Theor Appl Genet (1998) 97: 323–326

R. P. Adams · L. H. Rieseberg

The effects of non-homology in RAPD bands on similarity and multivariate statistical ordination in *Brassica* and *Helianthus*

Received: 6 December 1997 / Accepted: 11 December 1997

Abstract In order to estimate the impact of mis-coding non-homologous, co-migrating DNA bands as homologous, two sets of data were utilized. Analyses were conducted using three *Helianthus* species in which each co-migrating band had previously been confirmed. Comparisons of the similarities between these three *Helianthus* species using the original 177 RAPD bands and the corrected, homology verified, 197 RAPD band data set revealed that the triangular relationship among these three species was almost identical in both data sets. The non-homology errors in the *Helianthanus* data sets were found to be random. These random errors merely reduced the absolute similarities, but not the relative similarities nor the relationships among the taxa, in principal-coordinate-analysis ordination. Analyses of RAPDs for the classical *Brassica* U triangle were made by inserting random non-homologies for 5, 10, 15 and 20% of the original 220 RAPD bands. These analyses revealed a progressive decrease in similarities and less loading on the first two axes in principal coordinate analysis (PCO). However, the basic U triangle of relationships among these six *Brassica* species was maintained. It appears that if errors in homology of co-migrating DNA bands are random, this will have little effect on the relative similarities and on PCO ordination. This helps explain the successful use of RAPDs at the specific level.

Key words RAPD · DNA fingerprinting · Homology · Statistics · Classification

Communicated by H. F. Linskens

R. P. Adams (✉)
Plant Biotechnology Center, Baylor University,
P.O. Box 669, Gruver, TX 79040, USA

L. H. Rieseberg
Biology Department, Indiana University,
Bloomington, IN 47405, USA

Introduction

It is now apparent that when numerous DNA bands are obtained by DNA fingerprinting (random amplified polymorphic DNAs, RAPDs), and other related techniques, some of the co-migrating bands are not homologous. For example, Rieseberg (1996) recently analyzed the homology of RAPD bands using molecular techniques. Analyses of 220 co-migrating bands generated among three sunflower species revealed that 9% of the bands that co-migrated were not homologous.

Williams et al. (1993) found that 10% of co-migration bands were not homologous among several species of *Glycine*. Intergeneric analyses between *Brassica* species and *Raphanus sativa* revealed that about 20% of the co-migrating bands were not homologous (Thormann et al. 1994). However, Thormann et al. (1994) also showed that, within species, all co-migrating bands were homologous. It is intuitively obvious that at the infraspecific level, co-migrating bands will be homologous, and yet between species and genera homology will be less.

This is exactly the same situation faced by chemosystematists some 30 years ago. Adams (1972) examined this problem in the case where two co-running (or more properly "co-eluting") gas chromatographic peaks, assigned the same identity might, in fact, not be homologous (i.e., be different compounds). It was found that even when a compound accounted for 20% of the total character matching weight, it had minimal effects on the Gower metric of similarity (Adams 1972).

It is illustrative to recall that flavonoid work began by comparisons of spots and colors and eventually led to the identification of compounds. Terpene research originally utilized just retention times on gas chromatographic charts. With the advent of combined gas chromatography – mass spectroscopy – computer data systems (GC – MS – CS) almost every peak can

now be identified (Adams 1995). Even when compounds can not be completely identified, the mass spectra pattern can be used to clearly denote that the peaks in question are the same unknown compound. It should be noted that even with the easy GC – MS – CS analyses of today, no one identifies the components for every individual in a population. It is obvious when one examines gas chromatograms that all the individuals within a population vary by quantitative differences.

The purpose of the present paper is to examine the effects of errors in the homology assignments of co-migrating bands. Two data sets were utilized: three *Helianthus* species in which the homology of each co-migrating RAPD band has been verified by molecular techniques (Rieseberg 1996); and six *Brassica* species (Demeke et al. 1992) that define the classical U triangle (U 1935) for which progressively 5, 10, 15 and 20% of the co-migrating RAPD bands were randomly re-coded as non-homologous.

Materials and methods

Two data sets were constructed from RAPD analysis of three *Helianthus* species (*H. annuus* L., A; *H. petiolaris* Nutt., P; and the putatively stabilized hybrid-derived *H. anomalus* Blake, O). The first data set consisted of the 177 RAPD bands as originally scored (0 = absent, 1 = present). The second data set of 197 RAPD bands contained an additional 20 bands that proved to be non-homologous (Rieseberg 1996).

These data were coded into a matrix of taxa by character values. Similarity measures were computed using absolute character-state differences (Manhattan metric) divided by the maximum observed value for that character over all taxa (= Gower metric; Gower 1971; Adams 1975). Principal coordinate analysis (PCO) of the similarity matrix and ordination follows Gower (1966).

The second kind of data utilized were the RAPD data for six *Brassica* species (*B. campestris*, CM; *B. carinata*, CR; *B. juncea*, JU; *B. napus*, NP; *B. nigra*, NI; *B. oleracea*, OL) from Demeke et al. (1992), except that the bands were re-coded as: 0 = absent, 1 = present (in contrast to the original work where the bands were coded quantitatively between 0 and 6). The original data set for the six *Brassica* species contained 220 RAPD bands. Bands that co-migrated were randomly changed to reflect non-homology, such that 5, 10, 15 and 20% of the total bands for each species were re-coded to generate four additional data sets that departed from the original data. Similarities and PCO were performed on these data sets in the same manner as with the *Helianthus* data. The PCO program is available from R.P.A.

Results and discussion

Analyses of the original set of data for the three *Helianthus* species utilized 177 RAPD bands. PCO resulted in two principal coordinate axes accounting for 58 and 42% of the variation. The ordination is shown in Fig. 1 (dashed lines). PCO of the homology verified data utilized 197 bands (20 of the original 177 bands were found to be non-homologous, Rieseberg 1996). PCO resulted in two axes accounting for 56 and 44% of the variation. Notice that these two triangles of relationships are essentially the same (Fig. 1).

It is interesting to note that the similarities based on the original data (containing 11.3% erroneous band-matchings) are each larger than the corresponding similarity base on the corrected data (Fig. 1). This would be expected as the coding of mis-matches as matches will increase the similarity.

The second data set, *Brassica*, was utilized by intentionally introducing random non-homologies in co-migrating bands in order to examine several levels of "errors". The classical U triangle (U 1935) of relationships among these six Brassica species is depicted in Fig. 2 (upper left). PCO using the data as originally coded (adapted from Demeke et al. 1992) is shown in Fig. 2 (upper right). Notice the close agreement between the classical, chromosomally defined, U triangle and the triangle utilizing the original RAPD data (Fig. 2, upper left vs upper right).

Coding in 5, 10, 15 and even 20% non-homology has only a small effect on the triangular relationship (Fig. 2). It is interesting to note that as more random "errors" are introduced into the data sets, the amount of the variance removed by the first two coordinate axes decreases from 63% (original data) to 51% (20% non-homology). With six objects in PCO, it should take five or less axes to completely depict the relationships in hyper-space. The random "noise" introduced resulted in lower loading onto the first two axes and more variation expressed on the remaining three axes.

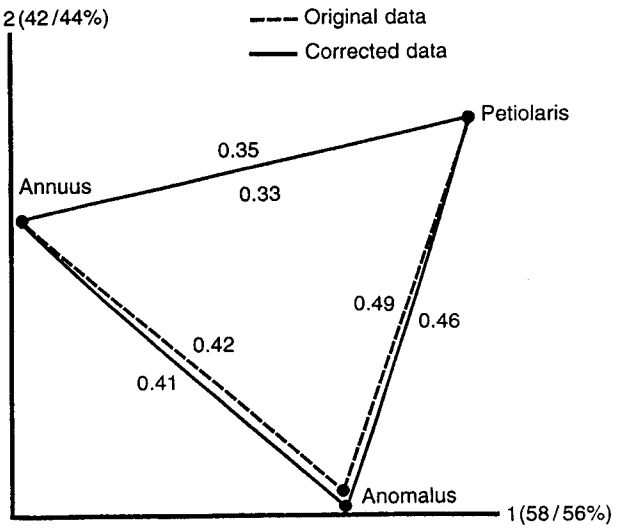

Fig. 1 Triangular relationships among three *Helianthus* species are revealed by PCO. Essentially the same triangular relationship was found using either the original data or the corrected (homology verified) data. The *numbers* inside the triangle are the similarities between OTUs for the original data and the *numbers* outside the triangle are similarities for the corrected data. The first number on each axis is the variance removed from the original data set and the second number is the variance removed using the corrected data set

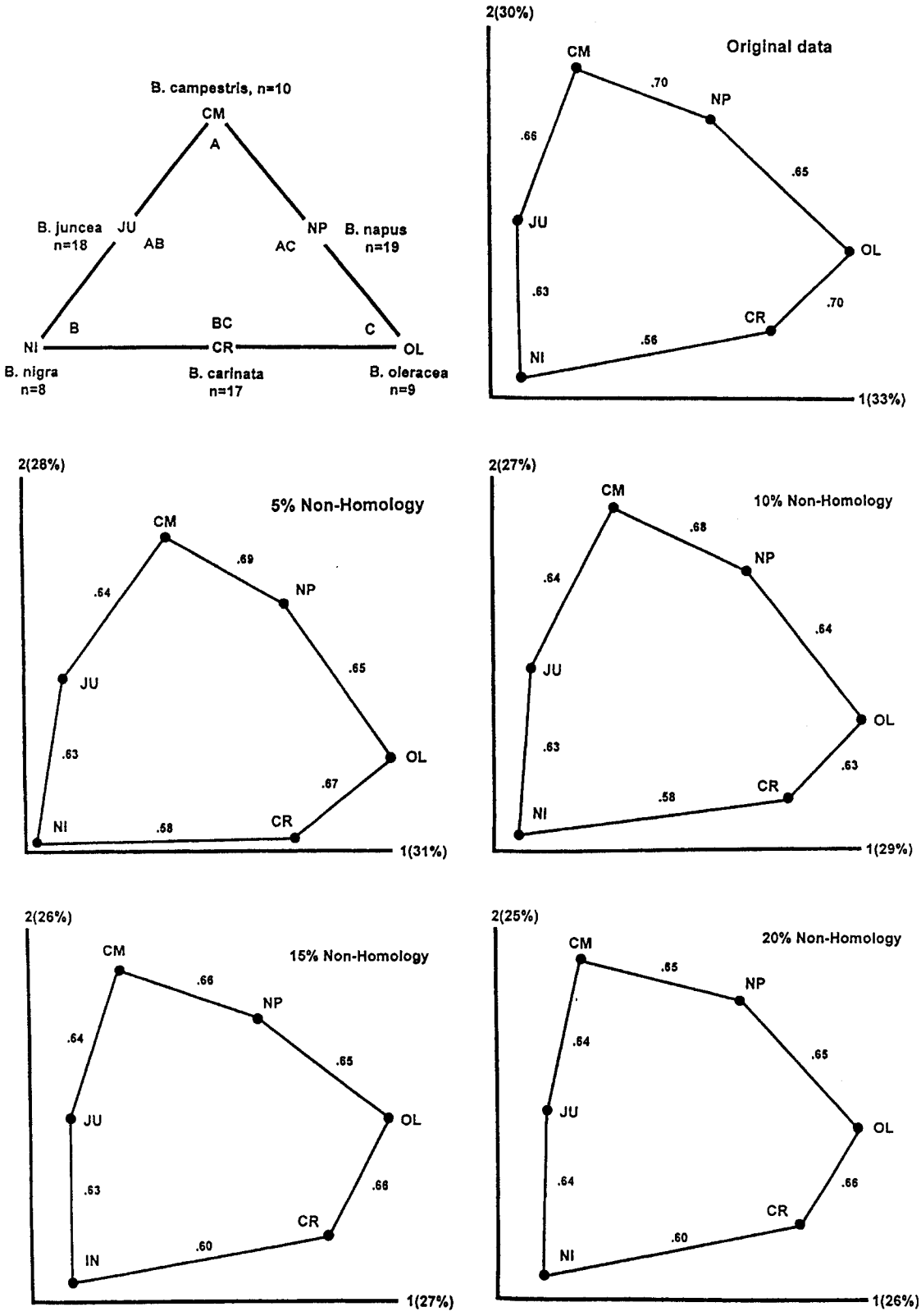

Fig. 2 (Upper left) classical U triangle (from U 1935) of relationships in *Brassica*; (upper right) PCO using the original RAPD data set (from Demeke et al. 1992); (middle left through lower right) PCO using data sets in which 5, 10, 15 and 20% of the bands have been re-coded as non-homologous. The similarities between OTUs are denoted by the *numbers* between OTU codes

In order for homology to present a problem in principal coordinate analysis, the "false homologies" would have to be non-random. If two co-migrating bands are not homologous this will be a random event. That is, any two co-migrating bands are as likely to be homologous (depending on how closely related the individuals are) as any other two co-migrating bands. It is illogical to assume that individuals A and B will have a non-random set of "false homologies". These random events contribute to the "error variance" in principal coordinate analysis and can be seen by the asymptoting of eigenvalues after the first few eigenroots are removed. This is obvious by comparing the percentage of variation removed in, for example, a comparison between the original and the 20% non-homology data sets:

original data set: 33.2; 29.9; 13.8; 12.5; 10.6%, 20% non-homology: 26.3, 24.9, 16.9, 16.2, 15.7%.

There is also heuristic evidence that homology is not a serious problem in principal coordinate analysis of RAPDs. For example, analyses of RAPD data almost perfectly elucidated the classical "U" triangle of relationships among *Brassica* species (Demeke et al. 1992). Likewise, the three classical morphologically delineated *sub-genera* (italics ours) of *Juniperus* (*Caryocedrus, Juniperus, Sabina*) were revealed in the RAPD data (Adams and Demke 1993). In addition, both RAPDs and terpene patterns were successfully used to re-classify putative *Juniperus excelsa* individuals from Arabia as, in fact, members of *J. procera* from Africa (Adams et al. 1993).

We should, however, add a cautionary note. We are not implying that DNA band data used for phylogenetic studies utilizing the common distance and parsimony algorithms will be affected in the manner of the aforementioned examples. It is not yet clear how robust these approaches are to the addition of random error, such as that likely to be found in many RAPD data sets. On the other hand, our data demonstrate that multivariate ordination methods (such as PCO) can assimilate random errors with little effect on relative similarity ordinations. Thus, multivariate ordination appears to be a robust method for the analyses of DNA band data at both the population and specific levels.

References

Adams RP (1972) Numerical analyses of some common errors in chemosystematics. Brittonia 24:9–21

Adams RP (1975) Statistical character weighting and similarity stability. Brittonia 27:305–316

Adams RP (1995) Identification of essential oil components by gas chromatography/mass spectroscopy. Allured Publications, Carol Stream, Illinois, USA

Adams RP, Demeke T (1993) Systematic relationships in *Juniperus* based on random amplified polymorphic DNAs (RAPDs). Taxon 42:553–571

Adams RP, Demeke T, Abulfatih HA (1993) RAPD DNA fingerprints and terpenoids: clues to past migrations of *Juniperus* in Arabia and east Africa. Theor Appl Genet 87:22–26

Demeke T, Adams RP, Chibbar R (1992) Potential taxonomic use of random amplified polymorphic DNAs (RAPDs): a case study in *Brassica*. Theor Appl Genet 84:990–994

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 315–328

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–874

Rieseberg LH (1996) Homology among RAPD fragments in interspecific comparisons. Mol Ecol 5:99–105

Thormann CE, Ferrieira ME, Camargo LEA, Tivang JG, OsbornTC (1994) Comparison of RFLP and RAPD markers to estimating genetic relationships within and among cruciferous species. Theor Appl Genet 88:973–980

U N (1935) Genomic analysis of *Brassica* with special reference to the experimental formation of *B. napus* and its peculiar mode of fertilization. Jpn J Bot 7:389–452

Williams JGK, Hanafey MK, Rafalski JA, Tingey SV (1993) Genetic analysis using random amplified polymorphic DNA markers. Methods Enzymol 218:704–740