USING SEQUENCE INVERTED REPEATS (RAPDs) DATA IN SYSTEMAICS: POTENTIAL, PROBLEMS AND SOLUTIONS

Robert P. Adams
Baylor University, Gruver Lab, P.O. Box 727, Gruver, TX 79040 USA
email Robert_Adams@baylor.edu

ABSTRACT

Several cases are presented from *Juniperus* in which the use of RAPD (Random Amplified Polymorphic DNA = Sequence Inverted Repeats SIRs) have been very concordant with DNA sequence and biogeographic data. The use of the term 'Random' should be discouraged because the method is not 'Random' but depends on numerous inverted sequences found in DNA that accounts for hairpin loops important to define the tertiary structure of RNA and thence enzyme activity and specificity. SIRs (RAPDs) require very precise attention to laboratory procedures. Several suggestions are presented to aid in these procedures. Unlike DNA sequence data, SIRs (RAPDs) data should be analyzed by a multi-variate statistical method such as Principal Components or Principal Coordinates Analysis that can account for variation within taxa and treat this variation as error terms. Running replicates and/or sibs is critical to determine if lab supplies and equipment are operating at the peak efficiency.

KEY WORDS: Random Amplified Polymorphic DNA (RAPD), Sequence Inverted Repeats (SIRs), systematics, methods, *Juniperus*.

Obtaining reproducible RAPDs (Random Amplified Polymorphic DNA) or SIRs (Sequence Inverted Repeats) patterns can be very difficult. This has severely impacted the reputation of RAPD data. In fact, I recently had a manuscript rejected with only one comment "RAPDs are known to have problems with reproducibility. I find it inappropriate to base a key on these data" (actually the key used only morphological data, no RAPD data).

DNA fingerprinting methods (i.e., producing a bar-code of DNA bands) that utilize inverted repeats include RAPD (Random Amplified Polymorphic DNA), ISSR (Inter Simple Sequence Repeats) and SSR (Simple Sequence Repeats, when using a single primer). Table

1 compares the basis, application and sequence knowledge needed for the

Table 1. Examples of DNA technology based methods for the analysis of genetic, breeding, and biodiversity studies [based in part from Henry, et al.(1997)].

Gene targeted	Primers	PCR bands	Application	Sequence data needed?
unknown	sequence inverted repeats (RAPDs) (SIRs)	several	bio-diversity, cultivar id., ssp/ var. id. mapping breeding	no, data mining from GenBank should lead to more general primers
various and inter-genic regions	inter-simple sequence repeats (ISSRs)	many	similar to RAPDs (above)	yes, based on known, sequence repeats but not for each taxon
unknown	M13F M13R (AFLPs)	many	similar to RAPDs (above)	no, but DNA must be cut with an enzyme and ligated to M13
unknown	consensus (intron/exon) (promoter/ exon)	many	similar to RAPDs (above)	yes, based on GenBank data mining, but not needed for each taxon

Table 1 (contd.)

Gene targeted	Primers bands	PCR	Application	Sequence data needed?
short simple sequence repeats [ex. $(GA)_{52}$] SSRs = STRs= microsatellites	simple repeats	one to a few	gene flow, parent id, hetero-zygosity estimates to find these	yes for the region bounding the SSR. Costly project
various genes	SNPs, single nucleotide polymorphisms	few	gene flow, parent id, biodiversity	yes, sequence needed for each sample
individual genes	based on sequence data from the taxon	one	species id, phylo-genetics	yes, sequence needed for each taxon

major kinds of DNA technology methods. One can see that those methods that don't require sequence knowledge are generally mostly utilized in gene mapping, populational studies, infraspecific variation, cultivar identification, etc. Studies concerning higher levels of relationships (between genera, families, etc.) almost exclusively utilize DNA sequencing.

It is important to examine the basis for the existence of sequence inverted repeats (SIRs) in DNA. Sequence inverted repeats (SIRs) in ssDNA form hairpin loops that are important for the control of gene transcription and subsequent protein processing (Brown, 2002). In addition, SIRs are extremely important in determining the tertiary structure of RNA. Figure 1 shows the structure of 16S rRNA (based on

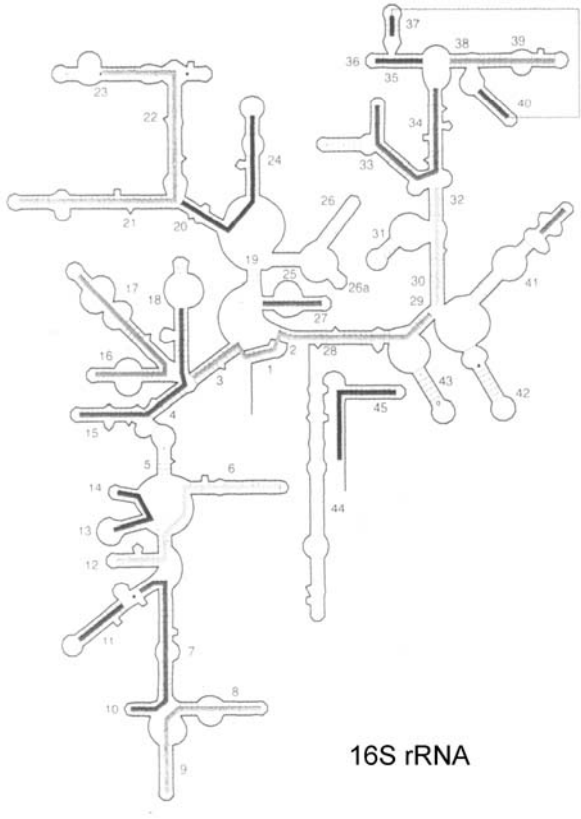

Fig. 1. Primary structure of 16s rRNA (from Noller, 2005). Note the prevalence of hairpin loops in the structure.

Noller, 2005). Hairpin loops are the dominant features of 16S rRNA primary structure. Interestingly, most of these hairpins are secured by inverted repeats of only 3 to 6 bp (Fig. 1). For 16S rRNA (Fig. 1), there appears to be only one 9 bp SIR. No SIR in 16S rRNA appears to be greater than 9 bp long. However, the frequency of hairpin structures in RNA is extensive (Noller, 2005), so SIRs of 10 bp and longer should be expected.

Figure 2 shows, diagrammatically, how sequence inverted repeats in DNA relate specifically to the formation of hairpin loops in RNA (Fig. 2). UBC 212 primer (shown in Fig. 2) is one of the 20 best

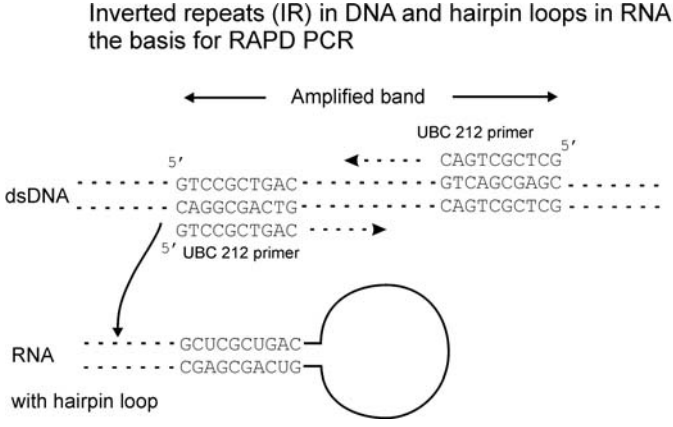

Fig. 2. Diagrammatic representation of an inverted repeat PCR and the relationship of the inverted repeat to a RNA hairpin loop.

primers found in our assays of 500 UBC primers screened on numerous plants, deer, fish and human DNAs (Table 2). PCR using the single UBC 212 primer results in an amplified band from this section of DNA (Fig. 2). The distance between the inverted repeats determines the size of the amplified band and also the hairpin loop size in the RNA (in this example). Of course, an additional priming site(s) may be present much farther downstream (even in an intron or in an inter-genic region) and would result in an additional, larger amplified band(s).

The use of single primers (inverted repeats) was co-discovered by Welsh and McClelland (1990) and Williams, et al. (1990). It is unfortunate that the terms 'random' and 'arbitrary' were used to describe the sequences of these primers, because we have discovered that the sequences are definitely neither 'random' nor 'arbitrary'. Beginning in 1990, our lab (Adams, Baylor University) began to screen 10 bp RAPD and 17-21 bp ISSR primers available in kits from the University of British Columbia (UBC). We have evaluated 500 RAPD and 100 ISSR primers for their ability to: 1. amplify DNA (from various sources, both

plants and animals); 2. obtain reproducible bands in replicate runs; 3. produce many bands; and 4. produce bands that are polymorphic between closely related species. These screenings revealed about 20 RAPD primers (4%) (Table 2) and 6 ISSR primers (6%) that met those criteria. It is now quite apparent that only certain sequences of IRs are common in genomes (about 4% of these tested).

Table 2. List of the most useful primers obtained from screening the UBC primer kits.

Name Sequence	Name Sequence
Best 20 primers:	Very variable (sensitive) primers:
116 TAC GAT GAC G	234 TCC ACG GAC G
134 AAC ACA CGA G	265 CAG CTG TTC A
153 GAG TCA CGA G	327 ATA CGG CGT C
184 CAA ACG GCA C	
204 TTC GGG CCG T	Other good primers
212 GCT GCG TGA C	131 GAA ACA GCG T
218 CTC AGC CCA G	237 CGA CCA GAG C
239 CTG AAG CGG A (conservative)	268 AGG CCG CTT A
	346 TAG GCG AAC G
244 CAG CGA ACC G	352 CAC AAC GGG T
249 GCA TCT ACC G	399 TTG CTG GGC G
250 CGA CAG TCC C	412 TGC GCC GGT G
338 CTC TGG CGG T	432 AGC GTC GAC T
347 TTG CTT GGC G	482 CTA TAG GCC G
375 CCG GAC ACG A	498 GAC AGT CCT G
376 CAG GAC ATC G	499 GGC CGA TGA T
389 CGC CCG CAG T	
391 GCG AAC CTC G	
413 GAG GCG GCG A	
431 CTG CGG GTC A	
478 CGA GCT GGT C	

Can these DNA bands be used in systematic studies? Figure 3 shows a comparison of two classifications of *Juniperus* species based on nrDNA (ITS) sequences and RAPD data. The correlation between these classifications was 0.95. Notice that whereas the ITS sequence data failed to resolve *J. macrocarpa*, *J. oxycedrus* and *J. o*. var. *badia*, these taxa were resolved in the RAPDs data (Fig. 3). Our experience in using several gene sequences for the phylogeny of *Juniperus* project (Schwarzbach et al., 2008), is that, in *Juniperus*, we still do not have gene sequences that can resolve very closely related species or many varieties.

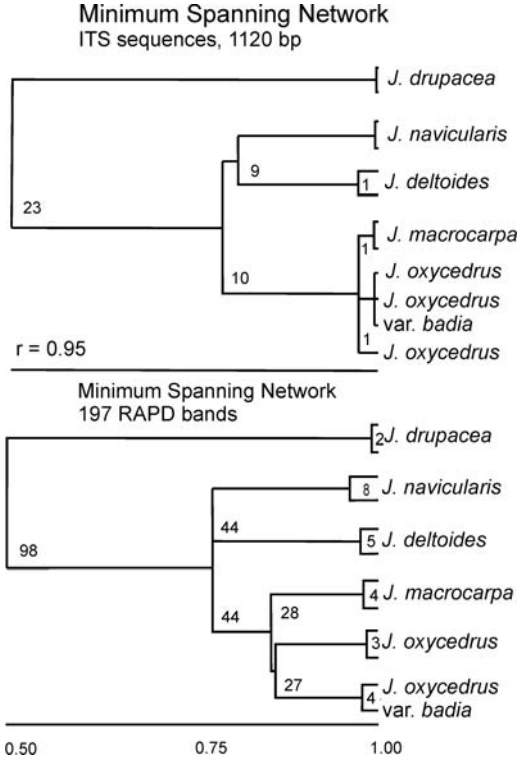

Fig. 3. Comparison of classifications based on ITS (nrDNA) sequences and RAPDs data (adapted from Adams et al., 2003). The correlation between the classifications is 0.95.

The juniper from the southwestern mountains of the Arabian peninsula has been called *J. excelsa* or *J. procera*. Principal coordinate analysis (PCO) using 121 RAPD bands to compute measures of similarity resulted in a very strong trend (Fig. 4, axis 1, 54%) that separated *J. excelsa* from the *J. procera* populations.

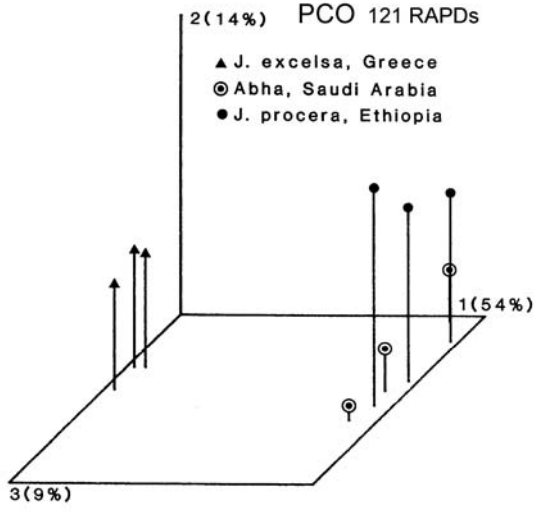

Fig. 4. PCO based on 121 RAPD bands for *J. excelsa*, Greece, putative *J. procera* from Abha, Saudi Arabia and *J. procera*, Ethiopia (adapted from Adams, et al., 1993).

Notice that 54% of the variation (axis 1) is due to the separation of *J. procera* from *J. excelsa* and that the putative *J. procera* plants from Abha all group with *J. procera* from Ethiopia.

Figure 5 shows a RAPD gel and sesquiterpenoids for *J. excelsa* (Greece), Abha, Saudi Arabia and *J. procera*, Ethiopia. It seems apparent that RAPDs can give the same kind of information as seen in the terpenoids (Fig. 5).

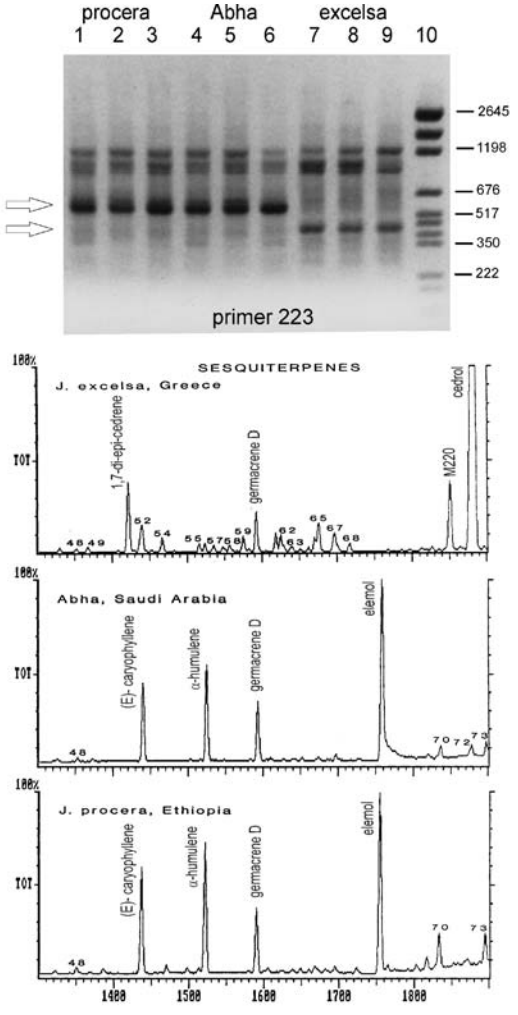

Fig. 5. Comparison of RAPD gel (primer 223) and sesquiterpenoids for taxa from Greece, Saudi Arabia, and Ethiopia (adapted from Adams, et al., 1993).

Figure 6 shows that RAPDs (Demeke, et al., 1992) analyzed by PCO perfectly reflect the famous U triangle (U, 1935) of relationships among *Brassica* species (based on chromosomal data).

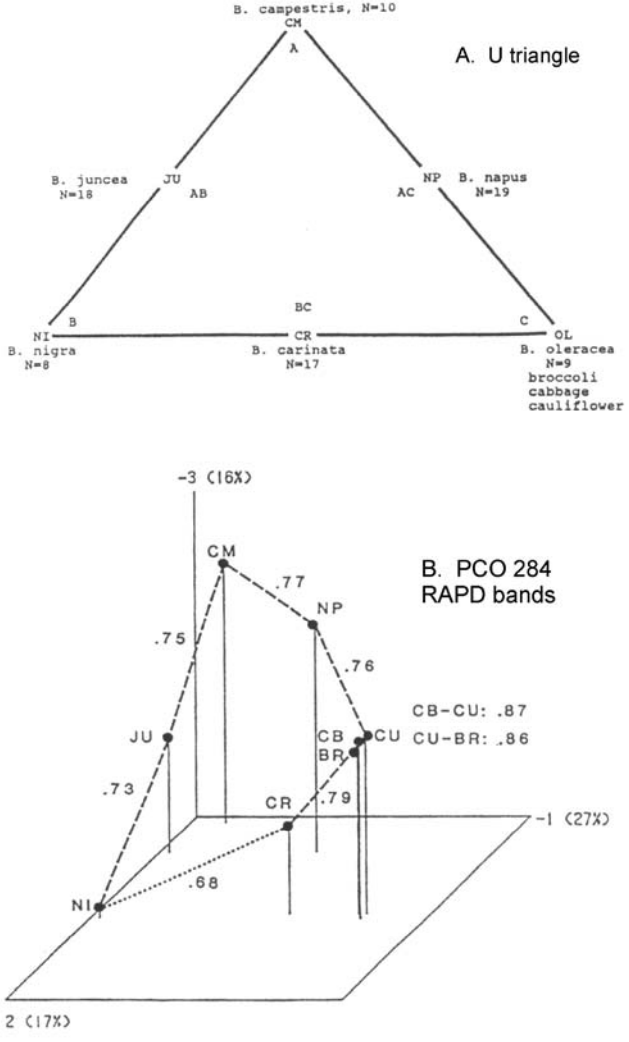

Fig. 6. Comparison of the U triangle and PCO based on RAPDs.

The U triangle of relationships among *Brassica* was based on chromosomal studies (U, 1935). Genotypes (with haplotypes) are represented by AA, AB, BB, BC, CC, and AC. B. PCO of the same *Brassica* taxa based on 284 RAPD bands. Notice the U triangle is perfectly represented by the RAPD data.

SOLUTIONS

There seem to be several factors that have enabled our lab to be able to utilize these kinds of data. First, we assume that there will be errors. There are some who think that data must be perfect in order to contain useful information. No data set (if very large) is likely to be free of errors (even DNA sequence data).

Second, we have developed laboratory methods that minimize errors. These methods are exhaustively discussed in Adams et al. (1998) and the reader is referred to www.juniperus.org to obtain a reprint. But it is worthwhile to consider a few of the major problems and solutions. The idea that RAPDs are not useful for systematics appears to have originated from a study by Penner et al. (1993) who investigated reproducibility in RAPDs using the same target DNA and primers in different laboratories. They found that the problems with reproducibility were mainly due to differences between PCR machines.

Another very influential study was that of Jones, et al. (1997) who sent *Populus* DNA, 2 primers, *Taq* polymerase (DynaZyme), and agarose to eight labs that used their own water, PCR tubes, and thermocyclers (3 of the most critical factors in RAPD PCR). They found only about a 75% reproducibility. Initially, in their AFLP tests, they had profiles that had 50% of the bands missing, but through practice, that improved to nearly perfect reproducibility (Jones et al., 1997). It seems odd that the reproducibility of RAPDs did not improve. However, it should be noted that there is no mention of water having been sent to each of the labs. The quality of water used in reactions is extremely critical. In addition, different thermocyclers were used and apparently not calibrated by external thermal meters.

We recently had to change suppliers for PCR tubes. Comparisons of tubes from 6 suppliers revealed that only 2 of the 6 manufacturers' PCR tubes gave the same results as our old 'discontinued' PCR tubes! In addition, we surveyed 4 sources of Taq and found tremendous variation in their products. So it seems that the Penner et al.

(1993) and Jones et al. (1997) studies did not adequately account for all variables that needed to be considered. In addition, variation among laboratory techniques alone likely accounted for considerable amounts of the variance among laboratories.

A very critical factor in PCR is the mixing of reagents. Pipetting is a source of many errors. Mixing reagents before and after pipetting is another critical step. To minimize errors due to pipetting very small amounts, Adams et al. (1998) investigated the stability of large amounts of RAPDs stock (ddwater, $MgCl_2$, 10x buffer, dNTPs and an entire vial of *Taq*). To maximize the potential for deterioration, the stock was stored at 22º C. They reported (Adams et al. 1998) no change in the RAPD amplification pattern after 4 days and only a slight reduction of band intensity after 60 days at 22º C! Of course, we store our RAPD stock at 4º C. The use of a large RAPD stock solution (enough to run all samples for one primer) and completely using an entire vial of *Taq* has greatly reduced variation (missing bands) in our replicated runs.

A second critical factor can be illustrated by my experience with a new post-doc from China. For several days he experienced failure to amplify for about 15 of 60 reactions. Finally, one day he obtained 60/60 perfect runs. Then the same success (60/60) for the next 5 days. I asked him if something was different. He replied, "I am now vortexing each PCR tube twice instead of once." Mixing of *Taq* and other reagents is extremely critical.

Because our lab has done considerable RAPD and sequencing, it is interesting to note that failure for a sequence to be generated is not unusual (or unexpected), but because the sequencing reaction did not give a credible sequence, one re-sequences the DNA without much thought. With sequencing, one can quickly see that the results are not credible and thus, the sample must be re-amplified and re-sequenced. But this is, intrinsically, more difficult to ascertain in PCR fingerprinting methods. Thus, it is important to have standard DNAs that generate standard profiles for each primer so these can be run as controls when new reagents are made, when a new thermocycler is used, etc. And it is critical to have multiple, genetically near-identical samples of each taxon to act as a reference to check that the amplification is credible. Obviously, if bands are not present or the larger bands are missing, one needs to re-run the RAPD analysis (we re-run in triplicate). It is almost impossible to overemphasize the attention

to detail that is required to do excellent RAPD analyses (see Adams, et al., 1998 for a very detailed discussion).

Thirdly, polysaccharides and other inhibitors (Pandey, et al. 1996; Adams et al., 1998) can cause problems in band amplifications. This is immediately apparent if the larger fragments are missing (2000-3500 bp). The DNA from *Juniperus flaccida*, extracted by the hot CTAB method, contains considerable inhibitors (Adams, et al., 1998) and could only be successfully amplified at a concentration of 250 pg of DNA (or less) per 15 µl PCR reaction. We now routinely use 300 pg of DNA per 15 µl PCR reaction. If the reaction does not amplify, we make serial dilutions of the DNA and run these until they amplify. In spite of better extraction kits, the easiest solution for inhibitor problems is to dilute the DNA.

Even if all these precautions are taken, there is still the problem of obtaining uniform, repeatable thermocycling conditions. We monitor every thermocycler with an external chart recorder and a temperature probe inside a control tube. This has enabled us to detect thermal cycling problems and correct them, and has also reduced our variation.

The lack of resolution of similar molecular weight bands on agarose (homology) is a problem (Rieseberg, 1996), but the use of similarity measures based on character differences, coupled with multivariate methods such as principal coordinates analysis (PCO), effectively eliminates this problem (see Adams and Rieseberg, 1998 for a detailed discussion).

In addition, it is important to recognize that not all bands generated are useful. There are some bands that, in replicated analyses, just tend to vary. By running several individuals from each taxon, collected in the same population, one can determine which bands are not representative of a population, variety or species. These can be eliminated. Our policy is, if in doubt, don't score the band. It is common that we discard 30 - 40% of the bands as being inconsistent, or just difficult to score. Demeke, et al. (1992) found in a study of *Brassica* that if fewer than 100 bands were utilized, the PCO ordination began to lose its correspondence to the U triangle (U, 1935). So it is necessary to start with 150 - 200 bands (generally using 15 - 18 primers that have been selected by intense screening, see Table 2).

Finally, it should be emphasized that multivariate statistical methods have the capability of accounting for error variance and are highly desirable for analysis. The movement in systematics to

parsimony tree building using sequence data has caused many to lose perspective that other kinds of data may require different methods for analyses. Certainly, those of us who have worked many years with secondary compound data are well aware of error variance and the need to factor data to remove (and account for) error variance. Perhaps a large part of the prejudice against the use of RAPD data for systematics is the result of a new generation of systematists who were not trained in the analysis of sampling errors.

It should be noted that not every person nor every lab can do this kind of analyses. I have had several students visit my lab that could just not do this kind of exacting work. I have had three students come from my colleague's lab in zoology, on separate occasions, for training in my lab. In each case, they obtained good results, but upon returning to their lab, they could never obtain reproducible results and abandoned the methods. Whether the problems were with their reagents, PCR tubes, water or thermocyclers was not determined.

Can RAPD data be used for systematics? From the examples above, I think it is impossible to explain the correlation between DNA sequence data and RAPD data classifications as chance. Laboratory procedures must be conducted at the highest standards using replicated analyses within each data set. Clearly, other kinds of data, including RAPDs, AFLP, ISSR, SSR, etc. can be utilized in systematics, but the methods of analyses need to be appropriate for the kinds of data concerned.

ACKNOWLEDGEMENTS

This research was supported in part with funds from NSF grant DEB-316686 (A. Schwarzbach and R. P. Adams) and funds from Baylor University.

LITERATURE CITED

Adams, R. P., T. Demeke, T. H. Abulfatih. 1993. RAPD DNA Fingerprints and terpenoids: Clues to past migrations of *Juniperus* in Arabia and east Africa. Theoretical Applied Genetics 87: 22-26.

Adams, R. P., L. E. Flournoy, R. N. Pandey. 1998. Obtaining reproducible patterns from random polymorphic DNA amplification (RAPDs). In Adams, R. P., Adams, J. E. (Eds.), Conservation of Plant Genes III: Conservation and utilization of African plants. Missouri Botanical Garden Press, p. 229-236.(St. Louis).

Adams, R. P., L. H. Rieseberg. 1998. The effects of non-homology in RAPD bands on similarity and multivariate statistical ordination in *Brassica* and *Helianthus*. Theoretical Applied Genetics 97: 323-326.

Adams, R. P. A. Schwarzbach, R. N. Pandey, R. N. 2003. The Concordance of Terpenoid, ISSR and RAPD markers, and ITS sequence data sets among genotypes: An example from Juniperus. Biochemical Systematics Ecology 31: 375-387.

Brown, T. A., 2002. Genomes, 2nd ed., John Wiley & Sons., Inc. (NY).

Demeke T., R. P. Adams, R. Chibbar. 1992. Potential taxonomic use of random polymorphic DNA (RAPDs): A case study in *Brassica*. Theoretical Applied Genetics 84: 990-994.

Demeke, T., R. P. Adams. 1994. The use of PCR-RAPD analysis in plant taxonomy and evolution. In: Griffin, H. G., A. M. Griffin, A. M. (Eds.), PCR Technology: Current Innovations, CRC Press, p. 179-192. (Boca Raton, FL).

Henry, R. J., H. L. Ho, S. Weining. 1997. Identification of cereals using DNA-based technology. Cereal Foods World 42: 26-29.

Noller, H. F. 2005. RNA structure: Reading the ribosome. Science 309: 1508-1513.

Pandey, R. N., R. P. Adams, L. E. Flournoy. 1996. Inhibition of random amplified polymorphic DNAs (RAPDs) by plant polysaccharides. Plant Molecular Biology Reporter 14: 17-22.

Penner, G. A., A. Bush, R. Wise, W. Kim, L. Domier, K Kasha, A. Laroche, G. Scoles, S. Molnar, G. Fedak, 1993. Reproducibility of random amplified polymorphic DNA (RAPD) analysis among laboratories. PCR Methods and Applications, 2: 341-345.

Rieseberg, L. H., 1996, Homology among RAPD fragments in interspecific comparisons. Molecular Ecology 5: 99-105.

Schwarzbach, A. E., R. P. Adams, J. A. Morris. 2008. Phylogeny of *Juniperus* based on nrDNA and trnC-trnD sequences. (in prep.)

U, R., 1935. Genomic analysis of *Brassica* with special reference to the experimental formation of *B. napus* and it peculiar mode of fertilization. Japan J. Bot. 7: 389.

Welsh, J.; M. McClelland. 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Research 18, 7213–7218.
Williams J. G. K., A. R. Kubelik, K. J. Livak, J. A. Rafalski, S. V. Tingey. 1990. DNA polymorphism amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18: 6531–6535.