# AFLP data and the origins of domesticated crops

**Robin G. Allaby and Terence A. Brown**

**Abstract:** Amplified fragment length polymorphism (AFLP) datasets have been used to construct neighbor-joining trees from which monophyletic origins for crops such as einkorn wheat, barley, and emmer wheat have been inferred. We simulated several different multiple domestication scenarios for an imaginary cereal crop and examined the resulting domesticated populations. The simulations showed that the population biology aspects of the domestication process can result in independently domesticated populations merging in such a way that a monophyletic origin is erroneously inferred when the resulting population is examined by AFLP genotyping and neighbor-joining analysis. The results bring into question the use of this method to infer the origins of real crops.

*Key words:* AFLPs, agriculture, neighbor-joining, plant domestication.

**Résumé :** Les données sur le polymorphisme de longueur des fragments amplifiés (AFLP) ont souvent été employées pour construire des arbres phylogénétiques par analyse de type « neighbor-joining ». À partir de tels arbres, il a été déduit que l’engrain, l’orge et le blé amidonnier avaient une origine monophylétique. Les auteurs ont simulé plusieurs scénarios différents de domestication pour une espèce céréalière imaginaire et ils ont examiné les populations domestiquées qui pourraient en résulter. Ces simulations montrent que certains aspects de la biologie des populations du processus de domestication peuvent faire penser que des populations domestiquées indépendamment ont une origine monophylétique. On pourrait arriver à une telle conclusion erronée en procédant au génotypage par AFLP et en analysant les données par l’approche « neighbor-joining ». Ces résultats soulèvent un doute sur la justesse de l’emploi de cette approche pour déduire l’origine de vraies espèces cultivées.

*Mots clés :* AFLP, agriculture, ‘neighbor-joining’, domestication des plantes.

[Traduit par la Rédaction]

## Introduction

The use of amplified fragment length polymorphism (AFLP) datasets to construct neighbor-joining trees is increasingly being applied to the study of crop domestication (e.g., Heun et al. 1997; Badr et al. 2000; Özkan et al. 2002). These studies have resulted in apparently monophyletic groupings for domesticates of einkorn (*Triticum monococcum* L.), barley (*Hordeum vulgare* L.), and emmer (*Triticum turgidum* L.), implying that the domestication of each of these crops was a unique event whose geographical location can be identified by comparisons with the AFLP genotypes of wild populations. Badr et al. (2000) are particularly assertive, stating that a neighbor-joining tree presented in their paper “closes the long-standing debate on the origin of barley”. Their conclusion is controversial, because the prevailing view has been that barley was domesticated on multiple occasions (Zohary 1996). This view was based on genetic evidence showing that the non-brittle rachis phenotype of cultivated barleys is coded by two different mutations and so arose on at least two occasions (Takahashi 1972). The validity of comparative analyses of DNA polymorphisms as a means of studying the wild origins of crop plants has been questioned by Abbo et al. (2001), who describe various limitations with this approach and point out inconsistencies with other types of genetic and archaeobotanical evidence. Here we report the results of simulations that show that the use of neighbor-joining analysis of AFLP datasets to infer crop origins is flawed and can lead to erroneous conclusions.

## Materials and methods

We simulated several different multiple domestication scenarios for an imaginary cereal crop and applied the same analysis as used by Heun et al. (1997), Badr et al. (2000), and Özkan et al. (2002) to the resulting domesticated populations. The simulations were based on 200 imaginary allele characters, each of which represented a different AFLP band on an electrophoresis gel. It was assumed that the underlying loci giving rise to the AFLP bands were unlinked and that all character states were selectively neutral such that all alleles were independently subject to random genetic drift. The simulations assumed the existence of three pairs of hypothetical wild populations, each pair created using a different starting scenario. In scenario 1, the wild populations, w1 and w2, were assumed to have a common origin in the recent past. Each AFLP character was assigned a random mean frequency between 0 and 1. This value represented the Gaussian mean allele frequency for that particular AFLP character in all wild populations. The specific allele frequencies for every AFLP character in the wild populations of

Received 27 May 2002. Accepted 3 March 2003. Published on the NRC Research Press Web site at http://genome.nrc.ca on 16 May 2003.

Corresponding Editor: B. Golding.

**R.G. Allaby and T.A. Brown.**[1] Department of Biomolecular Sciences, University of Manchester Institute of Science and Technology, Manchester M60 1QD, U.K.

[1] Corresponding author (e-mail: terry.brown@umist.ac.uk).

doi: 10.1139/G03-025

**Table 1.** Summary of outcomes for each simulation.

(*a*) Comparison between w1, w2, d1, and d2

| | Divergence between w1 and w2 | | |
|---|---|---|---|
| Outcome | Recent (scenario 1) | Distant (scenario 2) | Infinite (scenario 3) |
| No. of simulations | 20 | 20 | 50 |
| w1 formed a single clade | 1 | 20 | 50 |
| w2 formed a single clade | 1 | 20 | 50 |
| d1 formed a single clade | 20 | 15 | 42 |
| d2 formed a single clade | 16 | 12 | 37 |
| Correct origins of d1 and d2 could be identified | 16 | 20 | 50 |

(*b*) Comparison between w1, w2, and hyb5050

| | Divergence between w1 and w2 | | |
|---|---|---|---|
| Outcome | Recent (scenario 1) | Distant (scenario 2) | Infinite (scenario 3) |
| No. of simulations | 20 | 20 | 50 |
| w1 formed a single clade | 1 | 20 | 50 |
| w2 formed a single clade | 1 | 20 | 50 |
| hyb5050 formed a single clade | 15 | 11 | 15 |

(*c*) Comparison between w1, w2, and hyb1090

| | Divergence between w1 and w2 | | |
|---|---|---|---|
| Outcome | Recent (scenario 1) | Distant (scenario 2) | Infinite (scenario 3) |
| No. of simulations | 40 | 40 | 100 |
| w1 formed a single clade | 3 | 40 | 100 |
| w2 formed a single clade | 3 | 40 | 100 |
| hyb1090 formed a single clade | 38 | 38 | 96 |

interest, w1 and w2, were determined by periodical sampling from the Gaussian distribution that had a mean equal to that assigned previously and a standard deviation of 0.1. In scenario 2, the wild populations were assumed to have a more distant origin, but still discernibly related to one another. For this scenario, the allele frequencies for w1 and w2 were assigned as described above, but using a standard deviation of 0.3 to reflect a wider variance in the Gaussian distribution owing to continued drift in allele frequencies over time. In scenario 3, it was assumed that the wild populations had diverged an infinitely long time ago and hence had unrelated allele frequencies. Each AFLP character was therefore assigned a random allele frequency between 0 and 1, independently for w1 and w2, using random numbers taken from a rectangular distribution.

The later analyses required that 10 individuals be taken from each wild population to compare with the domesticates. It was assumed that the allele frequency for each AFLP character in each wild population was equal to the probability that any single individual drawn from that population would have a particular character (Kimura 1962). Each individual was constructed, character by character, by generating random numbers. If the random number was equal to or lower than the assigned frequency for that character in the wild population, then the individual was deemed to possess that particular AFLP character; if the random number was higher than the population frequency, then the individual was deemed not to have that AFLP character. The process was repeated for each of the 200 characters and repeated 10 times to produce 10 individuals.

Domesticated populations d1 and d2 were produced from w1 and w2, respectively. The domestication process was assumed to involve a population bottleneck, followed by a period of random genetic drift (Tanksley and McCouch 1997). To simulate this bottleneck, 10 individuals were drawn from the wild population as described above. The initial allele frequencies for each AFLP character in the domesticated populations were set by the allele frequencies displayed by these 10 individuals (e.g., if 3 of the 10 individuals possessed a particular AFLP character, then the allele frequency in the initial domesticated population was set at 0.3). A population expansion was then simulated using these allele frequencies to construct a new generation of 100 individuals. The population expansion was followed by a period of random genetic drift over 20 generations. During this period, the population size was maintained at 100 individuals, each constructed as described above, using the observed allele frequency for each AFLP character in the population to determine the probability of an individual in the next generation possessing that character. For the subsequent analyses, 10 individuals were drawn from the final population.

We then simulated situations where two independently domesticated populations join to form a hybrid population. Two models were considered. In the first, d1 and d2 had equal inputs into the hybrid population, hyb5050. In the second model, the two domesticates had non-equal inputs, 10% from one domesticate and 90% from the other, resulting in hyb1090. AFLP character frequencies in the hybrid populations were calculated as:

$$\text{frequency} = (p_{\text{d1}} \times f_{\text{d1}}) + (p_{\text{d2}} \times f_{\text{d2}})$$

where $p_{\text{d1}}$ and $p_{\text{d2}}$ are the proportional inputs of d1 and d2, respectively, and $f_{\text{d1}}$ and $f_{\text{d2}}$ are the frequencies of the AFLP character in d1 and d2, respectively. The resulting AFLP character frequencies were used to construct a hybrid popu-

**Fig. 1.** Neighbor-joining trees constructed from similarity matrices using the DICE similarity coefficient calculated for the simulated AFLP data for two imaginary wild cereal populations (w1 and w2), the independently domesticated populations d1 and d2, and the hybrid domesticated populations hyb5050 and hyb1090. See the text for details regarding these populations. Each tree has been bootstrapped by carrying out 100 replicates. Using conventional phylogenetics as a guide, we regard bootstrap values of 75% and above as indicating a significant relationship. (A–C) Examples of trees constructed with w1, w2, d1, and d2. (A) Recent divergence (scenario 1). (B) Distant divergence (scenario 2). (C) Infinite divergence (scenario 3). (D–F) Examples of trees for w1, w2, and hyb5050. (D) Recent divergence. (E) Distant divergence. (F) Infinite divergence. (G–I) Examples of trees for w1, w2, and hyb1090. (G) Recent divergence. (H) Distant divergence. (I) Infinite divergence.

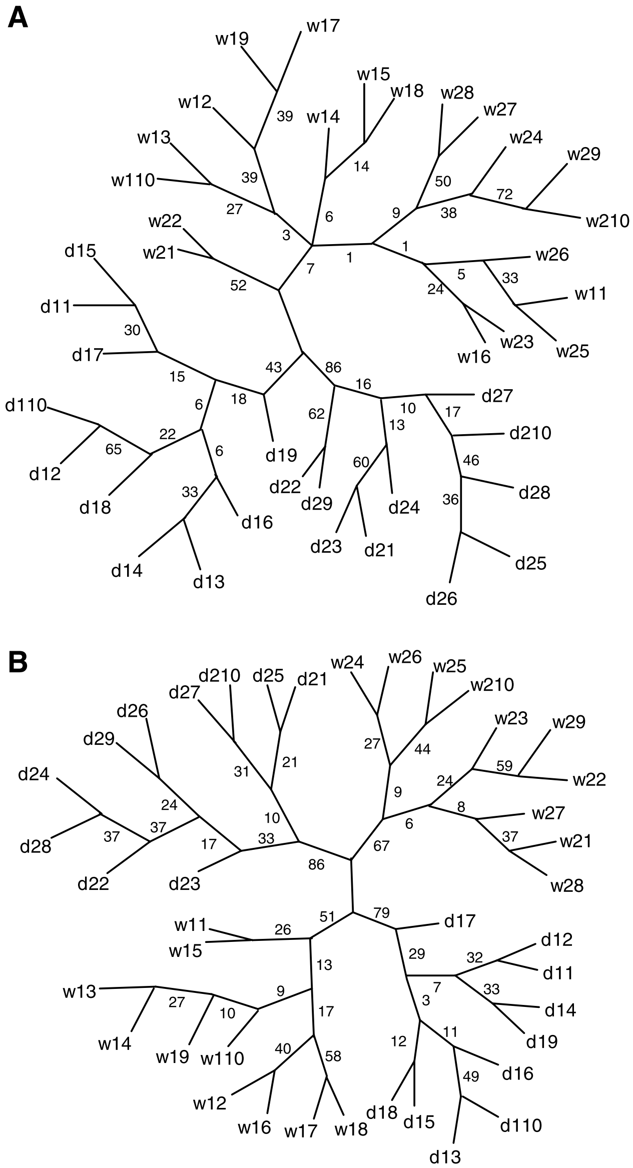
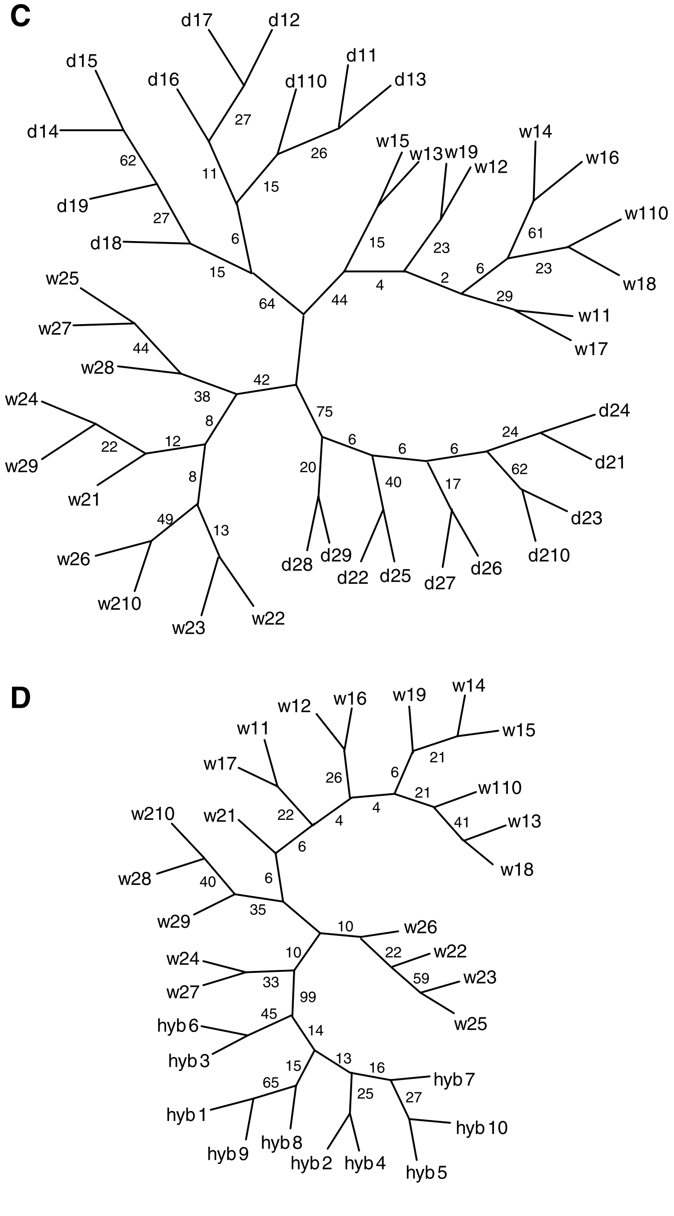


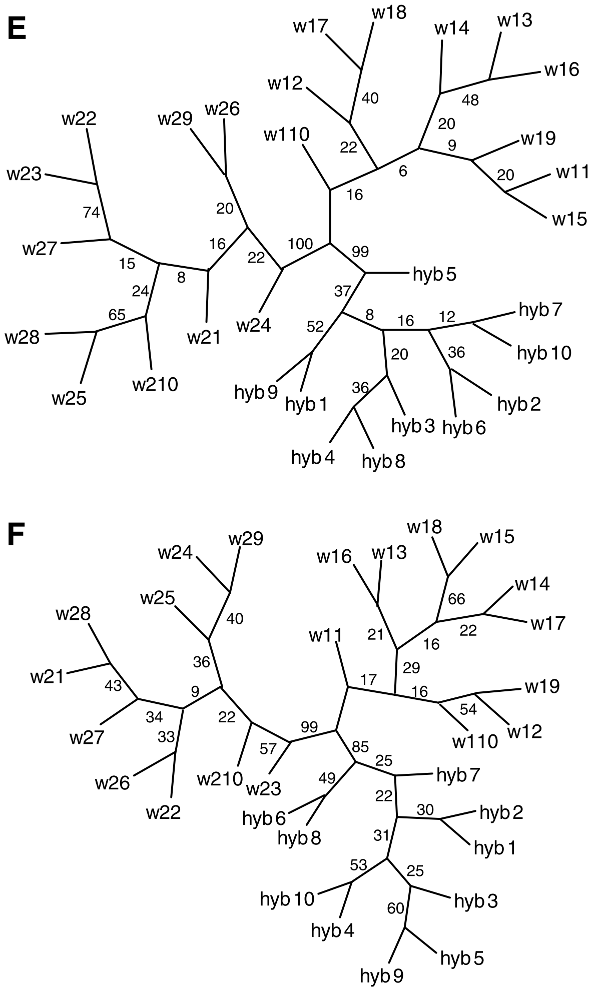
**Fig. 1** (*continued*).



lation of 100 individuals. Another period of random genetic drift was then simulated for 20 generations using the methodology described above for d1 and d2. Ten individuals were taken from the final population for the subsequent analyses.

DICE matrix generation and neighbor-joining analyses were carried out with the sets of 10 individuals taken from the wild and domesticated populations. Three analyses were made for each of the three starting scenarios: between w1, w2, d1, and d2; between w1, w2, and hyb5050; and between w1, w2, and hyb1090. Pairwise comparisons were made between individuals for each AFLP character. The total number of AFLP characters shared between a pair of individuals was calculated, as well as the number of characters unique to one or other of the individuals. These values were then

**Fig. 1** (*continued*).



used to calculate a coefficient of similarity between each pair of individuals (Dice 1945) as follows:

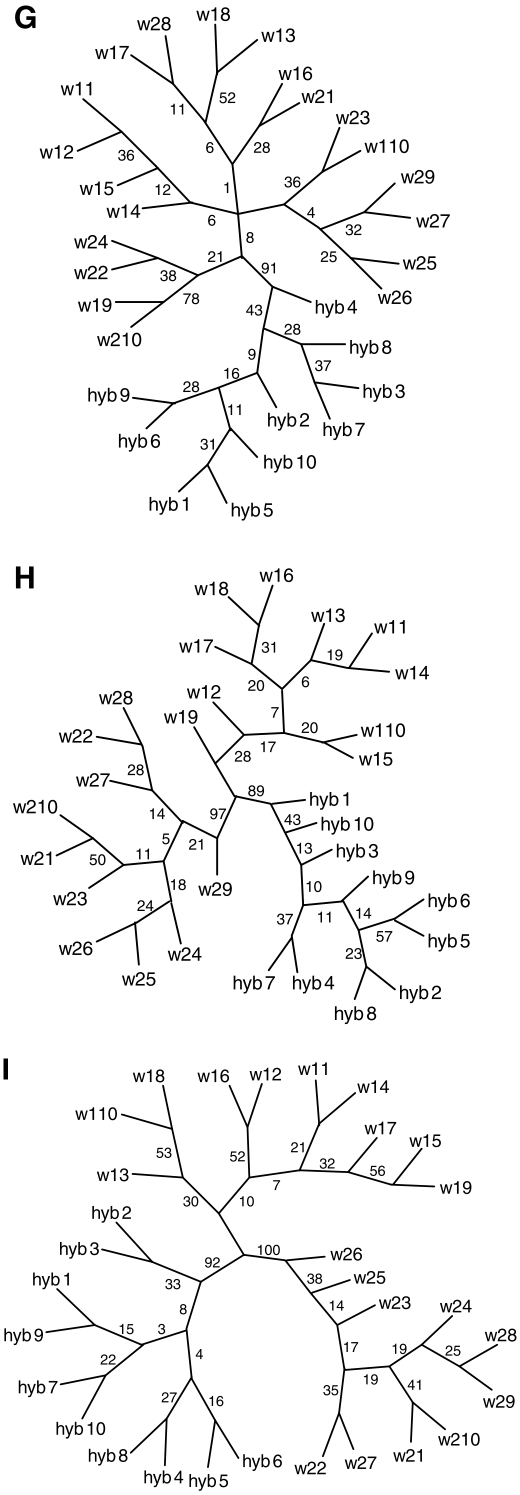
$$\text{similarity coefficient} = \frac{2a}{2a + b + c}$$

where *a* represents the number of characters shared by the pair of individuals, *b* represents the number of characters possessed by just one of the pair of individuals, and *c* represents the number of characters possessed by just the other member of the pair. The similarity values in a pairwise matrix were then converted to distance values using the following formula:

$$\text{distance value} = 1 - \text{similarity coefficient}$$

The distance matrix was then entered into the NEIGHBOR program of PHYLIP (Felsenstein 1991). The tree files were visualized using the UNROOTED tree viewing program. To evaluate the significance of the resulting trees, multiple runs of each simulation were carried out and individual trees were bootstrapped by carrying out 100 replicates.

**Fig. 1** (*continued*).



## Results

The results of the simulations are summarized in Table 1. Throughout the analyses, w1 and w2 rarely formed individual clades when they were assumed to have a common origin in the recent past (scenario 1), but always formed individual clades when they were assumed to be either more distantly related (scenario 2) or unrelated (scenario 3). In the first set of analyses, the two domesticated populations, d1 and d2, were non-hybridized (Table 1*a*). In the majority of the simulations, d1 and d2 formed monophyletic groupings and their independent origins were clearly evident. In all but 4 of the 90 simulations, the correct sister groups of d1 and d2 could be inferred from the relative positions of the domesticated and wild individuals in the tree. Examples of these trees are shown in Figs. 1A–1C. In Figs. 1B and 1C, the correct origins of d1 and d2 are evident and supported by reasonably high bootstrap values for the important branches; in Fig. 1A, the correct origins are less clear.

Archaeobotanically, the situation represented by the first set of analyses is unlikely, at least for wheat and barley. If a crop was domesticated twice in southwest Asia, then the individual domesticated populations would probably not have remained separate for long, as human movement would have brought them together thus leading to a hybrid population. The second and third analyses (Table 1*b* and 1*c*) are therefore more realistic. In most of the simulations involving hyb5050, the hyb5050 individuals formed a cluster around the center of the tree away from the wild individuals, and in almost half (41/90) of the simulations, this cluster formed a discrete clade (Table 1*b*). In these 41 simulations, the presence of this single clade could lead to the incorrect conclusion that the crop had a monophyletic origin. Furthermore, with many of these trees, one or other of the wild populations could erroneously have been inferred as the sister population, because the hyb5050 clade was usually located closer to, and sometime within, one of the wild clades. Figures 1D–1F show bootstrapped trees obtained with the hyb5050 populations. In all three trees, the bootstrap values give strong statistical support for the incorrect conclusion that the crop has a monophyletic origin. Figure 1D is an example of where topology further prompts the incorrect conclusion that w2 is the sister population to hyb5050.

If a crop has a polyphyletic origin, then the most likely situation is asymmetrical input from the wild populations, as modelled by 180 of our simulations (Table 1*c*). In 172 of these simulations, hyb1090 formed a single clade. In 90 of these simulations, the hybrid clade fell within one of the wild clades despite input from the other wild population. Examples are shown in Figs. 1G–1I. Again, all three trees indicate a monophyletic origin of the crop with high statistical support, and the tree shown in Fig. 1G additionally suggests that w2 is the sister population.

## Discussion

Our simulations show that the population biology aspects of the domestication process can result in independently domesticated populations merging in such a way that a monophyletic origin is erroneously inferred when the resulting population is examined by AFLP genotyping and neighbor-joining analysis. Furthermore, the tree topologies are such that, in some cases, a wild population would be incorrectly identified as the sister population to this non-authentic clade of domesticated plants.

Mutation of the AFLP loci, which was not considered by our simulation, would further obscure the polyphyletic origin of the hybrid domesticated population by introducing synapomorphies that would increase the genetic distance between the wild and domesticated plants. The fact that, in our simulations, hyb5050 and hyb1090 frequently appeared monophyletic even though the mutation rate was set at zero indicates that in the real world it is highly likely that a multiply domesticated crop would appear monophyletic when AFLP data are analysed by neighbor-joining.

Our work was prompted by the use of AFLPs and neighbor-joining to conclude that einkorn, emmer wheat, and barley were each domesticated on a single occasion at a geographical point that can be inferred from phylogenetic analysis (Heun et al. 1997; Badr et al. 2000; Özkan et al. 2002). These three studies have assumed an importance beyond plant genetics as they have been influential in the development of hypotheses regarding the human dynamics underlying the origins of agriculture in southwest Asia (e.g., Lev-Yadun et al. 2000; Diamond 2002). As such, we believe that it is important to be certain that their conclusions are correct. Implicit in the use of the neighbor-joining algorithm is the assumption that the markers being studied display complete linkage. In our model, we assumed that the markers are unlinked and show that if this is the case then neighbor-joining can produce erroneous results. It is of course true that within a collection of AFLPs some pairs of markers will display some degree of linkage, but we suggest that if a phylogenetic method like neighbor-joining is to be used, then it should first be established that the overall extent of linkage between the markers is sufficient for this choice of method to be valid. Until this is established for the AFLP datasets used by Heun et al. (1997), Badr et al. (2000), and Özkan et al. (2002), it is impossible, in our view, to consider that the conclusions they reach are proven. We do not suggest that cultivated einkorn, emmer, and barley must be polyphyletic, but neither do we believe that the AFLP studies establish their monophyly.

Although we specifically describe the markers that we use as “AFLPs” they are, in effect, “anonymous bands”. Markers such as RFLPs (restriction fragment length polymorphisms) and RAPDs (random amplified polymorphic DNA) would behave in an identical way in these simulations. Therefore, our conclusions regarding the possible invalidity of the results of phylogenetic analysis apply to any dataset obtained by anonymous band scoring.

## Acknowledgments

This research was supported by a grant from the U.K. Natural Environment Research Council. We thank the Editorial Board member and two anonymous referees for comments that have resulted in significant improvements to our paper.

## References

Abbo, S., Lev-Yadun, S., and Ladizinsky, G. 2001. Tracing the wild genetic stocks of crop plants. Genome, **44**: 309–311.

Badr, A., Müller, K., Schäfer-Pregl, R., El Rabey, H., Effgen, S., Ibrahim, H. H., Pozzi, C., Rohde, W., and Salamini, F. 2000. On the origin and domestication history of barley (*Hordeum vulgare*). Mol. Biol. Evol. **17**: 499–510.

Diamond, J. 2002. Evolution, consequences and future of plant and animal domestication. Nature (London), **418**: 700–707.

Dice, L.R. 1945. Measures of the amount of ecological association between species. Ecology, **26**: 297–302.

Felsenstein, J. 1991. PHYLIP (phylogeny inference package). Version 3.4. Department of Genetics, University of Washington, Seattle, Wash.

Heun, M., Schäfer-Pregl, R., Klawan, D., Castagna, R., Accerbi, M., Borghi, B., and Salamini, F. 1997. Site of einkorn wheat domestication identified by DNA fingerprinting. Science (Washington, D.C.), **278**: 1312–1314.

Kimura, M. 1962. On the probability of fixation of mutant genes in populations. Genetics, **47**: 713–719.

Lev-Yadun, S., Gopher, S.A., and Abbo, S. 2000. The cradle of agriculture. Science (Washington, D.C.), **288**: 1602–1603.

Özkan, H., Brandolini, A., Schäfer-Pregl, R., and Salamani, F. 2002. AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat in Southeast Turkey. Mol. Biol. Evol. **19**: 1797–1801.

Takahashi, R. 1972. Non brittle rachis 1 and non brittle rachis 2. Barley Genet. Newslett. **2**: 181–182.

Tanksley, S.D., and McCouch, S.R. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science (Washington, D.C.), **277**: 1063–1066.

Zohary, D. 1996. The mode of domestication of the founder crops of southwest Asian agriculture. *In* The origins and spread of agriculture and pastoralism in Eurasia. *Edited by* D.R. Harris. UCL Press, London, U.K. pp. 142–158.