as such, has similar potential drawbacks to other marker systems. For example, when these cases involve large ancestral population sizes and short times between divergence events, incomplete lineage sorting can often result in the phylogeny of a single locus not matching the species phylogeny (Degnan and Salter, 2005). For alignments of genes that have been sequenced and concatenated, new methods have recently been developed that estimate the species phylogeny taking into account the probability of gene trees under a coalescent model (Carstens and Knowles, 2007; Liu and Pearl, 2007). Such methods have not yet been developed for AFLP. Nevertheless, except in those hopefully rare cases where the most likely gene tree does not match the species tree (Degnan and Rosenberg,

All authors contributed equally to this work.

2006), it is expected that phylogenetic analysis of AFLP data should give a robust estimate of the species phylogeny (Sullivan et al., 2004; Koblmüller et al., 2007). This is essentially an issue of sample size; the hundreds or thousands of concatenated AFLP loci from a typical AFLP study are more likely, on average, to approximate the species phylogeny, and, except when some parametric conditions arise (Kubatko and Degnan, 2007), the individual effects of loci that have evolutionary histories different to that of the species are more likely to be diminished. And indeed, when phylogenies are estimated both from AFLP data and from robust, independent sources of evidence (e.g., phenotypic traits, behavior, and ecology), congruent results have been obtained (e.g., Marhold et al., 2004; Sullivan et al., 2004; Koblmüller et al., 2007; Pellmyr et al., 2007).

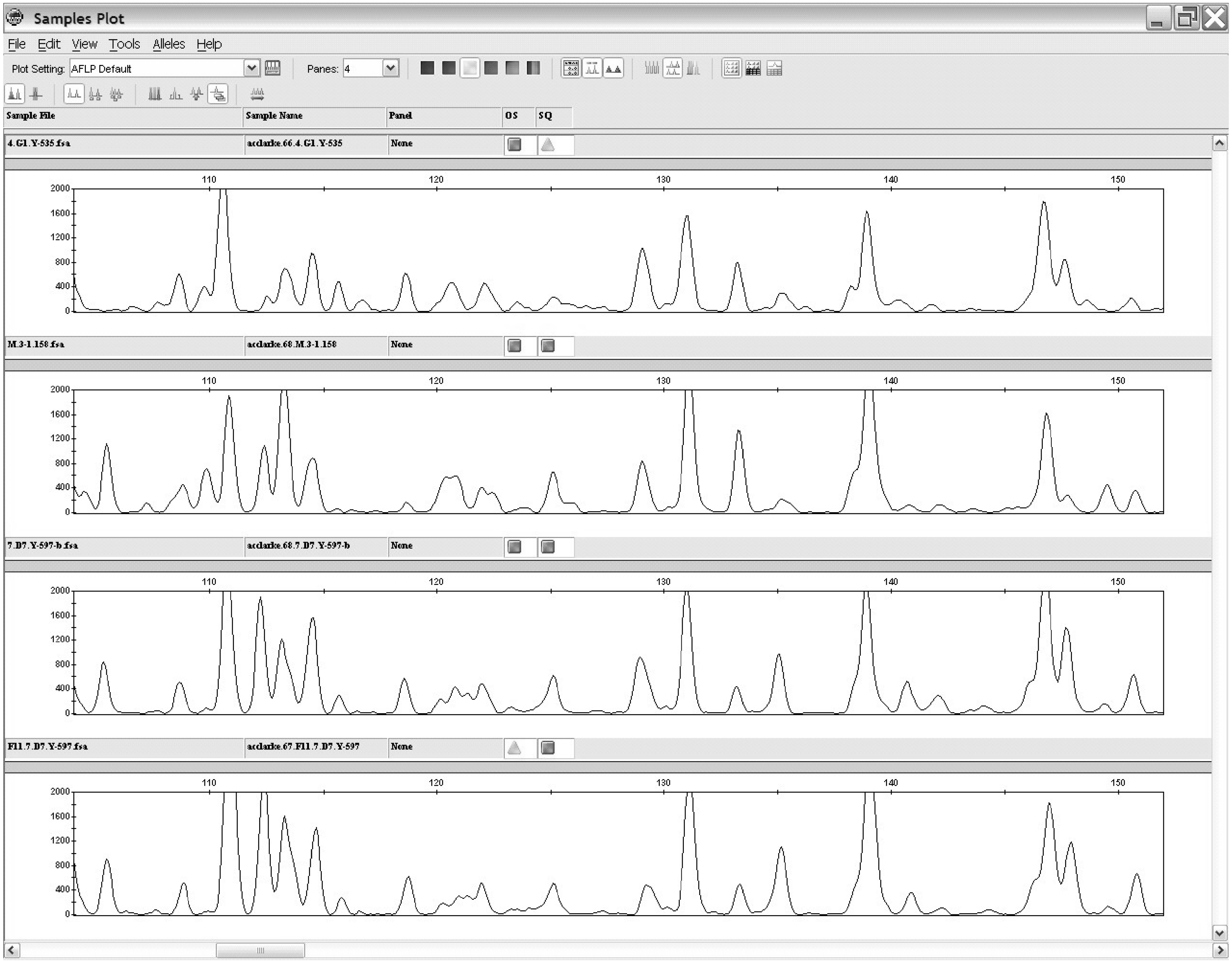
Although the appropriateness of AFLP data for phylogenetic reconstruction requires further study, it is nevertheless widely employed in phylogenetics and systematics (e.g., Marhold et al., 2004; Mendelson and Shaw, 2005; Koblmüller et al., 2007; Pellmyr et al., 2007). It is essential that we gain a better understanding of the situations in which AFLP data may be inappropriate (e.g., if incomplete lineage sorting is occurring), but it is outside the scope of this paper to address this further; instead, our focus is on optimizing AFLP scoring parameters to maximize the phylogenetic signal obtained from the raw data.

To generate AFLP data, a complete restriction endonuclease digestion is performed on total genomic DNA, followed by two rounds of selective polymerase chain reaction (PCR) amplification and separation of the fragments by electrophoresis (Vos et al., 1995; Meudt and Clarke, 2007). In capillary electrophoresis of fluorescently labeled AFLP fragments, the end result is the production of a profile like the ones shown in Figure 1. To convert the data for numerous profiles into a binary character matrix of 0's (peak absent; null allele) and 1's (peak present; plus allele), two types of decisions have to be made. First, when should a fragment be called as present (character state 1) versus absent (character state 0)? Second, when should two fragments be designated as having the same length (number of nucleotides) and therefore be treated as identical plus-alleles of the same locus? The ideal is to have all truly identical fragments recognized, scored as present, and assigned to the same column of the character matrix—and to have no nonidentical fragments assigned to the same column of the character matrix. In practice, this is not likely because some nonidentical fragments will have similar mobility by chance, identical fragments will have slightly different mobilities and peak heights due to random error (Vekemans et al., 2002), and shared absences (null alleles) may be derived in multiple, independent ways.

New capillary-based technologies allow more precise estimates of AFLP DNA fragment mobility (fragment length) and fluorescence intensity (amplitude) than traditional gel-based systems. Furthermore, analysis of capillary profiles with currently available automated scoring software (see table 1 in Meudt and Clarke, 2007) allows the user to control several parameters that influence the resulting data matrix. In contrast to manual scoring, automated scoring is objective, repeatable, and far less time-consuming. In fact, with increasingly large data sets, automated scoring is often the only feasible option, yet to our knowledge no experimental or theoretical studies have explored different automated scoring parameter settings and their effects on downstream analyses. Given that AFLP has many potential applications and that the automated scoring packages have many adjustable parameters, it is natural to ask: how can we measure the quality of the AFLP character matrix, and what is the best way to go about optimizing AFLP scoring parameters for phylogenetic studies? More specifically, when scoring a particular AFLP data set, which parameter settings will give the most accurate phylogenetic estimate?

Our aim is to find parameter settings for automated scoring software that lead to data matrices whose analysis allows us to accurately recover the true tree (i.e., the species phylogeny). However, there are difficulties in directly measuring the accuracy of phylogenetic estimates. This is because (1) in general the true tree is not known, and (2) the question is not amenable to study with a simulation-based approach, such as AFLP in silico (Qin et al., 2001; Bikandi et al., 2004), because the factors that influence bin width and peak height are not currently understood well enough to be simulated accurately. Because we cannot measure accuracy directly and simulation studies are not applicable, we use the resolution of the phylogenetic tree resulting from bootstrap analysis of the data matrices constructed with different parameter settings as a proxy for accuracy. The higher the resolution, the more information there is about phylogenetic relationships, and unless there is some systematic bias, high resolution should be correlated with accuracy (Hillis and Bull, 1993). For example, Taylor and Piel (2004) showed empirically that high bootstrap support was strongly correlated with accuracy in their study using a genome-scale yeast data set.

We explore the effect of parameter choice using two commonly available software platforms designed for automated AFLP scoring: GeneMapper v. 3.7 by Applied Biosystems and GeneMarker v. 1.51 by SoftGenetics. The parameters studied include: the minimum peak height threshold required for a peak to be called as present, the minimum fragment length at which a marker is scored and included as a character in the matrix, and the width of the marker bins in base pairs (bp). Each of these parameters influences the number of characters available for phylogenetic analysis and whether or not these characters represent homologous fragments (Fig. 2). Introducing more homologous characters should lead to higher resolution, but in practice by including more characters we also risk introducing errors. For each of the main parameters studied, we expect there to be a trade-off



Plot Setting: AFLP Default

Panes: 4

Sample File	Sample Name	Panel	OS	SQ
4.G1.Y-535.fsa	acclarke.66.4.G1.Y-535	None	■	▲
M.3-1.158.fsa	acclarke.68.M.3-1.158	None	■	■
7.D7.Y-597-b.fsa	acclarke.68.7.D7.Y-597-b	None	■	■
F11.7.D7.Y-597.fsa	acclarke.67.F11.7.D7.Y-597	None	▲	■

FIGURE 1. Example AFLP profiles. This screenshot from GeneMapper shows AFLP profiles for four samples of *Ipomoea batatas*, from top to bottom: Y-535 Society Islands, 158 cv. Mary Anne, Y-597 Peru B, and Y-597 Peru A. Thus, the first three rows show profiles from three different individuals, and the last two samples are replicate AFLP profiles obtained from a single individual (Y-597 Peru). Each profile consists of a plot of fluorescence (relative fluorescent units; rfu) versus fragment mobility/size (base pairs; bp) for one fluorescently tagged primer pair combination—in this case from approximately 105 to 150 bp using a NED-labeled primer. These raw fluorescence data are converted to binary data by first binning the data (grouping the similar-sized peaks from different accessions into a single character) and then scoring the peaks in that character as 1 (present) or 0 (absent). The resulting character matrix of 0's and 1's is then exported for phylogenetic analysis.

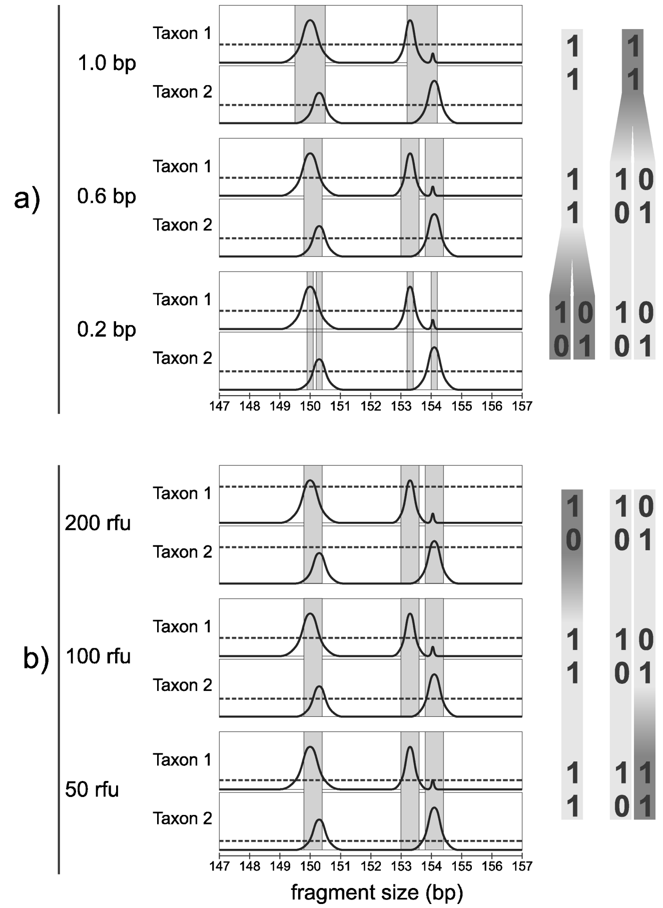


FIGURE 2. Theoretical effect of changing specific AFLP scoring parameter settings; (a) bin width and (b) peak height threshold. The binary matrix on the right shows the effects of changing the AFLP scoring parameters in the profiles on the left; correctly scored peaks are represented in the binary matrix in light grey and incorrectly scored peaks in dark grey. (a) Bin widths (BW) are shown as grey rectangles underneath the profiles; peaks that fall within a given bin width are scored as 1 (present) and outside as 0 (absent). Using a bin width that is too wide (1.0 bp) makes it more likely that identical alleles (whose mobilities differ only slightly because of random error) are treated as a single character, but it can also cause non-identical alleles to be incorrectly treated as one character. Although this problem is corrected by narrowing the bin width (0.6 bp), if the bin width is too narrow (0.2 bp), then even the identical alleles will be wrongly split into separate characters. (b) Peak height threshold (PHT) is shown using black dashed lines; a peak above this line is scored as 1 (present) and below as 0 (absent). If the PHT is set too high (200 rfu), then peaks that are present will be scored as absent (taxon 2, left peak). Although this is corrected by lowering the PHT (100 rfu), if the PHT is too low (50 rfu), then background noise or stutter peaks will be incorrectly scored as present (taxon 1, right peak). This simplified example shows two hypothetical taxa and three characters, but real data sets may contain hundreds of taxa and hundreds of characters; determining the optimum parameter settings over all taxa and all characters is much more complex.

between getting more characters of lower quality and fewer characters of high quality.

Specifically, these trade-offs include the following. First, lowering the minimum fragment length will increase the number of characters, but these characters may be of lower quality, as previous studies (Vekemans et al., 2002) indicate that smaller fragments are more likely to be homoplasious than larger fragments. Second, reducing the bin width will increase the number of characters, but as bin width is reduced, single characters may split into pairs of characters (Fig. 2a), which at the very least can reduce branch support in the resulting tree but could also potentially introduce error into the data set. Conversely, as we increase bin width, separate characters will be amalgamated. If these characters are not really identical, this could create character conflict in the resulting data matrix (Fig. 2a), which may reduce resolution. Third, lowering the peak height threshold will increase the number of characters. If peak height threshold is set too low, we will, by scoring background noise, call peaks present when they are really absent, and if it is set too high we will call peaks as absent when they are really present (Fig. 2b).

To study the effect of different parameter choices on automated scoring of AFLP data we used two example data sets with very different levels of divergence—these data sets represent the extremes of the phylogenetic problems to which AFLP can be applied. Both the empirical data sets contained a small number of known replicates so that the replicate error rate could be calculated. In addition, we analyzed a much larger data set of 25 replicate pairs ($n = 6110$ individual pairwise comparisons) to determine the average sizing error between truly identical fragments. The aims of the study are to (1) determine if the parameter settings that maximize phylogenetic resolution are the same for both our test data sets or if the parameters are data dependent; (2) determine how robust the resulting phylogeny is to changes in automated scoring parameters; and (3) stimulate more studies of automated scoring of AFLP data and encourage improvements to available software.

METHODS

Sampling Strategy

Two data sets, each of 30 AFLP profiles, were created: one from accessions of sweet potato (*Ipomoea batatas*; Convolvulaceae) and one from multiple individuals of several New Zealand species of the plant genus *Ourisia* (Plantaginaceae). Sweet potato is a single domesticated species, and as such it is expected to have relatively little genetic diversity compared to wild taxa. Because sweet potato is generally asexually reproducing, it is expected to evolve in a tree-like way. Conversely, *Ourisia* is a genus of 28 species found in high-elevation habitats mostly in the Southern Hemisphere. Probably the majority of *Ourisia* species are outcrossing, and therefore we expect relatively large amounts of genetic diversity. However, the New Zealand species of *Ourisia* exhibit low DNA sequence variation (Meudt and Simpson, 2006), and therefore AFLP markers might provide a suitable amount of variation for phylogenetic reconstruction. Putative hybrids were excluded from the *Ourisia* data set, so we expect the underlying evolutionary history to be at least approximated by a tree.

Each of these data sets represents a subset of a much larger data set produced to answer evolutionary questions in sweet potato (A.C.C.) and New Zealand *Ourisia* (H.M.M.). We reduced the data sets to 30 AFLP profiles each to give a representative sample of the larger data sets, allow comparison of data sets of the same size, and also decrease the time involved in preparing each of the character matrices and running the resulting analyses (see below). For a number of reasons, sampling strategy is extremely important in any AFLP study, particularly because it has been suggested that a thorough sampling strategy (i.e., sampling multiple individuals from multiple populations for each species under study) can actually improve the probability of coalescence near the tips of the tree and thus potentially increase the probability of capturing the true species tree (Degnan and Rosenberg, 2006). Especially for recently diverged taxa, sampling multiple individuals per species can provide additional information regarding species relationships (Carstens and Knowles, 2007). We therefore aimed to include in our respective larger unpublished studies in *Ipomoea* and *Ourisia* (where possible) multiple individuals from multiple populations from throughout the geographic ranges of all taxa involved. Likewise, for the smaller data sets used in the present study, we chose 30 AFLP profiles from each data set that represent the taxonomic diversity, genetic diversity, and geographic ranges of the organisms under study, in addition to the inclusion of a number of replicates. This sampling strategy resulted in an *Ourisia* data set containing 24 unique accessions (of 13 total species) and 6 replicates, and an *Ipomoea* data set containing 25 unique accessions (24 accessions of *I. batatas* and one accession of the outgroup *I. tiliacea*) and 5 replicates. For complete voucher information see Supplementary Table 1 (available at www.systematicbiology.org). AFLP profiles for all *Ipomoea* and three *Ourisia* replicate pairs were obtained from duplicate DNA extractions of the same leaf tissue. The three remaining *Ourisia* replicate pairs were independent profiles generated from the same DNA extraction.

Generation of Raw AFLP Data

AFLPs were generated based on the protocol of Vos et al. (1995) using an updated protocol for capillary detection of fluorescently labeled markers (see http://awcmee.massey.ac.nz/aflp/AFLP_Protocol.pdf). Briefly, DNA was digested with the restriction enzymes *Eco*R I and *Mse* I. Eco and Mse linkers were ligated to the restriction fragments and a subset of these were amplified using Eco+A and Mse+C preselective PCR primers. Selective amplifications were performed with four Eco+3/Mse+3 PCR primer combinations. Eco+3 primers were labeled with 6FAM (Sigma-Aldrich), VIC, NED, or PET (Applied Biosystems) fluorescent dyes. The fluorescently labeled selective amplification products were poolplexed, along with a GS-500 LIZ size standard, on a 3730 Genetic Analyzer (Applied Biosystems). Capillary electrophoresis was carried out at the Allan Wilson Centre Genome Service, Massey University.

Generation of Data Sets Using Different Automated Scoring Parameter Settings

We designed, and describe below, a procedure to optimize numerous automated scoring parameters. A flowchart showing the different methods used to investigate and optimize parameter settings is shown in Figure 3. For each of our two example data sets, we created 90 different character matrices in GeneMapper and 36 different character matrices in GeneMarker. Preliminary testing was performed to determine which parameters were most important. The most important parameters to be subsequently tested here were, for GeneMapper, peak height threshold (PHT), minimum fragment length (MFL), and bin width (BW); and for GeneMarker, PHT, MFL, stutter peak filter (SPF), and local and global detection percentages (LGDP). With respect to smoothing, "heavy" (GeneMapper), "enhanced" (GeneMarker), or "no smoothing" (both programs) all performed worse than the middle option of "light smoothing" (GeneMapper) or "smoothing" (GeneMarker), which we used here. In addition, in GeneMarker, the minimum peak score default of "fail < 1 check < 7 pass" performed worse than other settings with the second value less than 7. We therefore set the minimum peak score to "fail < 1 check < 1 pass" in which peaks below a score of 1 were discarded, and those above the score of 1 were automatically accepted, thus fully automating the scoring process. All other parameters, preliminary testing of which showed negligible effects on scoring, were left at their default values.

The 90 GeneMapper matrices (3 × 3 × 10) were obtained by setting PHT to 50, 100, or 200 relative fluorescence units (rfu), setting MFL to 50, 100, or 150 bp, and adjusting BW from 0.1 to 1.0 bp in increments of 0.1. (In GeneMapper, the allele-calling threshold was set to the same value as the peak height threshold.) The 36 GeneMarker matrices (3 × 3 × 2 × 2) were obtained by setting the PHT to either 50, 100, or 200 rfu, the MFL to either 50, 100, or 150 bp, SPF to either its default of 5% or turned off, and LGDP to its default of 1% (both local and global) or turned off. Note that the PHT and MFL parameters are common to both GeneMapper and GeneMarker. The algorithms used by GeneMarker automatically allocate different BW to different characters in the matrix, as opposed to setting one BW for all characters in the matrix in GeneMapper. Although BW in GeneMarker can be subsequently changed so that all characters have identical values, doing so does not appear to greatly alter the number of characters in the matrix or resulting character states, and thus we did not consider BW further in GeneMarker.

It may seem counterintuitive to consider bin widths less than one base pair. However, both gel- and capillary-based AFLP systems measure mobility and only *estimate* length. Due to differences in strand composition (sequence and secondary structure), mobility values, and thus the estimated length values, effectively vary continuously. When the mobility difference of nonidentical DNA fragments differs by 1 bp or less, there can be serious problems of identity assessment. In such cases, bin

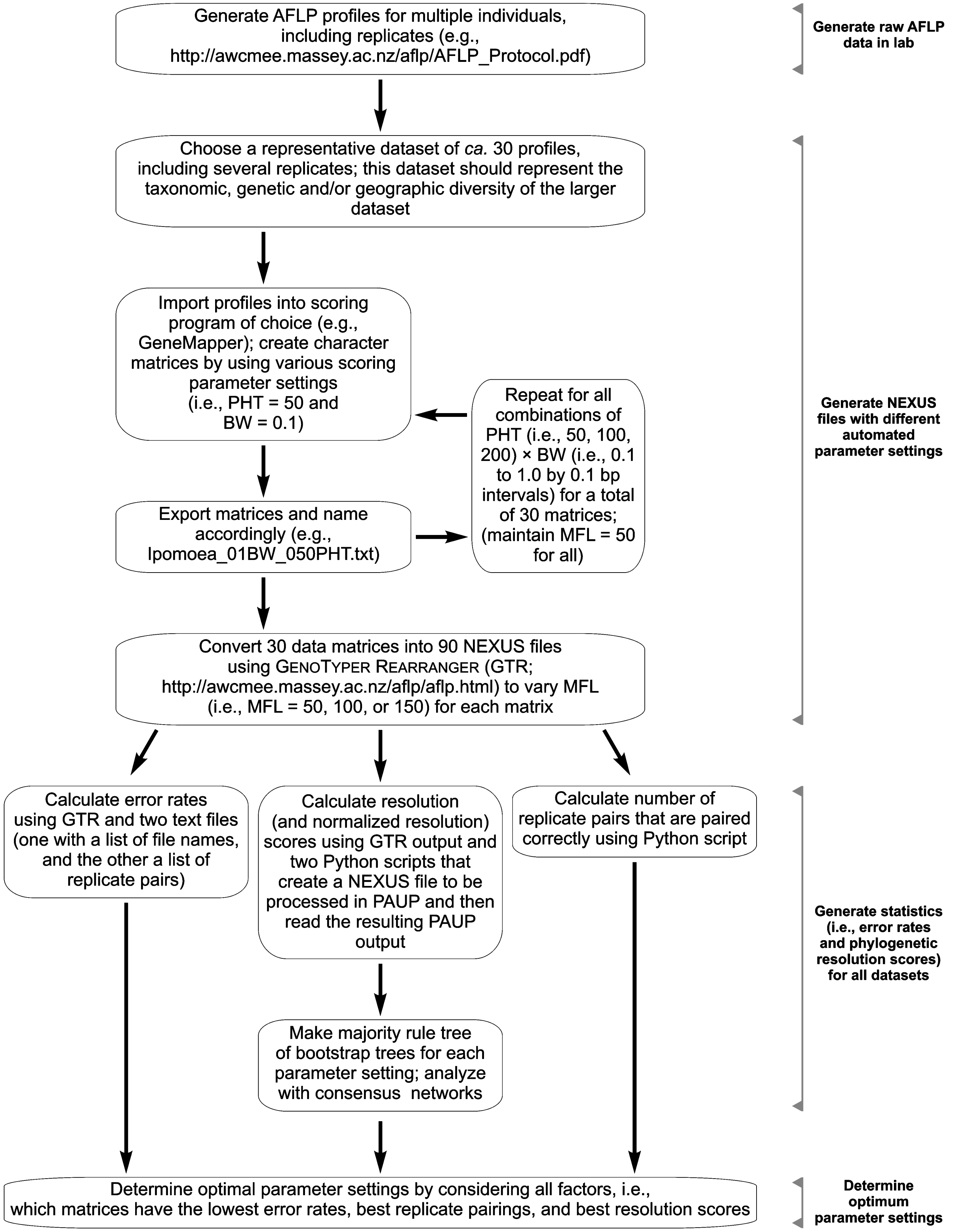

Generate AFLP profiles for multiple individuals, including replicates (e.g., http://awcmee.massey.ac.nz/aflp/AFLP_Protocol.pdf)
Generate raw AFLP data in lab
Choose a representative dataset of *ca.* 30 profiles, including several replicates; this dataset should represent the taxonomic, genetic and/or geographic diversity of the larger dataset
Import profiles into scoring program of choice (e.g., GeneMapper); create character matrices by using various scoring parameter settings (i.e., PHT = 50 and BW = 0.1)
Repeat for all combinations of PHT (i.e., 50, 100, 200) × BW (i.e., 0.1 to 1.0 by 0.1 bp intervals) for a total of 30 matrices; (maintain MFL = 50 for all)
Export matrices and name accordingly (e.g., Ipomoea_01BW_050PHT.txt)
Generate NEXUS files with different automated parameter settings
Convert 30 data matrices into 90 NEXUS files using GENOTYPER REARRANGER (GTR; http://awcmee.massey.ac.nz/aflp/aflp.html) to vary MFL (i.e., MFL = 50, 100, or 150) for each matrix
Calculate error rates using GTR and two text files (one with a list of file names, and the other a list of replicate pairs)
Calculate resolution (and normalized resolution) scores using GTR output and two Python scripts that create a NEXUS file to be processed in PAUP and then read the resulting PAUP output
Calculate number of replicate pairs that are paired correctly using Python script
Make majority rule tree of bootstrap trees for each parameter setting; analyze with consensus networks
Generate statistics (i.e., error rates and phylogenetic resolution scores) for all datasets
Determine optimal parameter settings by considering all factors, i.e., which matrices have the lowest error rates, best replicate pairings, and best resolution scores
Determine optimum parameter settings

FIGURE 3. Flowchart showing the steps involved in determining optimal scoring parameters using AFLP automated scoring software (e.g., GeneMapper) and methods and scripts described in this paper.

width settings <1 bp have the potential to separate nonidentical fragments with similar mobility.

Data matrices were exported from both programs and converted into NEXUS format files using the program GenoTyper Rearranger (GTR). The GTR program, a detailed protocol for using GTR, and all NEXUS files are available from http://awcmee.massey.ac.nz/aflp/aflp.html. Creating 126 matrices for each of the two 30-sample data sets via this streamlined process took approximately 4 hours.

Comparison of Data Sets to Determine Optimal Parameter Settings

Measures of accuracy.—For each character matrix we recorded the resolution score and the number of parsimony-informative characters. All phylogenetic analyses were carried out in PAUP* version 4.0b10 (Swofford, 2003) using both neighbor-joining (NJ) on uncorrected distances (dset dist = p) and with heuristic search using the parsimony optimality criterion (retaining all default settings in PAUP*). Both methods gave congruent trees with some local differences, but NJ gave higher resolution scores for all but 6 of the 252 parameter setting combinations tested (resolution was an average of 11.5% higher). For simplicity, we report only the NJ resolution scores. To calculate the resolution score for each character matrix, we performed 100 repetitions of 100 bootstrap replicates. For each replicate, all the bootstrap scores over 50% were summed and this number was divided by 27 (each data set had 30 samples so there were a maximum of 27 internal edges in each tree) to give a value between 0% and 100%. We then calculated the mean and standard deviation of the resolution score over the 100 repetitions.

We expect that both the quality of characters and the number of characters will have an effect on resolution and accuracy. If two data sets contain characters of the same quality, the data set with the most characters should give a more accurate phylogenetic estimate; if two data sets contain the same number of characters, the data set with the highest quality characters should give a more accurate phylogenetic estimate. To try and disentangle these two effects and get a measure of character quality independent of sequence length, we defined a normalized resolution score. We recorded the number of characters c_{min} in the smallest character matrix for each of the *Ourisia* and *Ipomoea* data sets (the data sets with PHT 200, MFL 150 bp, SPF 5%, LGDP 1% from GeneMarker). For each of the data sets and combination of parameter settings we created 100 new character matrices by sampling c_{min} columns of the original character matrix without replacement, thus creating many data sets of the same length. For each of these character matrices, we calculated the resolution score as outlined above. We then calculated the mean and standard deviation of this normalized resolution score over the 100 resampled alignments.

As another measure of accuracy, for each character matrix, we calculated the number of replicate pairs that were correctly assigned (i.e., as sister to one another).

For each character matrix, we calculated the replicate error rate (Bonin et al., 2004; Pompanon et al., 2005) as

$$\frac{N_{(0,1)} + N_{(1,0)}}{N_{(0,0)} + N_{(1,0)} + N_{(0,1)} + N_{(1,1)}},$$

where $N_{(0,0)}$ and $N_{(1,1)}$ represent the number of correct calls where a replicate pair both have no peak or both have a peak respectively; and $N_{(0,1)}$ and $N_{(1,0)}$ represent the number of incorrect calls where one half of the replicate pair has a peak and the other half does not. Each category is summed over all the replicate pairs in the data. This error rate is effectively the average Euclidean distance between replicate pairs. In the handful of animal and plant AFLP studies that have published them (see Bonin et al., 2004, and references therein), error rates have been calculated to be 2% to 5% using the above equation. However, the way in which this error rate is calculated makes interstudy comparisons very difficult. First, the formula includes a term for (0,0) calls in the denominator, which means that if the data set contains many plus alleles that are not present in a given replicate pair (i.e., scored as (0,0) in the replicates), then the error rate will appear to be lower. Thus, as more data are added to the data set or very divergent taxa are included, new, unique characters will be introduced and the apparent error rate will decrease. Second, the error rate includes both errors in the raw AFLP profiles themselves (e.g., PCR errors) as well as scoring errors. The number of scoring errors will vary widely between AFLP studies, depending on whether a manual, semiautomated, or automated scoring procedure is employed and which scoring software is used.

To check whether the size of our data sets changed the error rates, we recalculated the error rates for extended versions of both the *Ipomoea* and *Ourisia* data sets, containing 313 and 217 samples, respectively, using optimized parameter settings. To check if an increased number of (0,0) calls was masking an increased rate of (1,0) or (0,1) calls, we also defined an alternative error rate (effectively the average Jaccard distance between replicate pairs),

$$\frac{N_{(0,1)} + N_{(1,0)}}{N_{(1,0)} + N_{(0,1)} + N_{(1,1)}}.$$

Comparison of the standard (Euclidean distance) and alternative (Jaccard distance) error rates also allows us to test our prediction that a data set composed of divergent taxa will yield a lower standard error rate than a data set composed of the same number of closely related taxa.

To determine the relative contributions to the error rate of (1) errors in the raw AFLP profiles and (2) errors introduced during the automated scoring process, we used the program ReplicateError (which is available, along with a detailed protocol, from http://awcmee.massey.ac.nz/aflp/aflp.html). ReplicateError approximates the manual editing process by locating errors in a replicate pair (i.e., (0,1) or (1,0)) and testing to see if they can be corrected to (1,1) or (0,0) according to a set of predefined rules. ReplicateError detects three common types of (0,1) and (1,0) scoring errors: first, if a peak is detected but, because it falls below the PHT, it is scored as 0; second, if a peak is detected and is above the PHT, but because it does not meet all required quality criteria (e.g., peak shape), it is scored as 0; and third, if (0,1) and (1,0) errors comprise adjacent characters that are less than 0.5 bp apart. This third error type is caused by identical fragments that are only slightly different in length (due to random error) being binned as separate characters. See below for justification of why 0.5 bp is an appropriate range over which to amalgamate characters.

A number of scripts were written in Python to streamline the process of analyzing the PAUP* output and producing resolution scores and normalized resolution scores. All of these scripts along with detailed instructions on how to apply them to other data sets are available from http://awcmee.massey.ac.nz/aflp/aflp.html. The script to calculate error rates requires the set of NEXUS files created by GenoTyper Rearranger, a text file with a list of the file-names, and a text file with the list of replicate pairs. There are two scripts to calculate the resolution scores, the first uses the files that result from GenoTyper Rearranger and creates a NEXUS file to be processed by PAUP*. After PAUP* has been run, the second script reads in the resulting bootstrap trees and calculates the resolution scores. A similar process is used to calculate the normalized resolution scores. For each 30-sample data set, the NEXUS file used to compute the bootstrap trees for calculating the normalized resolution scores for 126 character matrices took approximately 3 hours to execute in PAUP* and the Python script took a further 30 min (on a Pentium 4). The resolution scores took less than an hour to compute in total.

Optimal parameter settings.—For each program, we determined optimal parameter settings based on the above analyses of our two data sets. For GeneMapper, to visualize how the error rate, resolution, and normalized resolution change with increasing bin width, we averaged over the nine possible parameter settings for MFL and PHT and plotted them for each bin width. For both programs, we also considered the results from the larger replicate study (see below) and trends in each of the measures of accuracy to find the optimal settings for each of PHT, MFL, and BW (GeneMapper) and PHT, MFL, SPF, and LGDP (GeneMarker).

In addition to the phylogenetic-based methods described above, the optimum bin width was independently investigated by analyzing a set of 25 pairs of replicate AFLP profiles composed of 6110 pairs of identical fragments from *Ipomoea* and *Ourisia* and measuring the average size difference (random error) between the peaks of identical fragments (raw data not shown). These peaks are known to represent identical fragments because they are from the same or replicate DNA extractions of the same tissue sample from the same individual.

Robustness of the phylogenies to changes in parameter settings.—To assess the robustness of the resulting tree to different parameter settings we constructed the majority-rule consensus tree of the bootstrap trees for each parameter setting. The sets of 90 trees from the GeneMapper character matrices and the sets of 36 trees generated from the GeneMarker character matrices for both the *Ourisia* and *Ipomoea* data sets were analyzed using consensus networks (Holland et al., 2005) as implemented in SplitsTree 4 (Huson and Bryant, 2006). For the GeneMapper data sets, we also made consensus networks of the 63 majority-rule bootstrap trees corresponding to bin width settings of 0.4 and above. Consensus networks also facilitated topological comparison of trees constructed using data sets scored with the software's default vs. optimized parameter settings and comparison of trees constructed using optimal parameter settings in GeneMarker versus GeneMapper.

RESULTS AND DISCUSSION

Measures of Accuracy

Phylogenetic resolution.—There is a wide variation in phylogenetic resolution depending on choice of parameters. Resolution scores range from 37% to 83% for *Ipomoea* and 36% to 83% for *Ourisia* (see supplementary data; available at www.systematicbiology.org for complete results; Table 1 compares the resolution scores for selected parameter settings). Importantly, in both

TABLE 1. Resolution scores and numbers of parsimony informative characters. Representative data from two different bin width settings in GeneMapper and two different detection percentage settings in GeneMarker are shown for all three minimum fragment length and peak height threshold settings for both *Ipomoea* and *Ourisia* data sets (see supplementary data for complete results; available online at www.systematicbiology.org). The maximum standard deviation for any of the resolution scores was 2.91 corresponding to a standard error of 0.29. MFL, minimum fragment length; PHT, peak height threshold; 1%, 5%, default values for local and global detection percentages (LGDP, 1%), and stutter peak filter (SPF, 5%); off,off, LGDP and SPF turned off; BW, bin width. Values for the software default settings are underlined, whereas the values we suggest are most appropriate (i.e., the overall optimal parameter settings) are indicated with an asterisk.

		GeneMarker		GeneMapper	
MFL	PHT	1%,5%	off,off	BW 0.5	BW 1.0
Ipomoea					
50	50	54% (301)	62% (284)*	77% (803)*	64% (572)
	100	56% (302)	58% (289)	66% (576)	56% (430)
	200	48% (275)	55% (267)	63% (365)	59% (293)
100	50	50% (237)	57% (220)	62% (634)	60% (463)
	100	49% (238)	47% (226)	53% (436)	45% (336)
	200	46% (212)	49% (204)	51% (264)	44% (216)
150	50	56% (160)	55% (150)	67% (459)	61% (349)
	100	49% (160)	40% (153)	49% (291)	45% (240)
	200	38% (136)	37% (132)	41% (163)	44% (141)
Ourisia					
50	50	64% (552)	70% (543)	73% (1548)	63% (1030)
	100	59% (534)	62% (528)	76% (1194)	68% (842)
	200	56% (483)	63% (485)	70% (840)	66% (627)
100	50	60% (449)	71% (442)	79% (1256)*	66% (855)
	100	56% (432)	62% (428)	77% (933)	68% (676)
	200	53% (384)	56% (387)	68% (640)	60% (485)
150	50	62% (328)	72% (328)*	74% (948)	59% (671)
	100	59% (312)	61% (313)	66% (662)	67% (498)
	200	50% (266)	54% (272)	54% (431)	56% (338)

programs, default settings are not optimal with respect to phylogenetic resolution. The highest (and the lowest) resolution scores were found by tuning scoring parameters away from the default settings. Default settings in GeneMapper (PHT 100, MFL 100, BW 1.0; shown in Table 1) give resolution scores of 45% (*Ipomoea*) and 68% (*Ourisia*). The highest resolution scores for GeneMapper are 83% for both data sets with parameter settings PHT 50, MFL 50, BW 0.4 (*Ipomoea*) and MFL 50, PHT 50, BW 0.2 (*Ourisia*). Default settings in GeneMarker (MFL 100, PHT 100, LGDP 1%, SPF 5%; shown in Table 1) give resolution scores of 49% (*Ipomoea*) and 56% (*Ourisia*). The highest resolution scores for GeneMarker are 62% (*Ipomoea*) and 72% (*Ourisia*), which occur with parameter settings PHT 50, MFL 50, LGDP off, SPF off (*Ipomoea*) and PHT 50, MFL 150, LGDP 1%, SPF off (*Ourisia*).

The normalized resolution scores did not vary as widely as the non-normalized resolution scores (Table 1; see supplementary data for complete results; Table 2 compares scores from selected parameter settings). This result indicates that most of the differences in resolution could be explained by a difference in the number of characters; i.e., as expected the presence of more characters leads to higher resolution. Assuming that higher resolution is correlated with higher accuracy, this means it is not always best to strive for error-free data sets at the expense of throwing away many characters. The approach of Althoff et al. (2007) that advocates eliminating all error may actually be counterproductive for phylogenetic applications of AFLP.

TABLE 2. Normalized resolution scores. Representative data from two different bin width settings in GeneMapper and two different detection percentage settings in GeneMarker are shown for all three minimum fragment length and peak height threshold settings for both *Ipomoea* and *Ourisia* data sets (see supplementary data for complete results). The maximum standard deviation for any of these values was 6.05 corresponding to a standard error of 0.61. For abbreviations, see Table 1. Values for the software default settings are underlined, whereas the values we suggest are most appropriate (i.e., the overall optimal parameter settings) are indicated with an asterisk.

		GeneMarker		GeneMapper	
MFL	PHT	1%,5%	off,off	BW 0.5	BW 1.0
Ipomoea					
50	50	46%	49%*	46%*	43%
	100	46%	46%	44%	42%
	200	40%	41%	44%	45%
100	50	47%	50%	44%	42%
	100	43%	43%	41%	39%
	200	40%	41%	40%	39%
150	50	55%	53%	45%	43%
	100	46%	38%	39%	41%
	200	39%	37%	35%	41%
Ourisia					
50	50	51%	55%	45%	46%
	100	49%	50%	48%	49%
	200	45%	48%	49%	50%
100	50	55%	62%	48%*	47%
	100	52%	56%	52%	52%
	200	47%	49%	48%	50%
150	50	60%	70%*	47%	47%
	100	58%	60%	50%	54%
	200	50%	53%	48%	51%

One interesting aside is that, in general, GeneMapper gives better non-normalized phylogenetic resolution than GeneMarker but worse normalized resolution. Thus, comparing data matrices from the two programs with identical MFL and PHT settings (and using the default settings for SPF and LGDP in GeneMarker, and BW 0.5 in GeneMapper) shows that GeneMapper data matrices contain from 1.2 to 2.9 times as many parsimony-informative characters as the equivalent GeneMarker data set. This implies that GeneMarker creates character matrices with higher quality characters than GeneMapper, but because the GeneMapper data sets contain more characters they give more highly resolved trees.

Correct assignment of replicates.—For both data sets, the number of replicate pairs that were correctly grouped as sister taxa (see supplementary data) provide evidence that BW values below 0.4 are not optimal. In addition, in GeneMapper using a PHT of 100 always gave more correctly assigned replicate pairs than PHT 50 or 200, although PHT 50 was almost as good. For the *Ourisia* data set all six replicate pairs were correctly grouped in all data sets from both programs except for GeneMapper data sets with bin width settings of 0.1 to 0.3 where one or two of the six replicate pairs were sometimes not grouped together. For the *Ipomoea* data set using GeneMapper, bin widths of 0.4 or 0.5 gave the most correct replicate pairs—3.33 (out of 5) on average. For the *Ipomoea* data set, using GeneMarker, all settings with PHT 100 or 200 gave only two correct replicates out of five, and settings with PHT 50 gave three correct replicate pairs (9 times) or two correct replicate pairs (3 times).

For the *Ipomoea* data, many settings incorrectly group the three replicate pairs 157 cv. Toka Toka Gold A/B, Y-622 Peru A/B, and Y-680 Colombia A/B. In fact, the replicate pair 157 cv. Toka Toka Gold A/B was never correctly recovered by either program for any parameter settings—one or other of the pair always grouped more closely with “158 cv. Mary Anne.” The cultivar “Mary Anne” is a recently derived vegetative mutant of “Toka Toka Gold” so it is perhaps not surprising that these two cultivars are indistinguishable based on AFLP. Analysis of the distance matrices (data not shown) shows that for each of the three replicate pairs that sometimes group incorrectly there is a third taxon that is also genetically very close. In contrast, in the *Ourisia* data, where the replicates are usually all correctly assigned, the distance between replicates is in the same range as in the *Ipomoea* data, but there are no other taxa that are genetically very close to the replicates. This is also seen in the *Ipomoea* data set in replicates of the outgroup, *I. tiliacea* K233-1 A and B, which are always grouped correctly in all analyses. The *Ipomoea* data set comprises cultivars of a single species plus the outgroup, and it appears that there is insufficient signal in the AFLP data to distinguish some ingroup accessions. In *Ipomoea* (for the optimal GeneMarker data set) the distances between replicates were 0.07, 0.08, 0.09, 0.10, and 0.20; the distances between a replicate and its closest nonreplicate ranged from 0.08 to 0.33 with a me-

TABLE 3. Replicate error rates for representative data sets. Representative data from two different bin width settings in GeneMapper and two different detection percentage settings in GeneMarker are shown for all three minimum fragment length and peak height threshold settings for both *Ipomoea* and *Ourisia* data sets (see supplementary data for complete results). For abbreviations, see Table 1. Values for the software default settings are underlined, whereas the values we suggest are most appropriate (i.e., the overall optimal parameter settings) are indicated with an asterisk.

		GeneMarker		GeneMapper	
MFL	PHT	1%,5%	off,off	BW 0.5	BW 1.0
Ipomoea					
50	50	12%	11%*	15%*	14%
	100	13%	13%	14%	13%
	200	14%	13%	15%	14%
100	50	11%	10%	15%	13%
	100	13%	13%	14%	12%
	200	14%	14%	16%	15%
150	50	11%	9%	15%	13%
	100	13%	13%	16%	13%
	200	15%	15%	18%	16%
Ourisia					
50	50	12%	13%	11%	13%
	100	12%	13%	9%	10%
	200	10%	11%	8%	8%
100	50	11%	13%	11%*	12%
	100	11%	12%	9%	9%
	200	10%	10%	7%	7%
150	50	10%	11%*	9%	11%
	100	9%	10%	7%	7%
	200	8%	8%	6%	6%

dian of 0.11. Similarly, in *Ourisia* the distances between replicates were 0.07, 0.07, 0.09, 0.11, 0.12, and 0.23; but in contrast to *Ipomoea*, the distances between a replicate and its closest nonreplicate tended to be larger, ranging from 0.13 to 0.34 with a median of 0.22.

Error rates.—The error rates range from 9% to 18% (*Ipomoea*) and 6% to 13% (*Ourisia*); see supplementary data for complete results (Table 3 shows the replicate pair error rates for both programs at selected parameter settings). The observed error rates are higher than those previously reported for AFLP data sets of (2% to 5%; see Bonin et al., 2004, and references therein). However, as discussed above (see Methods), we should be cautious regarding interstudy comparisons of error rates because the error rates may be affected by the level of divergence among the individuals included in the study, the number of individuals in the data set, and both errors resulting from the raw AFLP profiles themselves (e.g., PCR errors) and those resulting from the scoring process (and the type of procedure and software employed).

We found that all of these factors have affected the error rates in our data sets, some to a greater degree than others. First, the largely intraspecific *Ipomoea* data sets give higher error rates overall than the interspecific *Ourisia* data sets, which suggests that lower divergence among samples results in higher error rates. We tested this further by comparing the standard (Euclidean) and alternative (Jaccard) error rates. For the GeneMapper data sets, standard error rates are almost constant in the *Ourisia* data sets, whereas error rates are higher for smaller BW than larger BW in the *Ipomoea* data set as would be predicted (Fig. 4). To check if the difference in error rates and the difference in this trend in error rates in the *Ourisia* and *Ipomoea* data sets were due to an increased number of (0,0) calls masking an increased rate of erroneous (1,0) or (0,1) calls, we calculated the alternative (Jaccard) error rate. As shown in Figures 4a and b, this alternative error rate decreases sharply from a bin width of 0.1 to 0.5, after which it flattens off at around 30% for both data sets. The *Ipomoea* data set has an average standard error rate of 15% compared to 10% for the *Ourisia* data set. However, when the alternative error rate is used, the average error rate for *Ipomoea* is 38% compared to 40% for *Ourisia*, suggesting that the apparently higher standard error rate in *Ipomoea* may not be "real," but is instead a result of fewer (0,0) calls in the denominator.

Secondly, the higher error rates found in our study are also partly due to the small size (30 individuals) of our data sets. The recalculated error rates for the larger data sets of both *Ipomoea* and *Ourisia* scored with optimal parameter settings were indeed lower compared to the 30-taxon data sets. For the *Ipomoea* data set, the error rate dropped from 15% to 9% (GeneMapper) and from 10% to 9% (GeneMarker). For the *Ourisia* data set, the error rate dropped from 11% to 8% (GeneMapper) and from 12% to 11% (GeneMarker). Nevertheless, even though the error rates are lower when many more individuals are included, they are still not within the range reported in Bonin et al. (2004), which suggests that the nature of the errors is also an important factor. Finally, using the program ReplicateError, we were able to lower the error rate in the optimized GeneMapper data sets of *Ipomoea* and *Ourisia* from 15% (uncorrected) to 5% (corrected) and 11% (uncorrected) to 4% (corrected), respectively. This suggests that the majority of errors (the difference between the corrected and uncorrected rates) in our data set are scoring errors, whereas the resulting corrected error rates approximate the number of PCR errors. Because ReplicateError only locates errors between replicates, it is not possible to use this program to reduce errors in the data set as a whole, but it does indicate that there is significant potential to improve the accuracy of automated scoring so that error rates are comparable to those derived from manually scored data.

In summary, it is likely that our higher error rates are due in large part to the combined effects of smaller data sets and a fully automated scoring procedure. Including the (0,0) term in the denominator has a significant effect on error rates and makes comparison of error rates between data sets unreliable, especially if they contain different numbers of taxa and/or taxa with varying amounts of genetic diversity. The analyses using ReplicateError revealed that the error rate was significantly increased by errors introduced during the automated scoring process. Although it could be argued that this result supports manual scoring (or at least manual editing of automatically scored data), we still think that automated scoring is preferable because it is more time efficient, it makes it easier to maintain consistency in large data sets, and it removes both subjectivity and

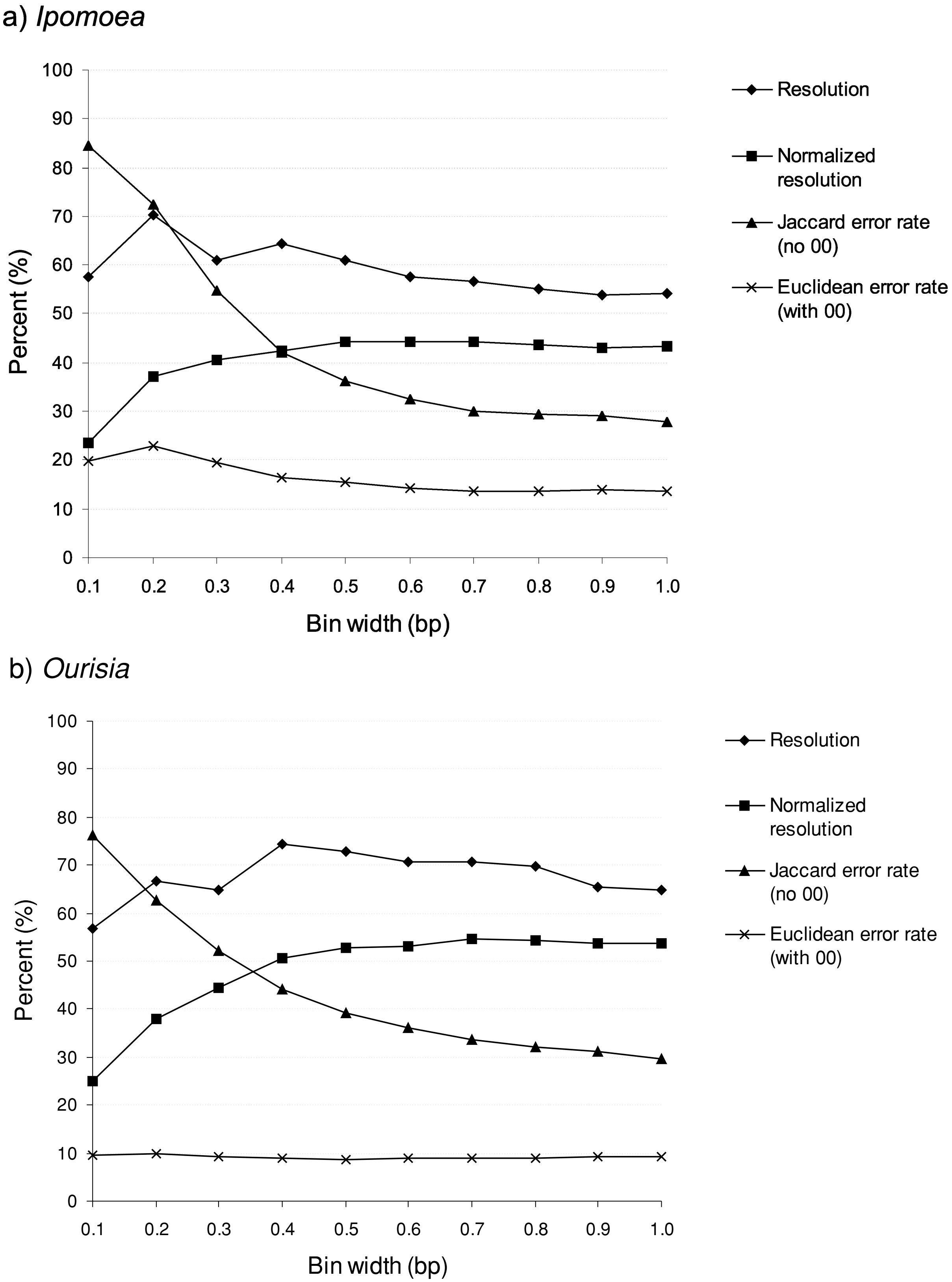


FIGURE 4. Resolution, normalized resolution, and error rate versus bin width for the GeneMapper analysis of the (a) *Ipomoea* data and (b) *Ourisia* data. All measures have been averaged over the nine possible settings for minimum fragment length (MFL) and peak height threshold (PHT).

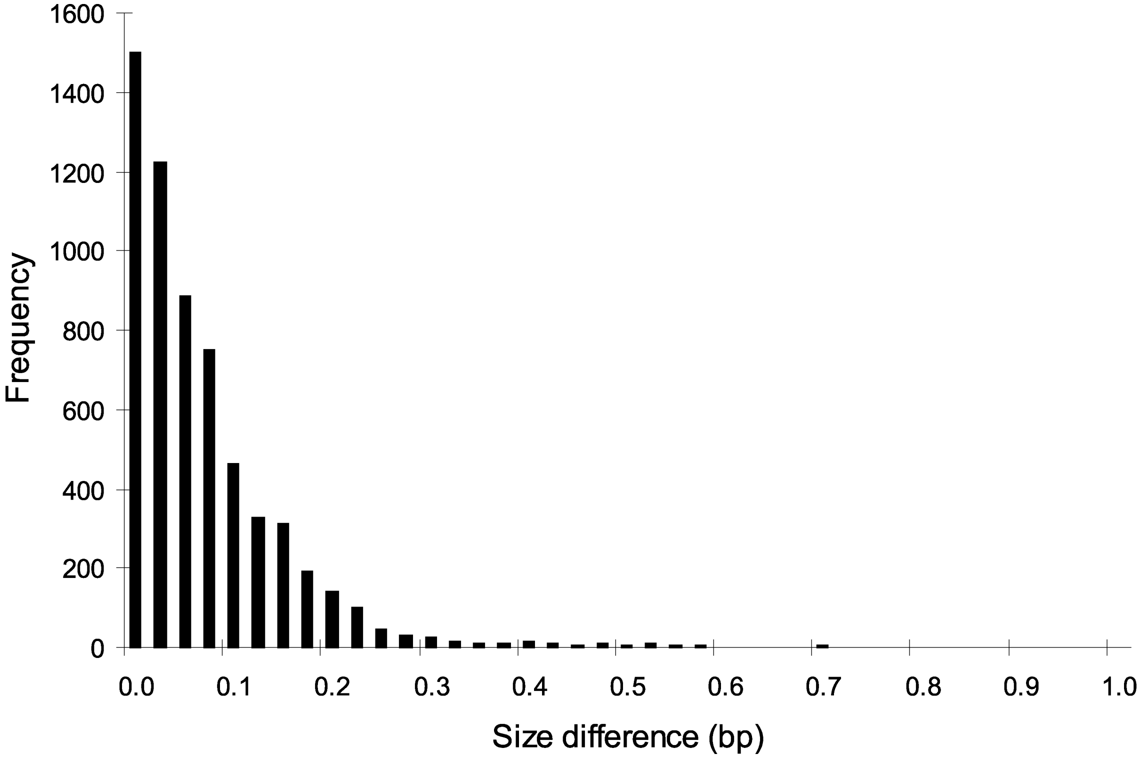


FIGURE 5. Size difference between identical alleles of replicate pairs. The size difference between 6110 pairs of identical alleles (fragments) from 25 replicate AFLP profiles was measured to determine the average sizing error between truly identical fragments. Together with error rates and resolution measures, these results help to determine the optimum bin width. The mean size difference between identical alleles was 0.08 bp with 99% of values falling within 0.42 bp and 99.9% falling within 0.66 bp. These results suggest that bin width can be set to less than 1.0 bp, without greatly increasing the risk of splitting identical alleles into separate characters. These values are for the experimental setup we used (see Methods) and may need to be determined empirically for other capillary instruments.

the potential to introduce bias into data sets (although there are also methods to remove bias when scoring manually).

Finally, Figure 5 shows the results of the larger replicate study, which demonstrates that almost all identical fragments fall within 0.4 bp of each other. This suggests that lowering the bin width setting below 1 bp could be beneficial as it would introduce only a small number of extra errors in the character matrix and may help distinguish between nonidentical fragments that differ in mobility by less than 1 bp. There is some measurable distance even between identical fragments, which means it is always possible for truly homologous fragments to fall into different bins. Using the empirical distribution of observed distance between identical fragments (Fig. 5), we did a simulation to estimate how many errors of this kind we would expect for different bin width settings. For each simulated pair of identical fragments, it was assumed that the position of the leftmost fragment of the pair was uniformly distributed within the bin; the distance to the rightmost fragment of the pair was then sampled from the empirical distribution (Fig. 5). We then recorded if the rightmost fragment was still in the same bin. This was repeated for 1,000,000 simulated fragments for each bin width between 0.1 and 1 (in steps of 0.1). The proportion of expected errors of this kind is shown in Table 4. Table 4 shows that if the bin width drops below 0.5 bp, the error rate starts to rise steeply.

TABLE 4. The empirical distribution of differences between identical fragments (Fig. 5) was used to simulate the expected number of errors where identical fragments are allocated to different bins for a range bin widths. Each error value is a proportion based on 1,000,000 random fragment pairs.

Bin width (bp)	Errors
0.1	0.59
0.2	0.37
0.3	0.26
0.4	0.20
0.5	0.16
0.6	0.14
0.7	0.12
0.8	0.10
0.9	0.09
1.0	0.08

Optimal Parameter Settings

GeneMapper.—In general, optimal parameter settings in GeneMapper for the two data sets were PHT 50, MFL 50 (*Ipomoea*) or 100 (*Ourisia*), and BW 0.5. For the *Ipomoea* and *Ourisia* data sets generated using GeneMapper, we can see how the error rate, resolution, and normalized resolution change with increasing bin width (Fig. 4). The values in these figures have been averaged over the nine possible parameter settings for MFL and PHT. For the *Ourisia* data, resolution peaks at a bin width of 0.4 bp, with a smaller peak at 0.2 bp. The *Ipomoea* data has a peak at 0.2 with a smaller peak at 0.4.

We propose that a BW of 0.5 is optimal for both data sets, for the following reasons. The normalized resolution scores show that the phylogenetic quality of the characters is increasing up to a bin width of 0.5, after which it is fairly stable. Taken together, the results from the larger replicate study (Fig. 5), the assignment of replicate pairs, and the trend in the normalized resolution and alternative error rate (Fig. 4) suggest that when scoring ABI 3730-derived AFLP data using GeneMapper, lowering the BW setting to 0.5 bp is beneficial. Although some errors are introduced, this is outweighed by the positive effect of having more characters. Decreasing the bin width below 0.5 bp results in even more characters but splits apart too many characters that are identical plus-alleles of the same locus.

A PHT setting of 50 appears to be optimal in both data sets. Although PHT 50 gives slightly higher replicate error rates in GeneMapper than the higher settings (Table 3), in most cases the resolution scores (Table 1) and normalized resolution scores (Table 2) are better for both data sets at this setting—especially at a bin width setting of 0.5. In contrast, the optimal setting for MFL appears to differ between the two data sets. The highest resolution (Table 1) and normalized resolution (Table 2) scores are found at MFL 50 for *Ipomoea* and MFL 100 for *Ourisia*. The error rate (Table 3) is not greatly affected by the choice of MFL, especially for *Ipomoea*.

GeneMarker.—Optimal parameter settings in GeneMarker for the two data sets were PHT 50, MFL 50 (*Ipomoea*) or 150 (*Ourisia*), and SPF off. Resolution, normalized resolution, and number of parsimony informative characters mostly increase with decreasing PHT (Tables 1 and 2). In most cases, for both data sets, setting the PHT at 50 gives better resolution than setting it at 100 or 200 (Table 1). Error rates decrease with decreasing PHT for *Ipomoea,* but the opposite trend is seen for *Ourisia* (Table 3). General trends regarding MFL are not as clear; for the *Ipomoea* data set, the highest resolution was found at a setting of 50, but for the *Ourisia* data set, the highest resolution was found at a setting of 150. Error rates were not affected by MFL for *Ipomoea* but decreased slightly with increasing MFL for *Ourisia* (Table 3). Turning off the stutter peak filter (SPF) caused a marked increase in resolution, but this could be artefactual. For instance, by including all the stutter peaks, support for some splits could be inflated. In contrast, LGDP had only a small effect on the resolution of the resulting trees.

Robustness of the Phylogenies to Changes in Parameter Settings

Consensus networks are a very useful way to visualize the robustness of the phylogenies to AFLP scoring parameter settings and to different software (in this case, GeneMapper and GeneMarker; see Supplementary Figs. 2 and 3, available at www.systematicbiology.org), and specifically examine the difference in phylogeny reconstruction of default versus optimized settings. Figure 6 shows the consensus network of the 63 GeneMapper majority-rule bootstrap consensus trees corresponding to bin width settings of 0.4 and above, and Figure 7 shows the consensus network of all 36 GeneMarker majority-rule bootstrap consensus trees. Significantly, many parts of the phylogenies in both Figures 6 and 7 are robust to parameter choice, and where the trees do differ the boxes in the consensus networks indicate specific areas of conflict due to local rearrangements rather than taxa shifting their position in the tree dramatically. One exception to this is the data sets from GeneMapper with low bin width settings. As indicated by Figure 4 and discussed above, setting the bin width lower than 0.4 probably creates many errors in the character matrix and may lead to the reconstruction of inaccurate trees. Indeed, this appears to be the case in our data sets, as the consensus networks constructed for the majority-rule bootstrap trees for all 90 GeneMapper data sets including those with BW below 0.4 (Supplementary Fig. 1; available at www.systematicbiology.org) show much more conflict than Figures 6 and 7. In spite of this, the consensus networks encouragingly show that regardless of parameter settings, the data sets are converging on very similar topologies whose accuracy is corroborated by independent sources of data (see below).

To investigate whether the two programs are converging on similar topologies, consensus networks comparing the majority-rule consensus trees using default parameter settings and using optimized parameter settings for GeneMapper versus GeneMarker were constructed (Supplementary Figs. 2 and 3). We emphasize that the intention of this exercise is not to compare the performance of GeneMapper versus GeneMarker per se, but to show the degree of topological congruence. For default settings (Supplementary Fig. 3), there are two conflicting edges in the *Ipomoea* data set and three conflicting edges in the *Ourisia* data set. For optimized settings (Supplementary Fig. 2) there are six areas of conflict in the *Ipomoea* data set (five affecting just single edges and one more complex area of conflict) and three conflicting edges in the *Ourisia* data set. In all cases, the conflict is confined to a few, local areas of the tree and does not represent large differences between the topologies recovered using the two programs. Therefore, in addition to differing parameter settings, topologies are also robust to software choice. Because very similar topologies were recovered using different software packages that use diverse algorithms and methods, this result provides further evidence that automated scoring of AFLP profiles results in accurate and robust phylogenies.

Finally, consensus networks also show that optimized parameter settings consistently show an increase in the number of edges with >50% bootstrap support relative to default settings. Figure 8 illustrates the difference in the majority-rule consensus tree between default parameter settings in GeneMapper (PHT 100, MFL 100, BW 1.0) versus optimized settings (PHT 50, MFL 50 or 100, BW

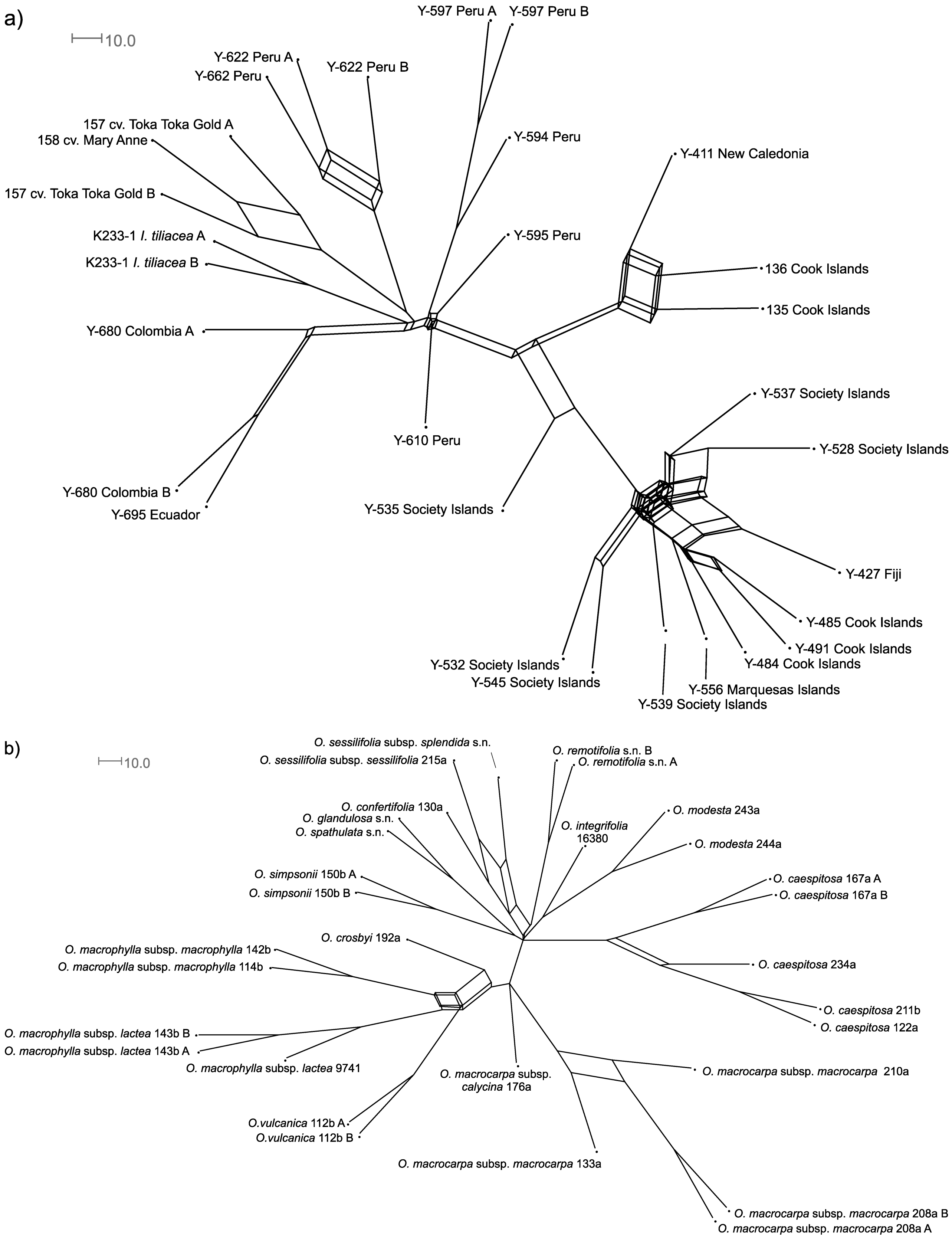


FIGURE 6. Consensus network of the 63 GeneMapper majority-rule bootstrap consensus trees that correspond to parameter settings with bin width >0.3 (i.e., 0.4 to 1.0) for (a) *Ipomoea* data and (b) *Ourisia* data. The consensus networks show all splits (edges) that appear in any of the 63 trees. Parallel edges represent the same split, and edge length is proportional to the number of trees that split appears in. Boxes represent areas of conflict among the trees generated using different parameter settings.

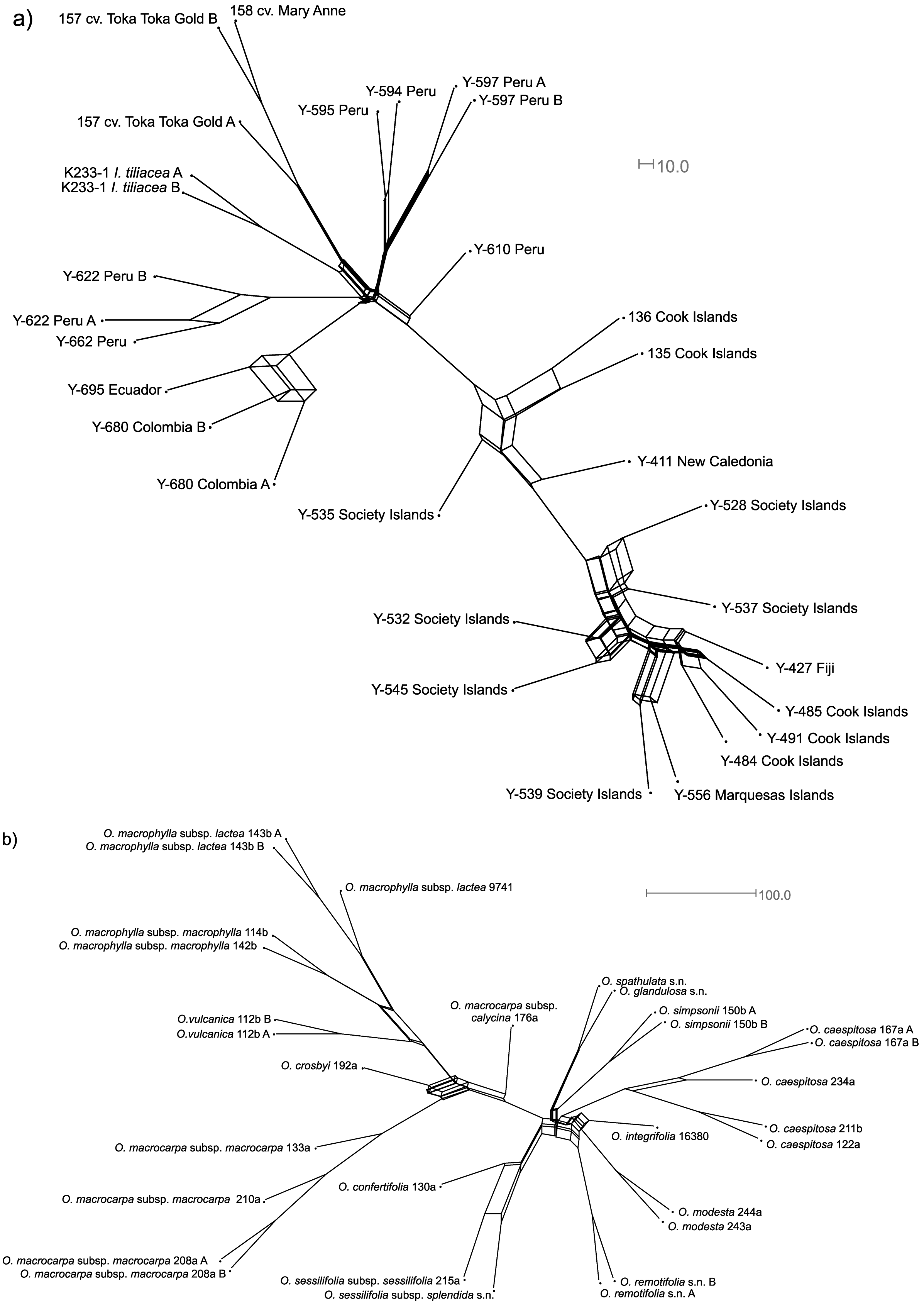


FIGURE 7. Consensus network of the 36 GeneMarker majority-rule bootstrap consensus trees for (a) *Ipomoea* data and (b) *Ourisia* data showing all splits (edges) that appear in any of the 36 trees. Parallel edges represent the same split, and edge length is proportional to the number of trees that split appears in. Boxes represent areas of conflict among the trees generated using different parameter settings.

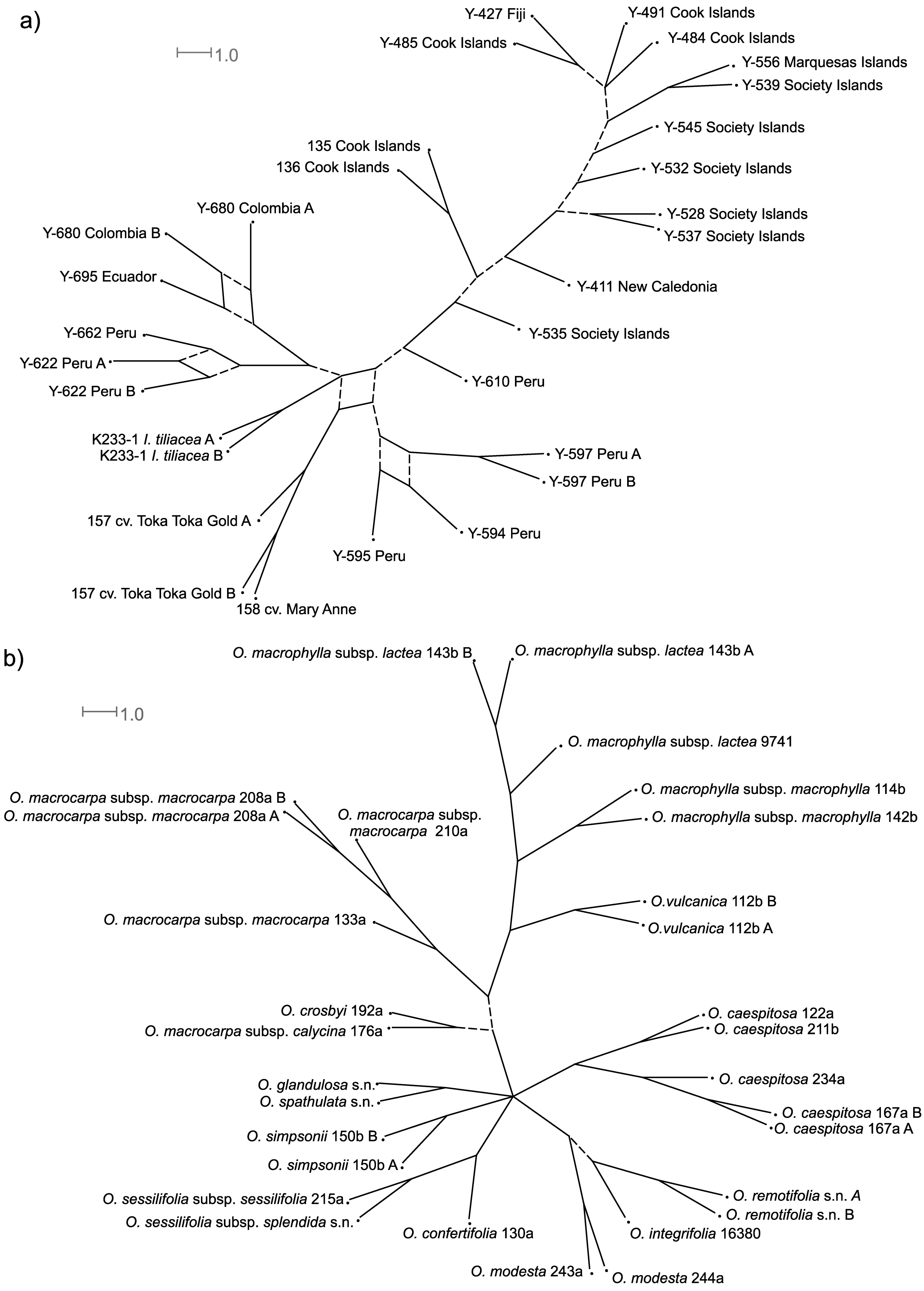


FIGURE 8. Majority-rule consensus networks of the default and optimized GeneMapper trees for (a) *Ipomoea* data and (b) *Ourisia* data. Parallel edges represent the same split, and edge length is proportional to the number of trees that split appears in. The dashed edges indicate splits that only appear in the optimized GeneMapper trees. Boxes represent areas of conflict between the trees generated using default vs. optimized parameter settings.

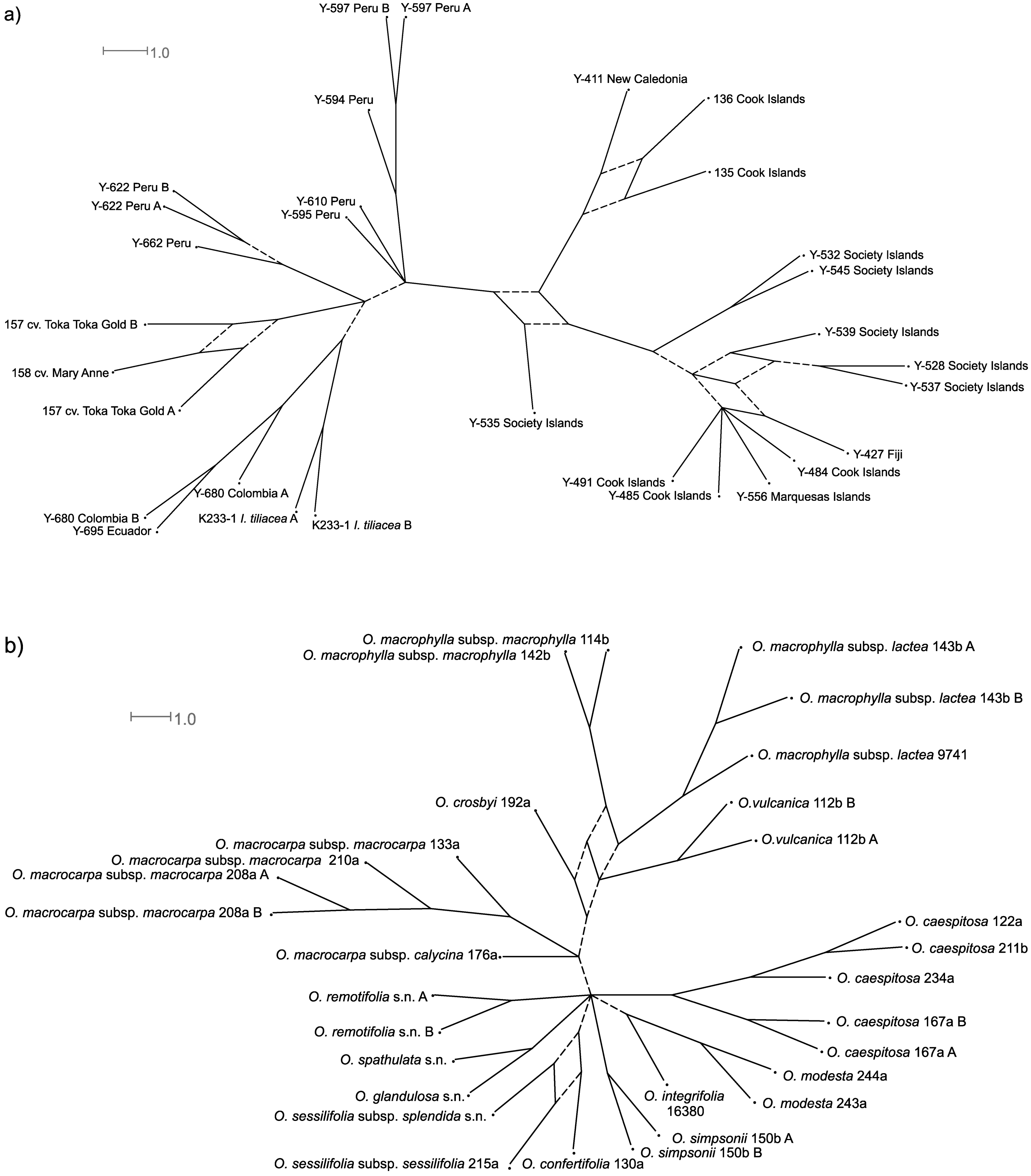


FIGURE 9. Majority-rule consensus networks of the default and optimized GeneMarker trees for (a) *Ipomoea* data and (b) *Ourisia* data. Parallel edges represent the same split, and edge length is proportional to the number of trees in which that split appears. The dashed edges indicate splits that only appear in the optimized GeneMarker trees. Boxes represent areas of conflict between the trees.

0.5). Edges shown with dashed lines appear only in the trees built with optimized parameter settings. Optimization of scoring parameters had the greatest effect in the *Ipomoea* GeneMapper data set, where the number of internal edges with >50% bootstrap increased from 14 to 25 (out of a possible maximum of 27) by optimizing scoring parameters. Along with these 11 new edges—that importantly do not conflict with the default setting tree—four additional edges are changed. For the *Ourisia* data, using optimized settings gives three new edges with bootstrap support above 50% that do not conflict with the default setting tree; it does not change any edges. The same plot was done for GeneMarker default (MFL 100, PHT 100, LGDP 1%, SPF 5%) versus optimized (MFL 50 or 150, PHT 50, LGDP off, SPF off) parameter settings (Fig. 9). For the *Ipomoea* data, using optimized settings gives five new edges with bootstrap support above 50% that do not conflict with the default setting tree; it also changes four edges. For the *Ourisia* data, using optimized settings gives four new edges with bootstrap support above 50% that do not conflict with the default setting tree; it also changes two edges. Thus, even though the topologies are largely robust to scoring parameter settings and choice of software, these consensus networks show that optimized parameter settings can improve resolution and increase support for the resulting phylogenetic trees relative to default parameter settings.

Additional Independent Evidence

In addition to resolution scores, another way to investigate the accuracy of trees constructed using the optimum parameter settings is to compare the tree topologies to other sources of data, such as morphological, ecological, or other characters. In the case of *Ourisia*, phylogenetic analyses of the AFLP data regardless of parameter settings are congruent with morphological, distributional, and taxonomic information (see Meudt, 2006; Meudt and Simpson, 2006). The optimal parameter settings, however, provide more resolution and more accuracy regarding species relationships than trees built with default settings, especially for data sets reconstructed in GeneMarker (Fig. 9b). In the case of *Ipomoea*, the AFLP-derived phylogenies are also congruent with other sources of information, in this case primarily morphological, horticultural, and anthropological evidence (see Green, 2005).

Conclusions

In any phylogenetic study using AFLP data, the main aim is to recover an accurate species phylogeny or to test a phylogenetic hypothesis. There has been much discussion in the literature regarding appropriate techniques for phylogenetic analysis of AFLP data (Albach 2007; Bonin et al. 2007; Meudt and Clarke, 2007, and references therein); by contrast, very little has been done with respect to the scoring of AFLP data and the effect this has on downstream analyses. To our knowledge, our study is the first quantitative, objective, and thorough investigation on the effect of different automated scoring parameter settings on phylogenetic resolution.

Optimizing the parameter settings for automated AFLP scoring significantly increased phylogenetic resolution in this study, allowing relationships to be resolved that were obscured when using default scoring parameters. We predict that similar improvements in resolution can be obtained in other phylogenetic studies and recommend that automated scoring parameters are optimized wherever possible. Interestingly, we found that it was not always best to choose the character matrices with the lowest error rates, as the benefits of increasing the number of characters could outweigh some reduction in character quality.

The optimal settings differed for the *Ourisia* and *Ipomoea* data sets, suggesting that, for at least some parameters, there are not universal optimal settings. The best settings for minimum fragment length varied for the two data sets in terms of both resolution and error rates. In contrast, there is a case for a universally optimal setting for bin width in GeneMapper; several lines of evidence suggest that 0.5 bp is a good choice. This is supported for both data sets by the high-resolution scores, the greater number of replicate pairs appearing as sister taxa, and the fact that almost all peaks of truly identical fragments in the replicate study fell within 0.5 bp. However, we caution that these results are based on analysis of data from two data sets that were run on the same capillary instrument. Therefore, rather than suggesting particular parameter settings, we recommend that users of AFLP use a similar procedure to that described here (Fig. 3) to investigate the effect of changing AFLP scoring parameters' settings on phylogenetic resolution, assignment of replicate pairs, and error rates. As more results are built up for different data sets, it will be possible to determine if there are some universally good settings. Nevertheless, it is clear that reducing bin widths from the default 1.0-bp setting is potentially beneficial.

One potential shortcoming of using the same data set to optimize parameter settings via bootstrap resolution and to make phylogenetic estimates is that this could upwardly bias support values in the phylogenetic analysis. A way to mitigate against this effect could be to use only a subset of the taxa of interest to optimize the parameters (indeed, this is what we have done here).

Error rates found in this study were higher than those previously reported for data sets generated using semiautomated scoring (2% to 5%; Bonin et al., 2004). We provide evidence that the calculation used to quantify genotyping error rate depends on the number of taxa in the data set and the genetic distance between them, but these effects are not sufficient to explain the difference in error rates between semiautomated and automated scoring. The results from ReplicateError suggest that the majority of errors were scoring errors (rather than PCR errors). This suggests that the incorporation of improved scoring algorithms into current software packages such as GeneMapper and GeneMarker would further increase their power and usefulness.

Despite error rates that are higher than ideal, automated scoring still produces character matrices that are phylogenetically informative and where most or all replicates are correctly assigned. When parameter settings are chosen carefully, character matrices can be produced using automatic scoring that result in well-resolved trees. As well as its application to phylogenetics, we predict that optimizing automated AFLP scoring parameters will provide increased resolution in other important applications of the technique such as linkage mapping and population genetics. In these cases, different measures of accuracy and resolution will be required, although in all applications of the AFLP technique measures of error rate from replicate samples are critical. Future studies should focus on calculating and publishing error rates, optimizing parameter settings prior to analysis, improving automated scoring algorithms, as well as taking a step back and thoroughly assessing the appropriateness of AFLP for phylogenetic reconstruction.

Acknowledgments

We thank Lorraine Berry (Allan Wilson Centre Genome Service, Massey University) for performing the capillary electrophoresis, Morore Piripi for performing a literature survey, Matt Phillips for suggesting the resolution measure, and Warwick Allen for developing the Genotyping Utilities GenoTyper Rearranger and ReplicateError. For helpful comments on the manuscript, we thank Simon Joly, David Penny, Vaughan Symonds, Lacey Knowles, and the anonymous reviewers. B.R.H acknowledges financial support from the FoRST Bridge to Employment scheme, A.C.C. from Massey University and the Allan Wilson Centre for Molecular Ecology and Evolution, and H.M.M. from the National Science Foundation International Research Fellowship Program. GeneMapper® is a registered trademark of Applied Biosystems and its subsidiaries in the USA and other countries. GeneMarker® is a registered trademark of SoftGenetics LLC.

References

Albach, D. C. 2007. Amplified fragment length polymorphisms and sequence data in the phylogenetic analysis of polyploids: Multiple origins of *Veronica cymbalaria* (Plantaginaceae). New Phytol. 176:481–498.

Althoff, D. M., M. A. Gitzendanner, and K. A. Segraves. 2007. The utility of amplified fragment length polymorphisms in phylogenetics: A comparison of homology within and between genomes. Syst. Biol. 56:477–484.

Bensch, S., and M. Åkesson. 2005. Ten years of AFLP in ecology and evolution: Why so few animals? Mol. Ecol. 14:2899–2914.

Bikandi, J., R. San Millán, A. Rementeria, and J. Garaizar. 2004. *In silico* analysis of complete bacterial genomes: PCR, AFLP-PCR and endonuclease restriction. Bioinformatics 20:798–799.

Bonin, A., E. Bellemain, P. B. Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet. 2004. How to track and assess genotyping errors in population genetics studies. Mol. Ecol. 13:3261–3273.

Bonin, A., D. Ehrich, and S. Manel. 2007. Statistical analysis of amplified fragment length polymorphism data: A toolbox for molecular ecologists and evolutionists. Mol. Ecol. 16:3737–3758.

Carstens, B. C., and L. L. Knowles. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from *Melanoplus* grasshoppers. Syst. Biol. 56:400–411.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:762–768.

Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

Després, L., L. Gielly, B. Redoutet, and P. Taberlet. 2003. Using AFLP to resolve phylogenetic relationships in a morphologically diversified plant species complex when nuclear and chloroplast sequences fail to reveal variability. Mol. Phylogenet. Evol. 27:185–196.

Green, R. C. 2005. Sweet potato transfers in Polynesian prehistory. Pages 43–62 *in* The sweet potato in Oceania: A reappraisal. Ethnology Monographs 19/Oceania Monograph 56 (C. Ballard, P. Brown, R. M. Bourke, and T. Harwood, eds.). Ethnology, University of Pittsburgh, Pittsburgh, Pennsylvania; United States/Oceania Publications; University of Sydney, Sydney, New South Wales, Australia.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

Holland, B. R., F. Delsuc, and V. Moulton. 2005. Visualizing conflicting evolutionary hypotheses in large collections of trees: Using consensus networks to study the origins of placentals and hexapods. Syst. Biol. 54:66–76.

Hollingsworth, P. M., and R. A. Ennos. 2004. Neighbour joining trees, dominant markers and population genetic structure. Heredity 92:490–498.

Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Kilian, B., H. Özkan, O. Deusch, S. Effgen, A. Brandolini, J. Kohl, W. Martin, and F. Salamini. 2007. Independent wheat B and G genome origins in outcrossing *Aegilops* progenitor haplotypes. Mol. Biol. Evol. 24:217–227.

Koblmüller, S., N. Duftner, K. M. Sefc, M. Aibara, M. Stipacek, M. Blanc, B. Egger, and C. Sturmbauer. 2007. Reticulate phylogeny of gastropod-shell-breeding cichlids from Lake Tanganyika—The result of repeated introgressive hybridization. BMC Evol. Biol. 7:7.

Koopman, W. J. M. 2005. Phylogenetic signal in AFLP data sets. Syst. Biol. 54:197–217.

Kosman, E., and K. J. Leonard. 2005. Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. Mol. Ecol. 14:415–424.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Marhold, K., J. Lihová, M. Perný, and W. Bleeker. 2004. Comparative ITS and AFLP analysis of diploid *Cardamine* (Brassicaceae) taxa from closely related polyploid complexes. Ann. Bot. 93:507–520.

Mendelson, T. C., and K. L. Shaw. 2005. Rapid speciation in an arthropod. Nature 433:375–376.

Meudt, H. M. 2006. Monograph of *Ourisia* (Plantaginaceae). Syst. Bot. Monogr. 77:1–188.

Meudt, H. M., and A. C. Clarke. 2007. Almost forgotten or latest practice? AFLP applications, analyses and advances. Trends Plant Sci. 12:106–117.

Meudt, H. M., and B. B. Simpson. 2006. The biogeography of the austral, subalpine genus *Ourisia* (Plantaginaceae) based on molecular phylogenetic evidence: South American origin and dispersal to New Zealand and Tasmania. Biol. J. Linn. Soc. 87:479–513.

Mueller, U. G., and L. L. Wolfenbarger. 1999. AFLP genotyping and fingerprinting. Trends Ecol. Evol. 14:389–394.

Pellmyr, O., K. A. Segraves, D. M. Althoff, M. Balcázar-Lara, and J. Leebens-Mack. 2007. The phylogeny of yuccas. Mol. Phylogenet. Evol. 43:493–501.

Pelser, P. B., B. Gravendeel, and R. van der Meijden. 2003. Phylogeny reconstruction in the gap between too little and too much divergence: The closest relatives of *Senecio jacobaea* (Asteraceae) according to DNA sequences and AFLPs. Mol. Phylogenet. Evol. 29:613–628.

Pompanon, F., A. Bonin, E. Bellemain, and P. Taberlet. 2005. Genotyping errors: Causes, consequences and solutions. Nat. Rev. Genet. 6:847–859.

Qin, L., P. Prins, J. T. Jones, H. Popeijus, G. Smant, J. Bakker, and J. Helder. 2001. GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP. Nucleic Acids Res. 29:1616–1622.

Spooner, D. M., K. McLean, G. Ramsay, R. Waugh, and G. J. Bryan. 2005a. A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc. Natl. Acad. Sci. USA 102:14694–14699.

Spooner, D. M., I. E. Peralta, and S. Knapp. 2005b. Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum* L. section *Lycopersicon* (Mill.) Wettst.]. Taxon 54:43–61.

Sullivan, J. P., S. Lavoué, M. E. Arnegard, and C. D. Hopkins. 2004. AFLPs resolve phylogeny and reveal mitochondrial introgression within a species flock of African electric fish (Mormyroidea: Teleostei). Evolution 58:825–841.

Swofford, D. L. 2003. PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates: Sunderland, Massachusetts.

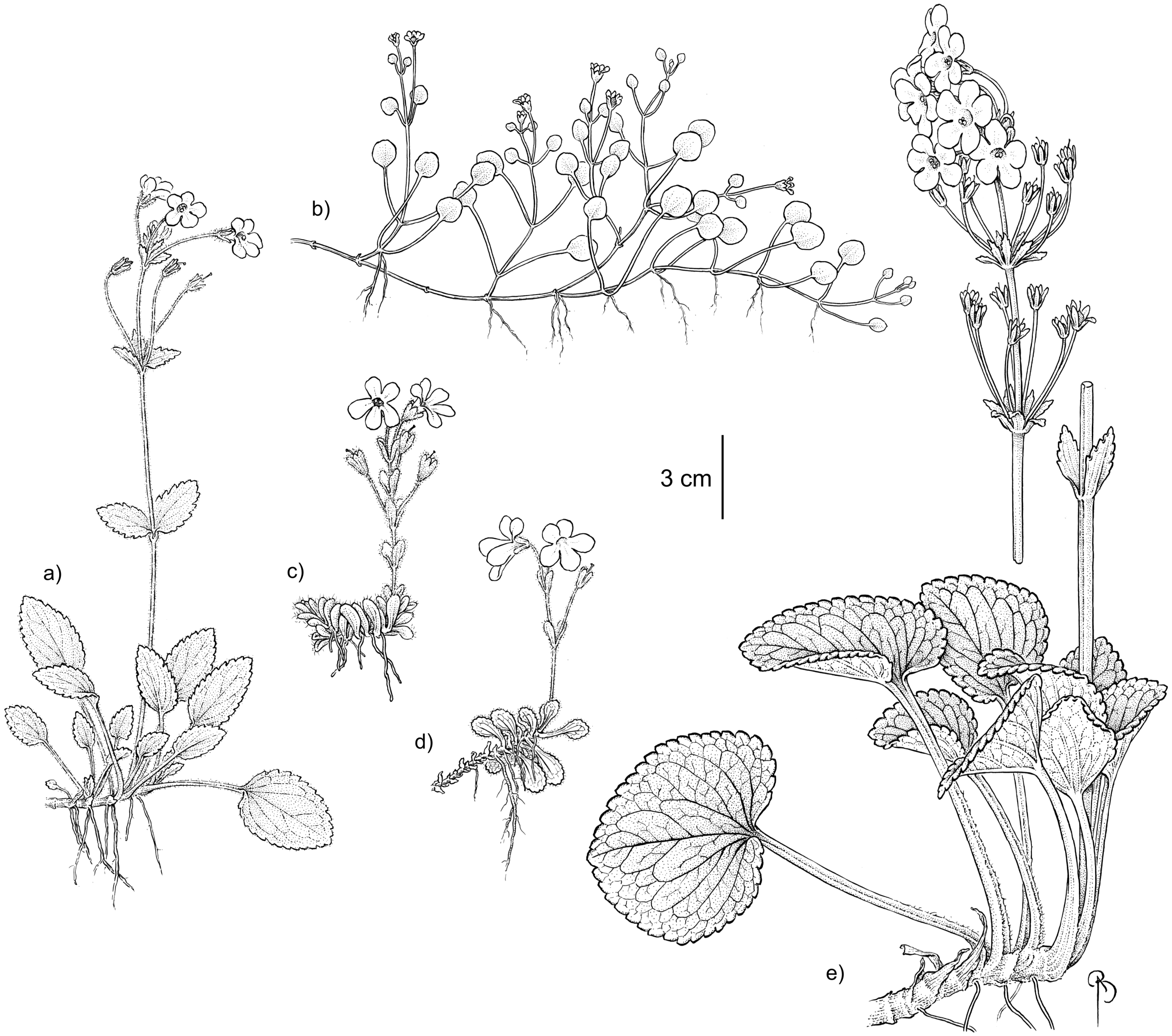
Taylor, D. J., and W. H. Piel. 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol. Biol. Evol. 21:1534–1537.

Tremetsberger, K., T. F. Stuessy, G. Kadlec, E. Urtubey, C. M. Baeza, S. G. Beck, H. A. Valdebenito, C. de Fátima Ruas, and N. I. Matzenbacher. 2006. AFLP phylogeny of South American species of *Hypochaeris* (Asteraceae, Lactuceae). Syst. Bot. 31:610–626.

Vekemans, X., T. Beauwens, M. Lemaire, and I. Roldán-Ruiz. 2002. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Mol. Ecol. 11:139–151.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau. 1995. AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res. 23:4407–4414.

First submitted 25 September 2007; reviews returned 27 November 2007; final acceptance 14 January 2008
Associate Editor: L. Lacey Knowles



Ourisia is a genus of 28 species found in high-elevation habitats mostly in the Southern Hemisphere, including these five New Zealand species: (a) *Ourisia crosbyi*, (b) *O. modesta*, (c) *O. glandulosa*, (d) *O. simpsonii* and (e) *O. macrophylla*. Line drawings by Bobbi Angell, first published in Systematic Botany Monographs (Meudt, 2006).