Supporting Information

Allaby *et al.* 10.1073/pnas.0803780105

SI Text

The following sections outline the specific methodology and reasoning of the population simulation program written by R.G.A.

Chromosome Phylogeny. To produce linked markers, one must have them arranged on chromosomes. Computationally, one could simply assign presence or absence of a marker to each locus randomly to produce a single genotype. However, this is problematic in at least two ways. Firstly, some loci may never become assigned with a marker present, resulting in a decrease in the markers actually used. Secondly, the expected frequency of all markers will be equal, and their presence independent (i.e., in perfect linkage equilibrium). The latter condition is not too problematic or unrealistic—although fairly contrary to the model condition here of maintaining markers in perfect linkage disequilibrium, but the former is quite unrealistic biologically. For instance, markers of greater antiquity may well have higher frequencies that those of more recent origin. Additionally, markers of recent origin may tend to occur with the markers that were present on the chromosome when they arose. To address these concerns of a realistic marker dependency as well as ensuring all loci became populated with a marker in at least one genotype, a chromosome phylogeny methodology was adopted which would simulate the evolutionary history of the chromosome. The following methodology was used: The phylogeny begins with the ancestral chromosome that had all loci set to 0 (no marker present). A locus was then randomly selected and mutated to 1, giving two genotypes. The process was iterated by first selecting a genotype, then a locus to mutate, and the resulting genotype added to the genotype list. Loci were only permitted to mutate once. The resulting list of genotypes then represented a randomized phylogeny. The simulation used a parameter of a haploid number of chromosomes of 20—each of these had a phylogeny independently generated.

1. Wild Population Genotype Frequencies. In recognition that each genotype may be at any point of fixation or loss within the population, each genotype was assigned a frequency. To avoid an unrealistically even spread of frequencies, the frequency allocation method used a finite population size. The method began with 10,000 individuals from which a random number was selected (with equal probability), without replacement, as the first frequency value. The frequency was assigned to a genotype (with equal probability). The second frequency was then selected from the remaining number of individuals and assigned to a second random genotype. This process was repeated until all of the individuals had been used up. When generating two wild populations in the same simulation, which would represent two separate cultivated crop origins, the same phylogeny was used but genotype frequencies were allocated independently. The resulting populations were differentiated by an Fst value of 0.25 on average, and each had a heterozygosity of 0.51 on average.

2. Individual Generation. A small number of individuals were drawn from these large wild populations. Individuals were 'built' by selecting a genotype for each chromosome. A genotype was selected from the genotype phylogeny of a particular chromosome with a probability equal to the proportion of the genotype frequency of the total population size (10,000). Each individual had two genotypes selected for each chromosome representing a diploid organism.

3. Generation Propagation. The virtual plants reproduced with panmixis in the following way. Two parent plants were selected at random. Each parent would then produce a gamete by random segregation. This was achieved by randomly choosing with equal probability one of the two genotypes available for each chromosome. During this process, meiosis was possible with a user-defined probability. A meiotic event was achieved by randomly selecting a point along the chromosome, as defined by the biallelic marker loci, with equal probability, which in effect was a random point between two adjacent addresses in the binary array that described the chromosome genotype. This method assumes loci are evenly distributed across the chromosome. The two arrays representing homologous chromosomes of an individual were then split and swapped at this point. Individuals were repeatedly generated in this way until the desired population size had been reached.

4. Hybrid Population. In the case of multiple origin-cultivated population scenarios, the above steps were carried out in parallel to generate a specified number of cultivated populations that had been independently drawn from different wild populations (but related by the same genotype phylogeny), subject to a bottleneck, and then expansion. The hybrid population was a panmictic population made up of the contributing populations in defined proportions. The contributing populations were used to make a hybrid population in the following way. Each population contributing to the hybrid population had a user-defined ratio of input. In the study a 50:50 input was used. Two populations were selected at random using the ratio proportions as the probability of selection. The same population could be selected twice using this method. Then two parents were selected at random from each of those populations and gametes made in the usual way described above. Individuals of the hybrid generation were drawn until the desired population size had been reached. Subsequent generations were propagated from these individuals in the usual way described above.

5. Analysis. At the end of the simulation a defined number of individuals from each population were selected at random for the analysis. In the case of the wild populations, this meant generation of new wild individuals using the original genotype frequencies. This is biologically reasonable, because it was assumed that the wild population was very large, and consequently would have changed negligibly in allele frequency over the timescale of the simulation. Individuals were drawn from the last independent cultivated populations available before hybrid populations were formed. Finally, individuals were drawn from the last generation of the simulation. In the case of the single origin, the last independent cultivated population was also the last generation of the simulation. Biallelic markers are naturally dominant—even if only one homologous chromosome carries a certain marker it will be identified as present in an analysis, such as AFLP. To accommodate this, the binary array pairs representing the chromosome genotypes of each individual were merged, with 1s indicating marker presence dominating 0s indicating marker absence. Pairwise comparisons were then made between individuals over all (merged) chromosomes, in which all loci with a marker present in one or both individuals contributed to the algorithm of Dice [similarity = $2a/(2a + b + c)$] where a is the number of loci with markers present in both individuals, and b and c are the number of loci in which a marker is only present in either one or the other individual respectively.

The similarity values generated from the Dice algorithm were inverted by subtracting from 1.

The simulation ended with the generation of the inverted similarity value matrix. The program was executed 100 times over using a command script generating 100 such matrices for each set of parameter conditions. A second command script input each inverted values matrix into the NEIGHBOR program to generate a tree, resulting in 100 tree files. Using a second program by R.G.A., tree files were then assessed for monophyletic clades by establishing that the cultivated individuals were united by a single branch to the exclusion of all other individuals in the analysis.

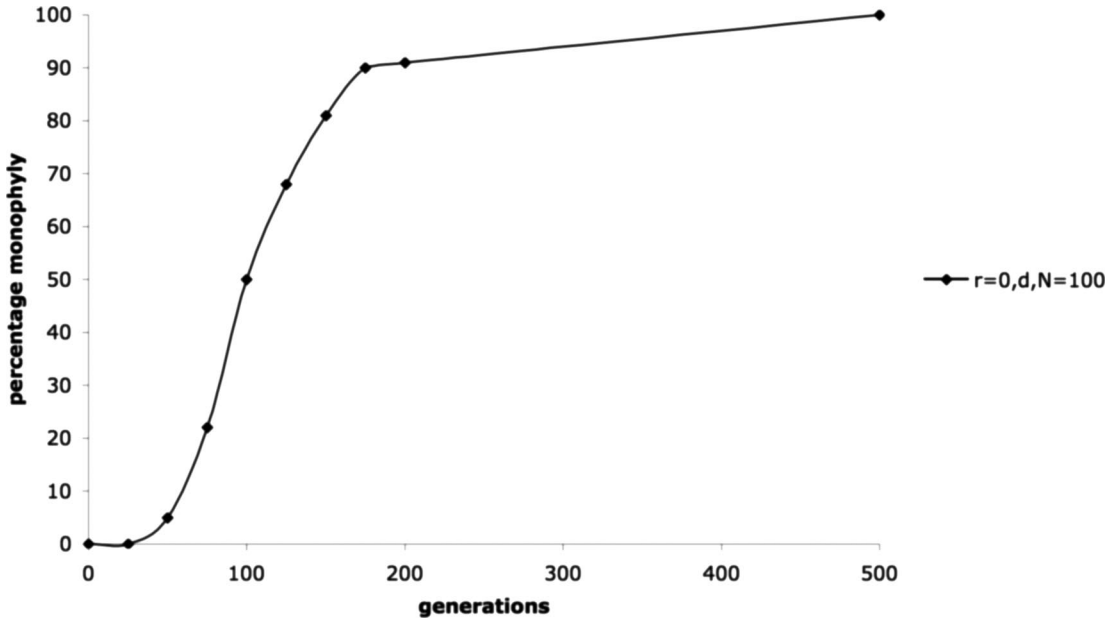

Fig. S1. Percentage of simulations that led to a monophyletic clade of cultivated plants over time. Conditions were set at no recombination, long-term population size 100 individuals, and a double origin of plants (key as in Fig. 2). After 500 generations, 100% of simulations result in monophyly, demonstrating that the plateau of the S curve is not an asymptote.

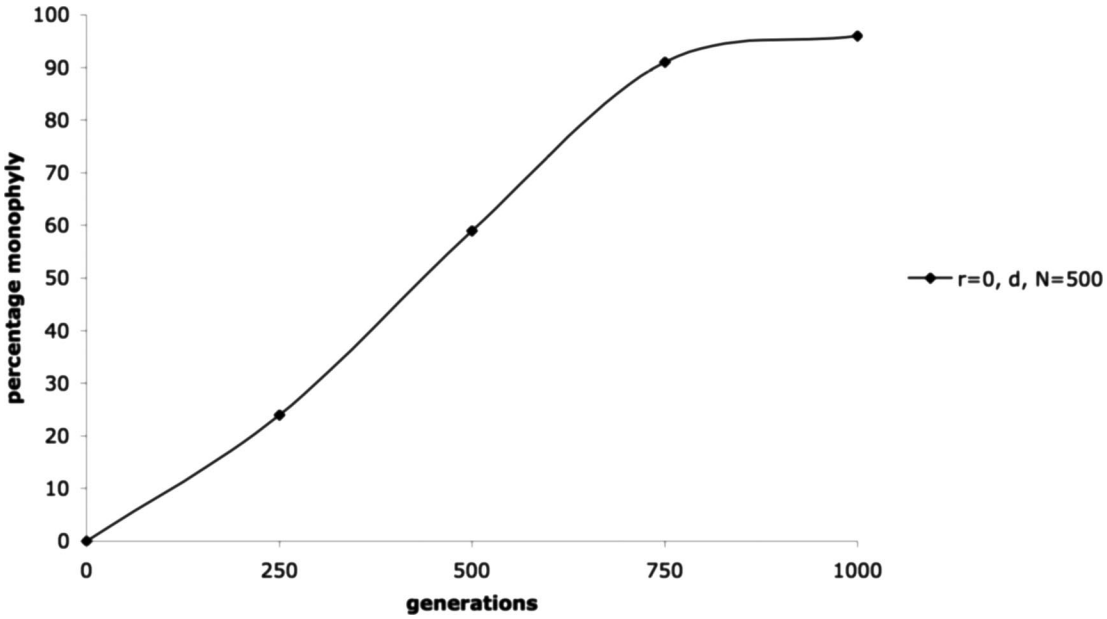

Fig. S2. Percentage of simulations that led to a monophyletic clade of cultivated plants over time. Conditions were set at no recombination, long-term population size 500 individuals, and a double origin of plants (key as in Fig. 2). A similar relationship between time in generations to monophyly and population size is observed as in Fig. 2, with substantially larger populations with a plateau reached by 2N generations.