Reevaluation of the use of multilocus markers to study crop origins

Brown T.A., Allaby R.G.

in

Molina-Cano J.L. (ed.), Christou P. (ed.), Graner A. (ed.), Hammer K. (ed.), Jouve N. (ed.), Keller B. (ed.), Lasa J.M. (ed.), Powell W. (ed.), Royo C. (ed.), Shewry P. (ed.), Stanca A.M. (ed.).
Cereal science and technology for feeding ten billion people: genomics era and beyond

Zaragoza : CIHEAM / IRTA
Options Méditerranéennes : Série A. Séminaires Méditerranéens; n. 81

2008
pages 19-24

Article available on line / Article disponible en ligne à l'adresse :

http://om.ciheam.org/article.php?IDPDF=800794

To cite this article / Pour citer cet article

Brown T.A., Allaby R.G. **Reevaluation of the use of multilocus markers to study crop origins.** In : Molina-Cano J.L. (ed.), Christou P. (ed.), Graner A. (ed.), Hammer K. (ed.), Jouve N. (ed.), Keller B. (ed.), Lasa J.M. (ed.), Powell W. (ed.), Royo C. (ed.), Shewry P. (ed.), Stanca A.M. (ed.). *Cereal science and technology for feeding ten billion people: genomics era and beyond.* Zaragoza : CIHEAM / IRTA, 2008. p. 19-24 (Options Méditerranéennes : Série A. Séminaires Méditerranéens; n. 81)

http://www.ciheam.org/
http://om.ciheam.org/

Reevaluation of the use of multilocus markers to study crop origins

T.A. Brown* and R.G. Allaby**
*Manchester Interdisciplinary Biocentre, University of Manchester,
131 Princess Street, Manchester M1 7ND, UK
**Warwick HRI, Wellesbourne, Warwickshire CV35 9EF, UK

SUMMARY – Crop origins are the subject of extensive debate, in particular with regards to the number and locality of the underlying domestication events. Multilocus studies using AFLPs or microsatellites have consistently inferred single origins for crops based on the monophyletic phylogenies of the domesticated varieties. Yet phylogenetic evidence from individual genetic loci suggests multiple origins for some of these crops. We have carried out simulations of crop domestication in which the behaviour of genome-wide multilocus systems are studied. We find that there is a tendency for multilocus analysis to assign a single origin to a crop even when that crop derives from multiple domestications. Indeed, crops that have multiple origins are actually more likely to give rise to a monophyletic phylogeny than those that have a single origin. Our results indicate that the probability of monophyly is largely dependent upon the size and age of the domesticate population, which means that multilocus systems can be used to establish the upper limit to the number of plants originally brought into cultivation.

Introduction

Considerable interest centres on the origins of crops, in particular regarding the number and locations of the domestication events. This information is important both to archaeologists, for whom the shift from hunting-gathering to agriculture represents one of the most significant transitions in human social evolution, and to crop biologists, who anticipate that a better understanding of the genetical and evolutionary events underlying domestication will aid breeding programmes aimed at supplementing the diversity of modern crops. Because of the relatively short period of time that has elapsed since the beginning of agriculture – c.10,000 years for the Southwest Asian cereals – it might be expected that modern crop plants will still display relatively close genetic affiliations with their wild progenitors, making it possible to identify these progenitors by examining the relationships between cultivated and wild populations. When relationships between populations are studied, it is commonly assumed that the accuracy of the information obtained increases in roughly direct proportion to the number of genetic loci that are examined. For this reason, in the post-genomics era, it has become commonplace to use multilocus systems to examine crop origins, these studies including one to several hundred separate genetic loci, possibly biallelic ones such as AFLPs (e.g. Heun *et al*., 1997 for einkorn; Badr *et al*., 2000 for barley; Ozkan *et al*., 2002 for tetraploid wheat; Spooner *et al*., 2005 for potato) or multiallelic microsatellites (e.g. Olson and Schaal, 2001 for cassava; Matsuoka *et al*., 2002 for maize). However, it is unclear how accurately molecular phylogenetic techniques such as tree-building, which have been developed for comparisons between single genetic loci within which the informative sites are closely linked, perform when applied to loci that reside on separate, segregating chromosomes, with even those loci present on a single chromosome possibly not displaying complete linkage because of recombination. To evaluate the veracity of these approaches to crop origins, we have examined the theoretical expectations of multilocus systems given different domestication scenarios. In this paper we summarise the results of this work and assess the implications that the results have for future studies of domestication.

Simulations of multiple domestication scenarios

One of the most striking features of the multilocus studies that have so far been carried is that all have indicated a monophyletic origin for the crop under study (Heun *et al*., 1997; Badr *et al*., 2000; Olson and Schaal, 2001; Matsuoka *et al*., 2002; Ozkan *et al*., 2002; Spooner *et al*., 2005). This is surprising because in several instances the prevailing view from more limited genetic studies had been that the crop in question was domesticated more than once. For example, domesticated maize is monophyletic according to the microsatellite analysis of Matsuoka *et al*. (2002), but the extensive cytogenetic work of McClintock *et al*. (1981) clearly shows that South American maize comprises two distinct germplasms, the Andean complex and the Venezuelan complex, a division which is also evident when morphological traits are compared. The existence of distinct complexes of germplasm had been widely taken as an indication that maize was domesticated more than once. Similarly, AFLP analysis suggests that cultivated potato (Spooner *et al*., 2005) and barley (Badr *et al*., 2000) both derive from single domestications whereas examination of chloroplast loci strongly support multiple origins for both these crops (Hosaka, 1995; Molina-Cano *et al*., 2005). We therefore began our study of the theoretical expectations of multilocus systems by evaluating how these give rise to a monophyletic appearance for a domesticated crop.

Multilocus systems can erroneously infer monophyly when it does not exist

Our initial simulations (Allaby and Brown, 2003) were based on 200 imaginary biallelic characters, each of which were taken to represent 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, one pair having a common origin in the recent past, one pair having a more distant origin but still discernibly related to one another, and one pair having diverged an infinitely long time ago. Domesticated populations were produced from the wild populations by a process assumed to involve a population bottleneck followed by a period of random genetic drift (Tanksley and McCouch, 1997). A population expansion was then simulated followed by a second period of random genetic drift during which the population size was constant. We then simulated situations where two independently domesticated populations join to form a hybrid population. Two models were considered, the first with pairs of domesticates having equal inputs in the hybrid population, and the second with one domesticate contributing 10% of the input into the hybrid and the other domesticate contributing the other 90%. Random genetic drift was then simulated for a further 20 generations.

Pairwise comparisons were made between individuals for each AFLP character and the total number of 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 used to calculate a coefficient of similarity between each pair of individuals, according to Dice (1945). The similarity values in a pairwise matrix were then converted to distance values, and the distance matrix entered into the NEIGHBOR program of PHYLIP (Felsenstein, 1991). 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.

The results of the simulations are summarized in Table 1. Throughout the analyses, the wild populations rarely formed individual clades when these starting populations were assumed to have a common origin in the recent past, but always formed individual clades when the starting populations were assumed to be more distantly related. If the two domesticated populations were kept separate (Table 1A) then in the majority of the simulations they formed monophyletic groupings, as expected, and their independent origins were clearly evident. In all but four of the 90 simulations the correct sister groups of the domesticated populations could be inferred from the relative positions of the domesticated and wild individuals in the tree. An example of this type of tree is shown in Fig. 1A. In this tree, the independent origins of the domesticates and the identities of their sister populations are evident and supported by reasonably high bootstrap values for the important branches.

Archaeobotanically, the situation represented by this first set of analyses is unlikely. If a crop was domesticated twice then the individual domesticated populations would probably not have remained separate for long, human movement bringing them together leading to a hybrid population. The second and third analyses, involving hybridisation between domesticates (Table 1B and C), are therefore more realistic. In most of the simulations in which the domesticates had equal input, individuals from the hybrid populations formed a cluster around the centre of the tree away from the wild individuals, and in almost half (41/90) of the simulations this cluster formed a discrete clade (Table 1B). 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 hybrid clade was usually located closer to, and sometime within, one of the wild clades. Fig. 1B shows a tree obtained with this scenario, with bootstrap values giving strong statistical support for the incorrect conclusion that the crop has a monophyletic origin and that *w2* is the sister population. This outcome is exacerbated when the hybrid population has an asymmetrical input from the original domesticates (Table 1C). In 172 of these 180 simulations, the hybrid population formed a single clade, and in 90 simulations the hybrid clade fell within one of the wild clades, despite the input from the other wild population. Am example is shown in Fig. 1C. Again, the tree indicates a monophyletic origin of the crop with high statistical support, and suggests that *w2* is the sister population.

Table 1. Summary of the outcomes of the multilocus simulations

Outcome†	Divergence between wild populations		
	Recent	Distant	Infinite
A. Comparison between *w1*, *w2*, *d1* and *d2*			
Number 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*			
Number 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*			
Number 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

†Abbreviations: *w1*, *w2*, wild populations; *d1*, *d2*, domesticated populations; *hyb5050*, hybrid domesticate with equal inputs from progenitor populations; *hyb1090*, hybrid domesticate with unequal inputs from progenitor populations.

According to multilocus systems, all crops tend towards monophyly over time

To investigate the behaviour of multilocus systems in greater detail, we carried out a second set of simulations (Allaby *et al*., submitted) in which we modelled the domestication of a monoecious, annual, amphidiploid, hypothetical crop plant with a basic haploid number of 20 chromosomes. Again, we simulated a biallelic system, equivalent to AFLPs. The basis of the model was to generate chromosomes for individuals of a population, and then to produce subsequent generations by uniting gametes generated from those individuals while subjecting the plants to processes associated with domestication such as bottlenecks and the amalgamation of unrelated populations. Several types of hypothetical crop plant were considered, differing in their tendency toward homologous recombination during meiosis. As in the previous simulation, at the end of the domestication process, individuals were randomly selected and trees built by neighbour-joining. The phylogenetic analyses show that, for each set of conditions that was tested, a hybrid domesticated population had a tendency towards monophyly that increased with generation number in an S-shaped curve. For plants displaying zero or 0.01 recombination frequency, the number of generations that elapsed before there was a 50% occurrence of monophyly was closely similar to the original population size of this hybrid domesticate. Plants that displayed recombination frequencies of 0.1 or 1 gave rise to monophyletic populations after fewer generations. A further set of simulations was carried out without the formation of hybrid populations, so that the final crop was the product of a single domestication. To our surprise, the populations derived from single domestications became monophyletic more slowly than the hybrid populations.

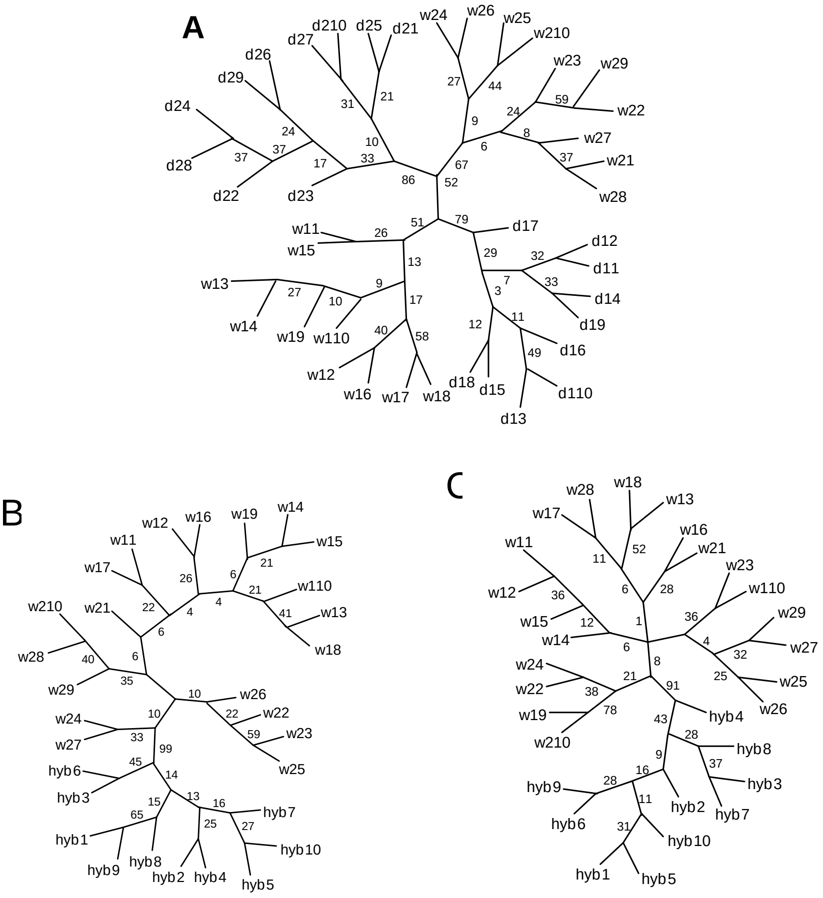

Fig. 1. Typical neighbour-joining trees obtained after multilocus simulation. (A) Domesticate populations kept separate. (B) Hybrid domesticate has equal input from progenitor populations. (C) Hybrid domesticate has 10% input from *w1* and 90% input from *w2*. The degrees of divergence of the wild populations were (A) distant, (B) and (C) recent. Bootstrap values were obtained by carrying out 100 replicates. Individuals are labelled d1n if taken from domesticated population *d1*, d2n if from *d2*, hyb if from a hybrid population, w1n if from wild population *w1*, and w2n if from *w2*.

Reasons for the problems with multilocus systems

Our simulations show that, when viewed through biallelic multilocus systems, crops that have undergone multiple domestication can appear to be monophyletic. We believe that the reasons for this lie with the methodology that has been used to analyse multilocus data with real crops (and which we have copied in our simulations), this methodology failing to recognise the nature of the loci being analysed and failing to take account of the population biology events underlying the domestication process. Implicit in the use of the neighbour-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 neighbour-joining can produce erroneous results. It is, of course, true that within a collection of genome-wide markers, such as AFLPs, some pairs of markers will display complete or a lesser degree of linkage, but this is a very different situation from the tight linkage that exists between *all* the informative sites within a single locus, this being the type of molecular information for which neighbour-joining has a proven track record.

The impact of the population biology events underlying domestication are illustrated by the superficially counterintuitive conclusion from our second set of simulations that, when viewed through multilocus systems, multiple domestications are actually more likely to result in monophyly than single domestication scenarios. In fact it is quite reasonable to suppose that a population derived from a single domestication will be slower to achieve monophyly than a multiply domesticated crop. This is because a population derived purely from a single wild source will differentiate from that wild source solely by changes in the frequencies of alleles that are present in both the wild and domesticated populations. In contrast, the amalgamation of independently domesticated populations results in a hybrid domesticate that contains some alleles that are present in only one of the two wild parent populations. The presence of these non-shared alleles means that immediately after amalgamation a multiply domesticated population will be more likely to appear polyphyletic than will a singly domesticated population, as was evident in our simulations, where we observed that during the first few generations the singly domesticated populations displayed a slightly increased number of instances of monophyly (Allaby *et al*., submitted). But the spread of the non-shared alleles through the hybrid domesticate will result in this hybrid subsequently appearing monophyletic more quickly than the population resulting from single domestication. When factors such as this are taken into consideration, it is not surprising that simplistic analyses such as neighbour-joining fail to identify the true origins of the hybrid crop populations generated during our simulations.

Implications for studies of the origins of real crops

As described above, each of the multilocus studies that have been carried out with real crops has inferred a monophyletic origin, despite the alternative genetic evidence suggesting that at least several of these crops are derived from multiple domestications. The results of our simulations suggest that these apparently contradictory findings can be explained by the tendency for multilocus analysis to assign a single origin to a crop even when that crop derives from multiple domestications. In short, multilocus systems, at least when analysed by tree-building, are unable to provide reliable information on crop origins, and the veracity of those studies that have utilised this approach should be reevaluated.

We should also reevaluate the way in which genetic data is being used to infer the human dynamics underlying the transition from hunting-gathering to agriculture. Even if the conclusions of the multilocus projects are correct, and many crops are monophyletic, it is not reasonable to extrapolate to the assumption that each of these crops was taken into cultivation just once. Part of the problem is that concepts such as "monophyly", which have clear meanings and implications when the evolution of species are studied, become much less determinative when applied to populations of a single species. A modern crop could appear to be monophyletic because it originated from a single domestication event, but monophyly could equally well arise due to events occurring after the initial cultivations. Salamini *et al*. (2002) point out that there are inconsistencies between the apparent monophyly displayed by the key founder crops of Southwest Asian agriculture and the gradual transition from gathering to cultivation to domestication than is apparent in the archaeological record for at least some of these crops. Discoveries such as those at Gilgal in the Jordan Valley, where granaries containing hundreds of thousands of wild barley and oat grains have been dated to over 11,000 years ago (Weiss *et al*., 2006), are making it increasingly clear that wild cereals were harvested well before the domesticated forms appear in the archaeobotanical record. This has led to suggestions that there might have been a substantial period of "predomestication cultivation" of crop plants in Southwest Asia (Weiss *et al*., 2006), meaning a period when plants were cultivated (i.e. deliberately planted and harvested by humans) but not yet domesticated (i.e. had not undergone the genetic selection, requiring reproductive isolation, that gives rise to the altered genotypes and phenotypes of the domesticated variety) (Jones and Brown, in press). A period of predomestication cultivation lasting a thousand years or more is likely to give rise to a non-monophyletic crop, because during this period the crop is not reproductively isolated from neighbouring wild populations. Hybridization with different wild populations as cultivation spreads through the natural range of the species would give the fully domesticated crop a genetic signature that is much more mixed than that anticipated in a monophyletic population. To solve this contradiction between predomestication cultivation and the monophyletic outcomes of multilocus studies, Salamini *et al*. (2002) suggest that genetic monophyly might arise if for each crop a superior landrace emerges from the variety of forms generated by the initial cultivations, this superior landrace subsequently spreading and becoming the progenitor of all the modern landraces and cultivars that are sampled. Whether or not this hypothesis is correct, considerations such as this show that there is difficulty in linking studies of the genetics of modern crops to archaeological questions regarding agricultural origins. In this particular example, it cannot be assumed that the superior landrace is descended from the first wild plants to be cultivated and it may not even be the first cultivated population to become domesticated. The geographical origin of this superior landrace therefore cannot identify the location of the farming communities that first took the wild plants into cultivation nor can it identify the location of those, possibly different, communities whose cultivated forms first became transformed into domesticated varieties. Indeed, if predomestication cultivation occurred then these questions might not even have archaeological relevance.

Conclusions

This paper takes a rigorous view of the contribution that genetics is making to our understanding of the origins of crops. Our simulations suggest that the monophyletic conclusions drawn from studies of multilocus systems in wheat, barley, etc., are unproven, and even if they are correct, it is questionable what interpretations can be drawn from these conclusions. However, the data from multilocus studies are not worthless, and a sound appreciation of the theoretical expectation of these systems will enable the real evolutionary meaning of the data to be determined. Once that stage has been reached, collaborations between plant geneticists and archaeobotanists will enable the evolutionary implications of multilocus studies to be placed in their correct archaeological context.

References

Allaby, R.G. and Brown, T.A. (2003). *Genome*, 46: 448–453.

Allaby, R.G., Merriwether, D.A. and Brown, T.A. (submitted).

Badr, A., Müller, K., Schäfer-Pregl, R., El Rabey, H., Effgen, S., Ibrahim, H.H., Pozzi, C., Rohde, W. and Salamini, F. (2000). *Molecular Biology and Evolution*, 17: 499–510.

Dice, L.R. (1945). *Ecology*, 26: 297–302.

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

Heun, M., Schäfer-Pregl, R., Klawan, D., Castagna, R., Accerbi, M., Borghi, B. and Salamini, F. (1997). *Science*, 278: 1312–1314.

Hosaka, K. (1995). *Theoretical and Applied Genetics*, 90: 356–363.

Jones, M.K. and Brown, T.A. (in press). In: *Rethinking Agriculture* (Denham, T.P. *et al*., ed.)

Matsuoka, Y., Vigouroux, Y., Goodman, M.M., Sanchez, J., Buckler, E. and Doebley, J. (2002). *Proceedings of the National Academy of Sciences*, USA 99: 6080–6084.

McClintock, B., T. A. Kato Y. and A. Blumenschein (1981). *Chromosome constitution of races of maize. Its significance in the interpretation of relationships between races and varieties in the Americas*. Colegio de Postgraduados, Chapingo.

Molina-Cano, J.-L., Russell, J.R., Moralejo, M.A., Escacena. J.L., Arias, G. and Powell, W. (2005). *Theoretical and Applied Genetics*, 110: 613–619.

Olsen, K.M. and Schaal, B.A. (2001). *American Journal of Botany*, 88: 131–142.

Özkan, H., Brandolini, A., Schäfer-Pregl, R. and Salamini, F. (2002). *Molecular Biology and Evolution,* 19: 1797–1801.

Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R. and Martin, W. (2002). *Nature Reviews Genetics*, 3: 429–441.

Spooner, D.M., McLean, K., Ramsay, G., Waugh, R. and Bryan, G.J. (2005). *Proceedings of the National Academy of Sciences*, USA 102, 14694–14699.

Tanksley, S.D. and McCouch, S.R. (1997). *Science*, 277: 1063–1066.

Weiss, E., Kislev, M.E. and Hartmann, A. (2006). *Science*, 312: 1608–1610.