*Annals of Botany* 1–13, 2007
doi:10.1093/aob/mcm151, available online at www.aob.oxfordjournals.org

OXFORD JOURNALS
OXFORD UNIVERSITY PRESS

# Tagging the Signatures of Domestication in Common Bean (*Phaseolus vulgaris*) by Means of Pooled DNA Samples

ROBERTO PAPA[1,*], ELISA BELLUCCI[1], MONICA ROSSI[1], STEFANO LEONARDI[2], DOMENICO RAU[1], PAUL GEPTS[3], LAURA NANNI[1] and GIOVANNA ATTENE[4]

[1]*Dipartimento di Scienze degli Alimenti, Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy,* [2]*Dipartimento di Scienze Ambientali, Università di Parma, Parco Area delle Scienze 11A, 43100 Parma, Italy,* [3]*Department of Plant Sciences, Section of Crop and Ecosystem Sciences, UC Davis, 1 Shields Avenue, Davis, CA 95616-8780, USA and* [4]*Dipartimento di Scienze Agronomiche e Genetica Vegetale Agraria, Università di Sassari, Via De Nicola, 07100 Sassari, Italy*

Received: 27 September 2006 Returned for revision: 4 December 2006 Accepted: 22 May 2007

• *Background and Aims* The main aim of this study was to use an amplified fragment length polymorphism (AFLP)-based, large-scale screening of the whole genome of *Phaseolus vulgaris* to determine the effects of selection on the structure of the genetic diversity in wild and domesticated populations.
• *Methods* Using pooled DNA samples, seven each of wild and domesticated populations of *P. vulgaris* were studied using 2506 AFLP markers (on average, one every 250 kb). About 10 % of the markers were also analysed on individual genotypes and were used to infer allelic frequencies empirically from bulk data. In both data sets, tests were made to determine the departure from neutral expectation for each marker using an $F_{ST}$-based method.
• *Key Results* The most important outcome is that a large fraction of the genome of the common bean (16 %; $P < 0{\cdot}01$) appears to have been subjected to effects of selection during domestication. Markers obtained in individual genotypes were also mapped and classified according to their proximities to known genes and quantitative trait loci (QTLs) of the domestication syndrome. Most of the markers that were found to be potentially under the effects of selection were located in the proximity of previously mapped genes and QTLs related to the domestication syndrome.
• *Conclusions.* Overall, the results indicate that in *P. vulgaris* a large portion of the genome appears to have been subjected to the effects of selection, probably because of linkage to the loci selected during domestication. As most of the markers that are under the effects of selection are linked to known loci related to the domestication syndrome, it is concluded that population genomics approaches are very efficient in detecting QTLs. A method based on bulk DNA samples is presented that is effective in pre-screening for a large number of markers to determine selection signatures.

**Key words:** AFLP, domestication, *Phaseolus vulgaris*, population genomics, QTLs, selection.

## INTRODUCTION

Over the last few years, the process of crop domestication has become of increasing interest among plant scientists, with quantitative trait locus (QTL) analysis and gene cloning providing the new tools to address better those questions related to the origins and duration of domestication and the identification of the genetics and molecular basis of domestication (Gepts and Papa, 2002; Gepts, 2004). However, crop domestication not only represents a milestone in the history of human beings and a unique model for the study of evolutionary processes, but recent studies have clearly indicated that the use of wild relatives can have a tremendous impact on crop improvement (Tanskley and McCouch, 1997; McCouch, 2004). For example, in tomato, the use of wild relatives has been shown to increase production by about 50 % compared with modern cultivars (Gur and Zamir, 2004). In the common bean (*Phaseolus vulgaris* L.), preliminary studies also show the potential to improve the yield of domesticated beans through genetic diversity from wild beans (Singh *et al.*, 1995; Kelly, 2004; Blair *et al.*, 2006).

The main reason for these accomplishments is the introduction of novel diversity that was not included during the domestication process, because of the reduction in diversity imposed by founder effects and selection at target loci. Not only have useful variants been fixed into the original domesticated pools, but additional genes that may be useful for plant breeding have lost their diversity through genetic drift. Whereas gene flow after domestication may have partially restored the diversity included in the domesticated pools (Jarvis and Hodgkin, 1999), selection by farmers for key traits of domestication, facilitated by the recessive nature of most of the domesticated alleles, has probably affected linked regions of the genome. The reduction of genetic diversity has also been more drastic for crops that had evolved as a different species from their respective wild progenitor(s) (e.g. *Triticum aestivum* and *Solanum lycopersicon*). These effects should be particularly strong in autogamous species that have restricted effective recombination, as compared with allogamous species, and because

* For correspondence. E-mail r.papa@univpm.it

autogamous species present a higher population structure that has probably enhanced the domestication bottleneck.

There are several problems associated with the introduction of exotic germplasm, such as the dominance of most wild alleles at domestication loci, the occurrence of deleterious traits that tend to mask useful variants, the presence of various levels of incompatibility between domesticated crops and their wild relatives, the need to develop an efficient identification system for the most promising wild genotypes to build segregating populations and the high segregation distortion that limits the amount of introgression from the wild source (Tanskley and McCouch, 1997). Thus, to exploit better the genetic diversity present in the wild relatives of a crop, a knowledge of the location of genes involved in the domestication syndrome and the proportion of the genome affected by domestication appears to be crucial. As has been shown in *P. vulgaris*, the regions linked to the domestication loci have probably been less exploited historically by farmers and breeders, and they are those where the highest diversity of the wild relatives is located (Papa *et al.*, 2005).

Because of the combined effects of selection and recombination, it is potentially possible to detect the signature of selection from variant patterns of allelic frequencies, as compared with neutral expectation (Cavalli Sforza, 1966; Lewontin and Krakauer, 1973). Lewontin and Krakauer (1973) were the first to propose the use of population divergence estimator-based methods (e.g. $F_{ST}$) to identify loci putatively under the effect of selection. Although their work has been highly criticized and was later abandoned, more recently it has been revised and re-evaluated as a promising tool for the identification of loci involved in the genetic control of adaptive phenotypic variants (for a review, see Beaumont, 2005). Several methods have been proposed for such an approach, which are based not only on divergence, such as $F_{ST}$, but also on diversity estimators (e.g. heterozygosity; for reviews, see Nielsen, 2005; Storz, 2005). These approaches that have recently been referred to as 'population genomics' have great potential for the detection of QTLs for adaptive traits, with a better resolution than conventional QTL analysis and without the need for *a priori* knowledge of the phenotypic trait or the candidate gene that may be responsible for the adaptive response.

DNA pools are an efficient way to pre-screen a large number of markers. This approach has been used in genome association studies between molecular markers [restriction fragment length polymorphisms (RFLPs) and single nucleotide polymorphisms (SNPs)] and loci controlling human disease (Arnheim *et al.*, 1985; Hinds *et al.*, 2004; Butcher *et al.*, 2005), and to analyse pools of genotypes (bulks) showing a divergent phenotype in segregant populations [bulk segregant analysis (BSA); Michelmore *et al.*, 1995]. This approach could thus also be useful in population genomics-based studies.

The common bean (*P. vulgaris*) has been domesticated independently in Mesoamerica and in the Andes (Gepts, 1998). The process of its domestication has been studied in detail and the major domestication traits have been mapped (Koinange *et al.*, 1996). Papa *et al.* (2005) recently showed that genes for domestication are located in regions of high divergence between wild and domesticated *P. vulgaris*, demonstrating that for the common bean, a genome scan for the signature of domestication is a promising approach.

The main objectives of this study were to: (*a*) determine the effects of selection on the structure of genetic diversity in wild and domesticated *P. vulgaris* from Mexico; (*b*) identify the loci under selection and evaluate the potential for population genomics combined with the use of DNA pools to identify QTLs in the common bean; and (*c*) using DNA pools, develop molecular markers that tag genomic regions harbouring domestication genes in *P. vulgaris*.

## MATERIALS AND METHODS

### *Plant material*

A subset of 38 genotypes was used that had been studied previously by Papa and Gepts (2003) and which includes 19 wild and 19 domesticated *P. vulgaris* L. individual plants collected during an expedition to the Mexican states of Chiapas, Oaxaca, Puebla and Jalisco, from sites where the two forms are present at different degrees of sympatry (Table 1). For these genotypes, genetic diversity data obtained with four amplified fragment length polymorphism (AFLP) primer combinations (Papa and Gepts, 2003) were also used.

### *AFLP analysis*

A total of 141 AFLP primer combinations involving 14 *Eco*RI and 48 *Mse*I primers were used, following the protocols described by Vos *et al.* (1995), with minor modifications. For restriction–ligation, total genomic DNA (250 ng) was restricted with *Eco*RI (5′-GAATTC-3′) and *Mse*I (5′-TTAA-3′) by incubation at 37 °C for 3 h. The digested products were ligated to *Eco*RI (5′-CTCGTAGACTGCGT ACC-3′ and 3′-CTGACGCATGGTTAA-5′) and *Mse*I (5′-GACGATGAGTCCTGAG-3′ and 3′-TACTCAGGACT CAT-5′) adaptors at 37 °C for 2·5 h. For pre-selective amplification, DNA from the restriction–ligation reactions was diluted 8-fold prior to amplification with the primers (*Eco*RI, 5′-GACTGCGTACCAATTC-3′; *Mse*I, 5′-GATG AGTCCTGAGTAA-3′) with a single selective nucleotide (i.e. *Eco*RI + N and *Mse*I + N). For selective amplification, the pre-selective amplification products were diluted 1:11 and used as templates for the second amplification. Cy5-labelled *Eco*RI primers were used for the last selective PCR, and the AFLP fragments were separated by 6 % polyacrylamide gel electrophoresis for 2 h at 50 W constant power, using a Genomix system (Beckman, CA, USA).

Data were scored visually for all of the gels and recorded as the presence or absence of the same co-migrating AFLP fragment.

### *Pooling strategy*

To analyse a large number of markers, screening was based on pooled DNA samples (bulk DNA or bulks). Using AFLP, the DNA of 14 bulks of genotypes representing wild and domesticated subpopulations of *P. vulgaris*

TABLE 1. *Wild (W) and domesticated (D) accessions of* P. vulgaris *from Mexico used for AFLP analysis (see also Papa and Gepts, 2003)*

| Accession no. | Accession name | Bulk | Form | State | Altitude (m a.s.l.) | Seed colour |
|---|---|---|---|---|---|---|
| 122 | CHWI6 | WCH1 | W | Chiapas | 1200 | Agouti |
| 128 | CHWI12 | WCH1 | W | Chiapas | 1200 | Agouti |
| 282 | CHWETE2 | WCH1 | W | Chiapas | 1200–1050 | Agouti |
| 284 | CHWETE10 | WCH2 | W | Chiapas | 1200–1050 | Black |
| 290 | CHWETE31 | WCH2 | W | Chiapas | 1200–1050 | White |
| 292 | CHWETE35 | WCH2 | W | Chiapas | 1200–1050 | Agouti |
| 300 | CHWERD7 | WCH3 | W | Chiapas | 1200–1050 | Agouti |
| 309 | CHWES2 | WCH3 | W | Chiapas | 1200–1050 | Agouti |
| 315 | CHWES8 | WCH3 | W | Chiapas | 1200–1050 | Agouti |
| 145 | JADW2 | WJA1 | W | Jalisco | 1740 | Brown |
| 148 | JADW5 | WJA1 | W | Jalisco | 1740 | Agouti |
| 152 | JADW11 | WJA1 | W | Jalisco | 1740 | Brown |
| 155 | JADW56 | WJA2 | W | Jalisco | 1740 | White |
| 156 | JADW45 | WJA2 | W | Jalisco | 1740 | Agouti |
| 255 | OXW5 | WOX1 | W | Oaxaca | Not available | Agouti/brown |
| 259 | OXW9 | WOX1 | W | Oaxaca | Not available | White/brown |
| 262 | OXW12 | WOX1 | W | Oaxaca | Not available | Agouti/violet |
| 183 | PUW7 | WPU1 | W | Puebla | 1280 | Agouti |
| 190 | PUW20 | WPU1 | W | Puebla | 1280 | Agouti |
| 71 | CHCEE3 | DCH1 | D | Chiapas | 1200–1050 | Black |
| 82 | CHCEE34 | DCH1 | D | Chiapas | 1200–1050 | Pink |
| 113 | CHCI30 | DCH1 | D | Chiapas | 1200 | Black |
| 116 | CHCI44 | DCH2 | D | Chiapas | 1200 | Black |
| 332 | CHCES10 | DCH2 | D | Chiapas | 1200–1050 | Pink |
| 337 | CHCES30 | DCH2 | D | Chiapas | 1200–1050 | Pink |
| 93 | CHCF12 | DCH3 | D | Chiapas | 1650 | 'Ojo de cabra' |
| 96 | CHCF16 | DCH3 | D | Chiapas | 1650 | 'Ojo de cabra' |
| 100 | CHCD13 | DCH3 | D | Chiapas | 1450 | Pink |
| 170 | JADC5(47) | DJA1 | D | Jalisco | 1740 | 'Flor de castilla' |
| 171 | JADC2 | DJA1 | D | Jalisco | 1740 | Yellow |
| 172 | JADC3(45) | DJA1 | D | Jalisco | 1740 | White |
| 174 | JADC1 | DJA2 | D | Jalisco | 1740 | Yellow |
| 177 | JADC8(11) | DJA2 | D | Jalisco | 1740 | White |
| 251 | OXC1 | DOX1 | D | Oaxaca | Not available | Black |
| 261 | OXC11 | DOX1 | D | Oaxaca | Not available | Black |
| 268 | OXC18 | DOX1 | D | Oaxaca | Not available | Black |
| 191 | PUC21 | DPU1 | D | Puebla | 1280 | Brown |
| 192 | PUC22 | DPU1 | D | Puebla | 1280 | Brown |

from Mexico (Mesoamerica) were analysed. Each bulk represents a different subpopulation and was obtained by pooling the first AFLP amplification products (Vos *et al.*, 1995) of two or three different individuals. For the analysis of the bulks, 2 μL of the selective amplification from each individual genotype were pooled, as preliminary analysis showed that the pooling of pre-selection amplifications was not affected by differences in the DNA concentrations of the samples, compared with pooling genomic DNA or restriction ligations (data not shown). Preliminary observations also showed that in the bulk analysis the occurrence of a given AFLP fragment present in only one of the pooled genotypes was not always visible when more than three genotypes were pooled together (data not shown). The choice was therefore made to pool no more than three sgenotypes. Thus, when a given AFLP fragment was absent in a given bulk, it was considered to be absent in all three of the genotypes making up the same bulk, while when it was present, the band could arise from one or more individuals of the bulk. No attempt was made to relate the intensity of the AFLP band to its frequency in the bulk after preliminary testing.

The genotypes were pooled in 14 bulks (seven wild and seven domesticated) from the same populations (Table 1). The bulks were constructed to maximize the similarities between individuals within each bulk and to minimize differences in the average similarities between wild and domesticated samples. Indeed, by using AFLP data from four primer combinations obtained from individual genotypes (Papa and Gepts, 2003), the similarities within bulks were 85·0 % for the wild and 87·8 % for the domesticated samples (data not shown). This compares with 67 and 76 %, respectively, from the same data, for within-bulks similarities if the bulks had been built by random sampling of the genotypes. Thus, compared with random sampling of genotypes for bulk construction, this approach increased the probability of sampling the same allele from individuals of the same bulk and made homogeneous the same probability between the wild and domesticated forms.

### *Individual genotype analysis*

After the screening of the bulks, all of the 38 genotypes were analysed individually, using 10 AFLP primer

combinations that showed diagnostic fragments. Thus, along with the information from Papa and Gepts (2003), the AFLP analysis on individual genotypes was based on 14 primer combinations, which produced a total of 250 polymorphic fragments.

*Allelic frequencies and population parameters inferred from the bulk frequencies, and their validation*

To exploit better the information contained in the data set, the frequencies of the AFLP markers in the two samples of wild and domesticated genotypes were estimated from all the eight combinations of presence vs. absence observed in seven bulks (from zero to seven bulks with the presence of the band). Allelic frequencies were estimated assuming complete homozygosity, because *P. vulgaris* has a predominantly autogamous behaviour (>98 % autogamy in most studies, e.g. Ibarra-Pérez *et al.*, 1997) and a strong population structure (e.g. high $F_{ST}$; Papa and Gepts, 2003), which reduce the heterozygosity in a sub-population through random genetic drift. Because the distribution of allelic frequencies is a characteristic of the markers and populations used (Crow and Kimura, 1970), approaches based on the expectation of their uniform distribution could be misleading. Indeed, real frequencies were first inferred using maximum likelihood approaches, but, as expected, a good fit with the observed data was not obtained because frequencies were skewed towards extreme values (data not shown). Thus, the observed frequencies were estimated by using the data from individual genotypes for 250 AFLPs obtained with the 14 primer combinations. AFLP markers were grouped by their eight possible combinations of presence or absence of the AFLP fragment (from zero to seven) in the wild and domesticated samples. For each combination, the average frequencies obtained for the same markers were then calculated for all of the 19 wild and 19 domesticated genotypes analysed individually. This approach was first conducted separately for the wild and domesticated samples, but eventually an average estimation of the AFLP frequencies of each bulk combination for both samples was made because the estimated frequencies were almost equal for every bulk combination in the two samples. Similarly, the data from Papa and Gepts (2003) and those originating from the analysis of the AFLP primer combinations that presented diagnostic markers between wild and domesticated forms were also evaluated separately. Because of the very high correlations between the two estimates, the data were used conservatively based on the individual genotypes of both wild and domesticated samples to estimate the average frequencies of the eight different bulk frequencies (from zero to seven).

The approach was validated with 50 re-samplings. Each time, half of the 250 markers were randomly selected, making two equally numerous but independent groups of markers. At each re-sampling cycle, the first group was used to calculate the average AFLP frequencies for all the eight combinations of bulks (from zero to seven). Then, the observed frequencies, computed on individuals from the second sample, were used to compare this with the estimated frequencies obtained using the average values of the first group of markers.

*Genetic structure and outlier detection*

Data were analysed assuming a haploid genome (i.e. complete homozygosity) because of the predominantly selfing mating system of *P. vulgaris*. $F_{ST}$ (Weir and Cockerham, 1984) was calculated as described by Weir (1996) for haploid populations and genetic diversity was calculated as unbiased heterozygosity (Nei, 1978). An *ad hoc* statistic ($\Delta H$) was also used to measure the loss of diversity in domesticated populations as proposed by Vigouroux *et al.* (2002): $\Delta H = 1 - (H_D/H_W)$, where $H_D$ and $H_W$ are the genetic diversity in domesticated and wild common bean, respectively. To detect the effects of selection, the approach used was that proposed by Beaumont and Nichols (1996), further developed by Beaumont and Balding (2004), and implemented in the FDIST2 software (http://www.rubic.rdg.ac.uk/~mab/software.html). This approach is based on the infinite island model, but it has been shown to be quite robust in different demographic scenarios (Beaumont and Balding, 2004). The method is usually based on the observed weighted $F_{ST}$ that originates from all of the markers. By coalescence simulation, the expected neutral distribution of $F_{ST}$ conditioned on heterozygosity is then obtained. A new $F_{ST}$ is then calculated by excluding all of the loci showing departure from the simulated expected neutral distribution. To determine the putative neutral $F_{ST}$, this process is iterated until no further locus falls outside the expected distribution. In the present case, considering the high number of markers, in each test of neutrality, a significance $P$ value of 0·01 was used to avoid the risk of underestimating the 'real' $F_{ST}$ value and, thus, an overestimation of the effects of selection. On the other hand, this may increase the risk of overestimating the number of loci that undergo homogeneous selection between populations which will show a lower $F_{ST}$ than expected. Loci subjected to homogeneous selection between populations have also been observed to be hardly detectable with this method (Beaumont and Balding, 2004). However, interest here was focused on the loci that are differentially selected between the wild and domesticated forms (showing a higher $F_{ST}$ than expected). Therefore, even if markers showing an $F_{ST}$ lower than that expected were eliminated during the iterative process, they were not taken into account as putative outliers, but they were considered as putative neutral loci (PN). Comparisons between the expected $F_{ST}$ neutral distribution and the observed $F_{ST}$ estimates allowed the detection of outliers that were potentially under heterogeneous selection between wild and domesticated populations. The same expected neutral distribution as compared with that of the $F_{ST}$ estimates from the bulks was then used to detect the frequency of outlier loci.

The data set from the individual genotypes was first used to determine the expected neutral value of $F_{ST}$, using a $P$ value of 0·01, and to detect the outlier loci from the same data set. Then, the same $F_{ST}$ was used to estimate the departure from neutral expectation for all of the other markers for

which inferred frequencies were used for the bulk analysis. Thus, for bulks, the iterative procedures of FDIST2 was not used, but instead the putative neutral $F_{ST}$ derived from the analysis of individual genotype data was used. The data from the individual genotypes were then divided into three groups: putatively neutral (PN; $P > 0{\cdot}05$, also including markers with a significantly lower $F_{ST}$ compared with the null distribution) and putatively under selection at $P < 0{\cdot}05$ ($PS_{0{\cdot}05}$) and at $P < 0{\cdot}01$ ($PS_{0{\cdot}01}$).

### *AFLP mapping*

A critical point of all methods that try to detect a signature of selection using both diversity and divergence methods is the risk of false positives. A validation of the approach is therefore usually recommended (Storz, 2005). For this purpose, the map localization of the loci was compared with previously known genes or QTLs for domestication. In the present case, many genes and QTLs relating to the genetic control of the set of traits selected during the domestication process have been localized in the common bean core map (Koinange *et al.*, 1996; Freyre *et al.*, 1998). Thus it can be expected that many of the loci under selection could be linked to those genes and QTLs, as was also seen by Papa *et al.* (2005). To map the AFLP fragments obtained from individual genotype analyses, a set of 61 recombinant inbred lines (RILs) mapping the BAT93 × Jalo EEP558 (BJ) population was used to locate those markers in the core map of *P. vulgaris* that was developed by Freyre *et al.* (1998). The map location was calculated using the framework markers (mainly RFLPs) used to develop the core map (Freyre *et al.*, 1998) employing the Mapmaker/EXP version 3·0 software, following standard procedures (Freyre *et al.*, 1998; Papa *et al.*, 2005). The core map also included markers used by Koinange *et al.* (1996) to map the gene and QTLs involved in the process of domestication in the common bean in the RILs population derived by crossing a domesticated Andean (MIDAS) and a wild Mesoamerican genotype (G12873).

### *Phylogeny*

For each data set, 1000 bootstrapped simple matching distances were generated to construct a neighbour-joining consensus tree condensed to 50 % of the bootstrap support using the Mega 3·1 software (Kumar *et al.*, 2004; http://www.megasoftware.net/).

## RESULTS

### *Bulk screening*

A total of 3619 fragments were identified using 141 primer combinations in the 14 DNA bulks analysed. An average of 26 fragments per primer combination was seen, with 3312 fragments in the wild and 3048 in the domesticated bulks. The number of polymorphic markers was 2509 (69 %), with an average of 18 polymorphic markers per primer combination. Considering that the size of the bean genome is about 600 Mb (Arumuganathan and Earle, 1991; Bennet *et al.*, 2000; Brougthon *et al.*, 2003), a genome-wide average density of about one polymorphic AFLP every 250 kb was obtained. The percentage of polymorphic markers was 75 % in the wild and 62 % in the domesticated bulks. Among the polymorphic markers, 878 (34 %) were private: 571 were present only in the wild and 307 only in the domesticated bulks. Finally, among the private markers, 58 were diagnostic (private and fixed in one of the samples), of which 27 diagnostic fragments were present only in the wild and 31 only in the domesticated samples.

The frequency of diagnostic fragments observed in the sample was tested using all possible permutations by randomly re-labelling each bulk sample as domesticated or wild. All of the combinations of bulks in the two groups of seven have been tested for the presence vs. absence of a diagnostic fragment, to verify the probability of obtaining the observed values just by chance. The permutations showed that 58 or a comparable number of diagnostic fragments could only be detected when the real wild and domesticated samples were contrasted. In all other cases, the number of diagnostic fragments ranged from zero to a maximum of nine. For the overall permutations, a mean frequency of 0·15 diagnostic fragments was seen, with 98 % of the permutations showing a maximum value of one diagnostic fragment.

### *Analysis based on individual genotypes*

The AFLP data from 14 primer combinations were first analysed. This set consisted of 250 polymorphic AFLPs, about 10 % of all of the markers analysed using the pooled DNA samples. An FDIST2 analysis was started from the mean estimated $F_{ST}$ (0·41), using all of the markers. Thus $F_{ST} = 0{\cdot}41$ was used as the input for the first step of the iterative process of FDIST2, using 100 000 simulations for each step. After three iterative processes that excluded at each step all of the markers departing from a neutral expectation at $P < 0{\cdot}01$, no more outliers at $F_{ST} = 0{\cdot}14$ were found. A final run was then performed using all the 250 markers assuming $F_{ST} = 0{\cdot}14$ and using 1 000 s000 simulations. Three subsets of data were constructed: the first with 155 PN AFLPs (with 145 markers falling within the expected distribution at $P < 0{\cdot}05$ and 10 with a significantly lower $F_{ST}$ than expected) and the second and the third with markers putatively under selection at $P < 0{\cdot}05$ ($PS_{0{\cdot}05}$) and $P < 0{\cdot}01$ ($PS_{0{\cdot}01}$), consisting of 95 and 76 AFLPs, respectively.

### *Diversity*

In Fig. 1, genetic diversity ($H$) and the reduction in diversity between wild and domesticated ($\Delta H$) are shown for the three subsets of markers. The genetic diversity is significantly lower ($P < 0{\cdot}01$, Wilcoxson, non-parametric test) in both wild and domesticated types for $PS_{0{\cdot}05}$ and $PS_{0{\cdot}01}$ markers compared with PN markers. However, this pattern appears to be more drastic in the domesticated

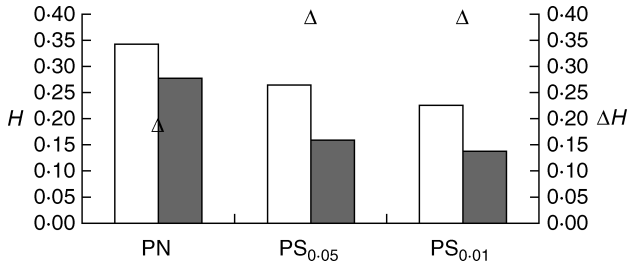


FIG. 1. Individual genotype analysis: genetic diversity ($H$) in wild (open bars) and domesticated (shaded bars) populations using putatively neutral AFLPs (PN) and those putatively under selection $P < 0{\cdot}05$ ($PS_{0{\cdot}05}$) and $P < 0{\cdot}01$ ($PS_{0{\cdot}01}$). The reductions in genetic diversity ($\Delta H$) are also shown (open triangles).

types, as the reduction in diversity observed is almost doubled for $PS_{0{\cdot}01}$ and $PS_{0{\cdot}05}$ compared with PN markers.

*Phylogenetic analysis*

In Fig. 2, the neighbour-joining trees obtained using the PN (Fig. 2A) and $PS_{0{\cdot}01}$ (Fig. 2B) AFLPs are shown. For the $PS_{0{\cdot}01}$ markers, a clear and highly significant separation of wild and domesticated genotypes into separate clades is seen, in contrast to the PN markers, where the wild and domesticated populations do not group into separate clusters. Nevertheless, for PN markers, the populations do tend to group in some cases by geographic origin, as can be seen easily for the Chiapas samples, which form a group of wild and domesticated forms separate from all the other accessions (Fig. 2A). The results using $PS_{0{\cdot}05}$ are not shown because they were almost identical to those of $PS_{0{\cdot}01}$.

*Mapping markers*

The loci putatively influenced by selection were validated by determining the proportion of putative outliers that mapped close to genes and QTLs of the domestication syndrome, as defined by Papa *et al.* (2005). Markers were located on the core map, and linkage distances between AFLP markers and QTLs were calculated to the adjacent markers with the highest LOD score and then classified as: D, linked to domestication loci (cM $< 30$); ND, linked to QTLs or genes not related to the domestication process (cM $< 30$); or UN, not linked to any identified genes or QTLs (cM $> 30$). A $\pm 30$ cM interval was used for the classification of AFLPs for the following reasons: (*a*) because the effects of selection may extend over large distances in the presence of linkage; (*b*) because the confidence interval of a QTL may be very large (Darvasi *et al.*, 1993); and (*c*) for consistency with Papa *et al.* (2005).

Overall, 70 AFLPs from the 250 used in individual genotypes could be mapped: 30 D, 10 ND and 30 UN. Among the mapped markers that are putatively under the effects of selection (29 $PS_{0{\cdot}05}$ and 23 $PS_{0{\cdot}01}$), most of them were D markers, linked to genes or QTLs related to the domestication syndrome (Fig. 3). The frequency of D markers was significantly higher for $PS_{0{\cdot}01}$ compared with PN. Indeed, while 65 % ($n = 15$) of the 23 $PS_{0{\cdot}01}$ were D markers, only 32 % ($n = 15$) of the PN markers were linked to QTLs related to the domestication syndrome

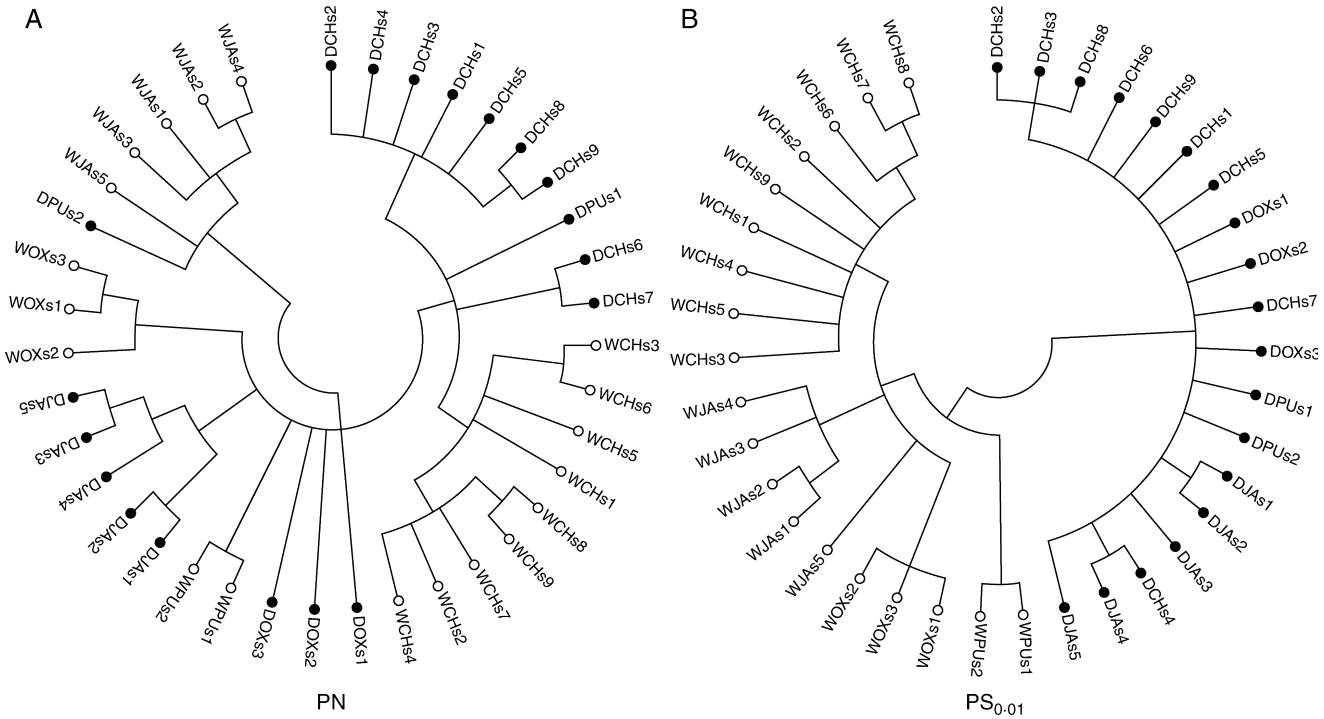


FIG. 2. Neighbour-joining consensus tree (condensed at $<50$ % bootstrap support) among individual genotypes using 1000 bootstrap re-samplings for PN and $PS_{0{\cdot}01}$ markers.

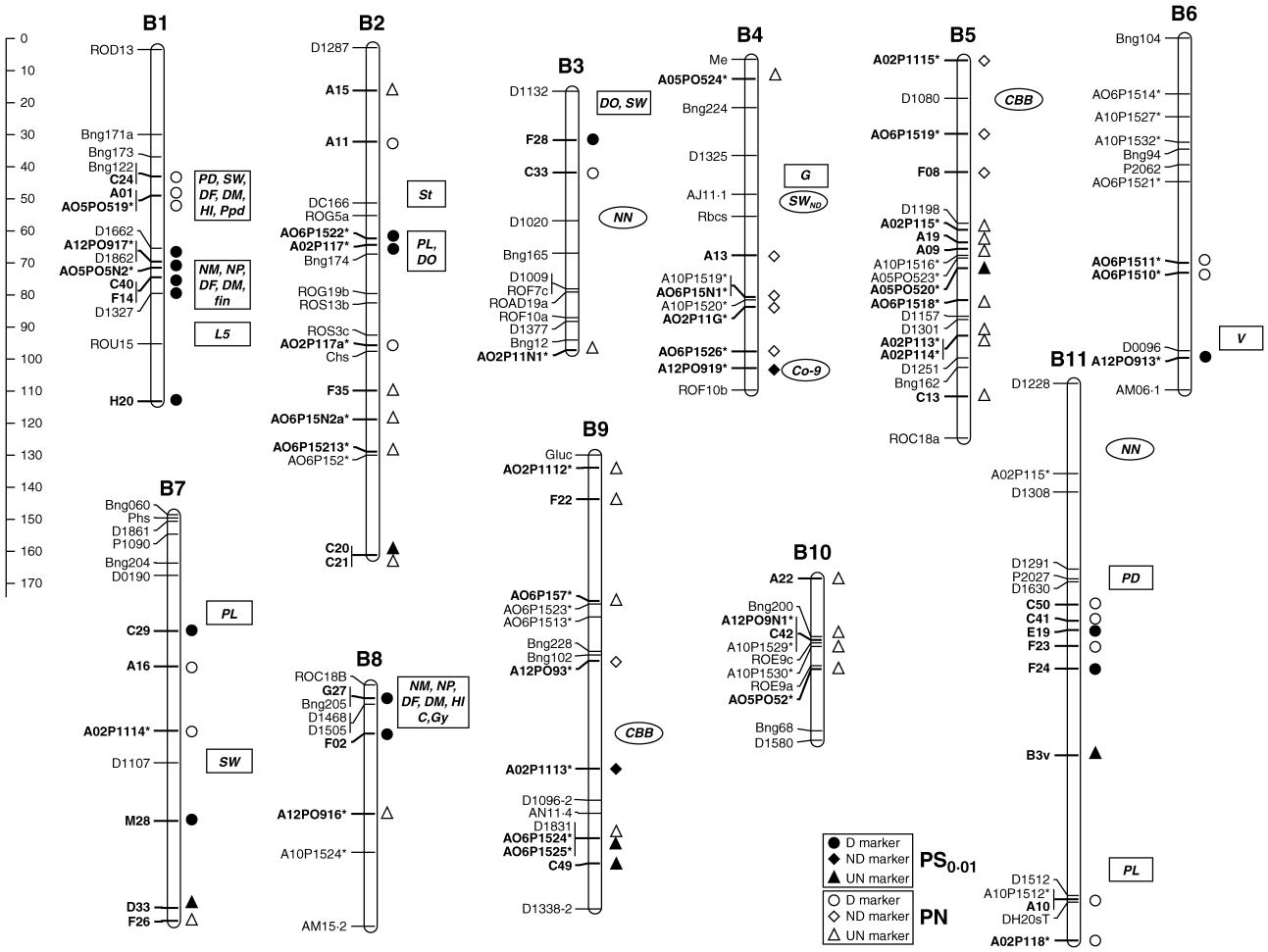


FIG. 3. Molecular linkage map of the common bean. To the right of each linkage group, the black symbols indicate the $PS_{0\cdot01}$ outliers detected with FDIST2 ($P < 0\cdot01$) and the grey symbols indicate the PN markers, mapped in the BAT93 × Jalo EEP558 population. AFLPs are classified as D (circle), ND (diamond) and UN (triangle). The mapped domestication genes and QTLs are in rectangles and other traits are in ovals. Genetic distances are in cM map units. QTLs or major genes: *CBB*, common bacterial blight resistance; *Co-9*, anthracnose resistance; DF, days to flowering; DM, days to maturity; DO, dormancy; *fin*, determinacy; *G*, *Gy*, *C*, seed colour; HI, harvest index; L5, length of the fifth internode; NM, number of nodes on the main stem; NN, rhizobium nodule number; NP, number of pods; PD, photoperiod sensitivity; PL, pod length; *Ppd*, photoperiod sensitivity; *St*, pod suture fibres; $SW_{DOM}$, seed weight, identified in a cross with wild bean; $SW_{ND}$, seed weight, identified in a cross between cultivars; *V*, flower and seed colour. To the left previously mapped framework markers (Freyre *et al.*, 1998), (*) AFLP markers mapped by Papa *et al.* (2005) and, in bold, AFLP markers polymorphic in this study are indicated.

(Pearson, $\chi^2 = 7\cdot0$, $P = 0\cdot008$). Also for $PS_{0\cdot05}$, the frequency of D markers (55 %; $n = 16$) was higher than for PN markers ($n = 14$; 34 %) but this difference was not statistically significant (Pearson $\chi^2 = 3\cdot1$, $P = 0\cdot08$).

Using a non-parametric test (Wilcoxon, rank sums), the average distances of $PS_{0\cdot01}$ and PN from QTLs or genes of the domestication syndrome were then compared (considering for each AFLP marker the closest domestication QTL or gene). This analysis was performed using (*a*) all markers (100 cM distance was used for all markers located in linkage groups where no domestication loci and QTLs were mapped: B5, B9 and B10); and (*b*) only markers located in chromosomes where genes and QTLs related to domestication have been detected. Considering all of the markers, the average minimum distance from QTLs of the domestication syndrome was significantly ($P < 0\cdot009$) higher for PN (58·4 cM) compared with $PS_{0\cdot01}$ (34·4 cM). A greater distance was also found for PN markers (30·2 cM) compared with $PS_{0\cdot01}$ (18·8 cM) ($P < 0\cdot02$) when only the markers located in the chromosome where domestication loci have been mapped were used. For $PS_{0\cdot05}$, the differences from PN were not significant (data not shown).

Finally, Fig. 4 shows the reduction of the frequency of $PS_{0\cdot01}$ at increasing distances between the markers and the domestication QTLs by comparing the frequency of $PS_{0\cdot01}$ in different groups of markers that were classified according to their distances to domestication QTLs. Five distance classes were considered (Fig. 4): $\leq$5 cM (average distance 3·1 cM, nine markers), $>$5 and $\leq$15 cM (average distance 10·0 cM, 11 markers), $>$15 and $\leq$30 cM (average distance 22·6 cM, 11 markers), $>$30 cM (average distance 53·0 cM, 15 markers) and markers located in linkage groups where no domestication QTLs

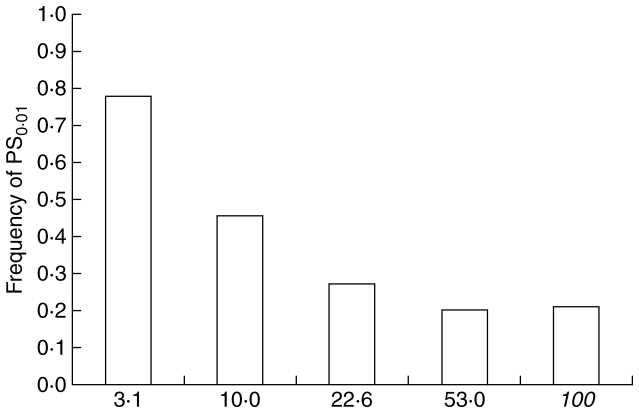


FIG. 4. Reduction in the frequency of $PS_{0\cdot01}$ at increasing distances between the markers and the domestication QTLs. Five distance classes were considered: $\leq$5 cM (average distance 3·1 cM, class 3·1), $>$5 and $\leq$15 cM (average distance 10·0 cM, class 10·0), $>$15 and $\leq$30 cM (average distance 22·6 cM, class 22·6), $>$30 cM (average distance 53·0 cM, class 53·0), and markers located in linkage groups where no domestication QTL or gene have been detected (class *100*).

or genes have been detected (class *100*, 24 markers). Overall, the different frequencies of $PS_{0\cdot01}$ seen among the distance classes (Fig. 4) were found to be significant (Pearson, $\chi^2 = 11{\cdot}9$; $P = 0{\cdot}018$; contingency table: $5 \times 2$, five distance classes and two classes of markers, $PS_{0\cdot01}$ and PN). For the class of markers showing the smallest average distance from domestication loci (3·1 cM), the frequency of $PS_{0\cdot01}$ was 0·78, significantly higher (with the only exception of class 10·0 cM) than all of the other classes, which ranged from 0·27 to 0·20 (class 22·6 cM, $P < 0{\cdot}025$; class 53 cM, $P < 0{\cdot}005$; and class 100 cM, $P < 0{\cdot}0025$). Class 10·0 cM presented an intermediate value (0·45 %) that was not significantly different from all of the other classes. For $PS_{0\cdot05}$, the overall test was not significant, and only the comparisons between class 3·1 cM, class 53 cM and class *100* were significant (data not shown).

### *Estimations of allelic frequencies and population parameters*

A total of 250 AFLP markers that were screened both in the bulks and in the individual genotypes were used to estimate the average frequencies for the eight combinations of presence vs. absence among the seven bulks. Then, using observed and estimated frequencies, the heterozygosity and the $F_{ST}$ was calculated for each marker. Figure 5 shows the agreement between the frequency calculated among individuals and those calculated in the bulks, whereas Fig. 6 shows the agreement for the $F_{ST}$ values. The agreement between the bulk and individual frequency estimates shows a substantial absence of bias: the slope of regression between the two series of values was 1·0 and the intercept was 0·0, with an $R^2$ of 0·93. For $F_{ST}$ and $H$, the agreement was also high: for $F_{ST}$, the slope was 0·91 and the intercept 0·06, with $R^2 = 0{\cdot}80$; for $H$, the slope was 0·972 and the intercept was −0·008, with an $R^2 = 0{\cdot}79$. This approach of estimating allelic frequencies was validated by 50 re-samplings of two sets of

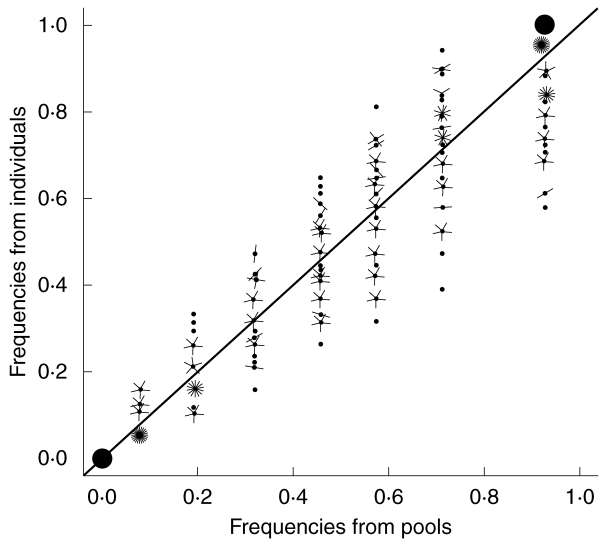


FIG. 5. Relationship between frequencies estimated from bulks and observed frequencies from individuals. The number of petals in the 'sunflower dots' is the number of overlapping points. The dashed line represents perfect correspondence, while the solid line represents the actual regression line. In this graph, the two lines overlap (see text).

independent markers (see Materials and Methods). The agreement between estimates from bulk and individuals was high: the slope of regression between the two series of values was 0·99 and the intercept was 0·00, with an $R^2$ of 0·93.

### *Analysis based on inferred frequencies from the bulk data*

The diversity ($H$) and divergence ($F_{ST}$) between wild and domesticated populations were estimated using the inferred frequencies from the 2506 AFLPs evaluated

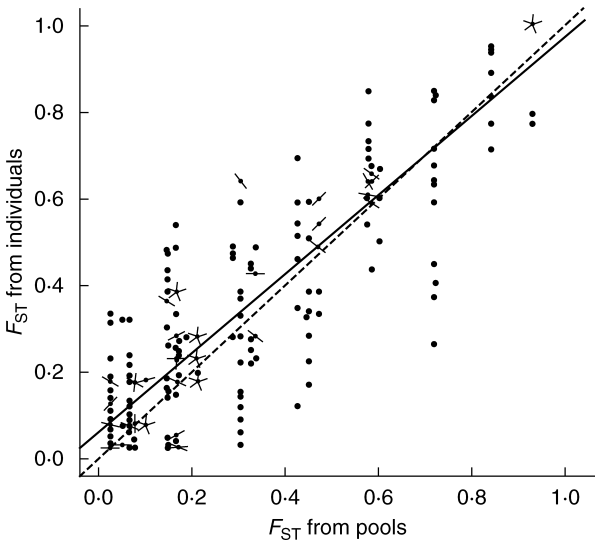


FIG. 6. Relationship between $F_{ST}$ estimated from bulks and real $F_{ST}$ estimated from individuals. The number of petals in the 'sunflower dots' is the number of overlapping points. The dashed lines represent perfect correspondence, while the solid lines represent the actual regression line (see text).

TABLE 2. *Diversity ($H_W$ wild; $H_D$ domesticated) and divergence (average weighted $F_{ST}$) estimated by putatively neutral AFLPs (PN) and putatively under selection for $P < 0.05$ ($PS_{0.05}$) and $P < 0.01$ ($PS_{0.01}$).*

| | No. | $F_{ST}$ | $H_W$ | $H_D$ | $\Delta H$ |
|---|---|---|---|---|---|
| All | 2,506 | 0·29 | 0·30 | 0·25 | 0·16 |
| PN | 2,036 | 0·12 | 0·31 | 0·27 | 0·14 |
| $PS_{0.05}$ | 470 | 0·65 | 0·28 | 0·20 | 0·29 |
| $PS_{0.01}$ | 404 | 0·68 | 0·26 | 0·18 | 0·32 |

using pooled DNA. Using all of the markers, the weighted average for $F_{ST}$ was 0·29, and for $H$, 0·30 for the wild, and 0·25 for the domesticated sample, with a resulting loss of diversity ($\Delta H$) of 0·16 (Table 2). Using FDIST2 (assuming a neutral $F_{ST}$ of 0·14 and $P < 0.05$), the number of markers showing a higher $F_{ST}$ compared with the neutral expectation was 470, corresponding to 19 % of all of the markers, and 404 with $P < 0.01$, corresponding to 16 % of all of the markers. Considering that the genome of the common bean is about 600 Mb, and assuming a uniform distribution of the markers over the genome, these results would correspond to average values of 114 Mb ($P < 0.05$), 97 Mb ($P < 0.01$) and 43 Mb (Bonferroni, $P < 0.05$).

In Fig. 7, the neutral distribution of $F_{ST}$ originating from 1 000 000 simulations using FDIST2 was compared with the observed distribution using inferred frequencies from DNA pools. In our materials, an excess of markers was observed that have a high $F_{ST}$ compared with the expected neutral distribution. The maximum value of the ratio of observed over expected neutral $F_{ST}$ was 7500 for the class $F_{ST} > 0.9$.

Using the $P < 0.05$ and the $P < 0.01$ thresholds, the data set was then split into three groups of markers: one of PN markers and two putatively under selection ($PS_{0.05}$ and $PS_{0.01}$). The results obtained with inferred frequencies from the bulks were then compared with the individual genotype data for the 250 markers for which both sets of information were available. Using the individual genotype data, 155 PN, 95 $PS_{0.05}$ and 76 $PS_{0.01}$ markers were obtained, while using the frequencies inferred from the bulks, 163 PN, 87 $PS_{0.05}$ and 66 $PS_{0.01}$ markers were obtained. Thus using the inferred frequencies, the numbers of markers subject to the effect of selection were underestimated, from 8·4 % for $PS_{0.05}$ to 13·2 % for $PS_{0.01}$. This indicates that with the inferred frequencies from the bulks, detection of the effects of selection was more conservative.

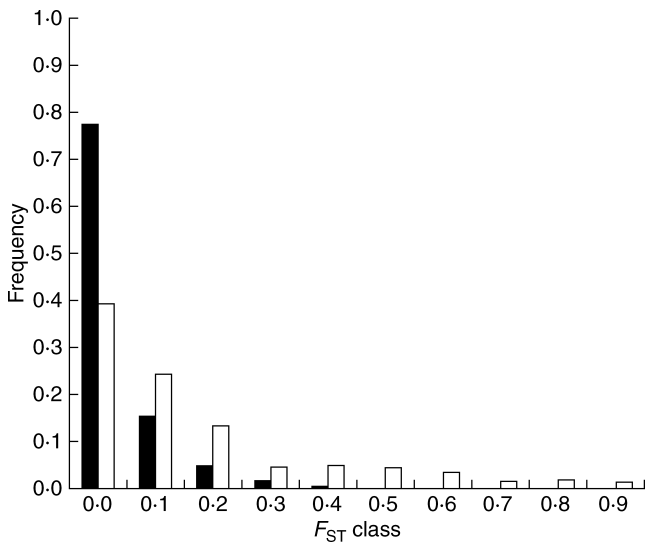


FIG. 7. Expected neutral distribution of $F_{ST}$ after 1 000 000 simulations assuming an average $F_{ST} = 0.14$ (filled bars) and an $F_{ST}$ derived from AFLP frequencies inferred from bulk data (open bars).

Table 2 shows the $H$, $\Delta H$ and $F_{ST}$ estimates for the PN and PS markers. The average weighted $F_{ST}$ for the $PS_{0.05}$ markers was 0·65, for the $PS_{0.01}$, 0·68, while for the PN markers it was 0·12. The wild samples showed a significantly higher level of diversity for all types of markers and for both the wild and domesticated types, $H$ was lower for the PS compared with the PN markers; however, the reduction in diversity was higher for the PS ($PS_{0.05}$ $\Delta H = 0.29$; $PS_{0.01}$ $\Delta H = 0.32$) than for the PN ($\Delta H = 0.14$) markers.

Figure 8 shows the neighbour-joining trees obtained using the PN (Fig. 8A) and $PS_{0.01}$ (Fig. 8B) AFLPs. As shown for the individual genotypes, two distinct groups were seen for the $PS_{0.01}$ markers that correspond to the wild and domesticated forms. For the PN markers, the wild and domesticated bulks did not separate into distinct clusters. The results using the $PS_{0.05}$ are not shown because they were almost identical to those of the $PS_{0.01}$.

## DISCUSSION

### *Statistical approaches*

The statistical approaches and the assumptions used in the present study raise the following questions.

(*a*) Was the method used to infer allelic frequencies from pooled DNA robust enough to support the inference on observed allelic frequencies? An empirical approach was used to estimate the allelic frequencies from the bulk data, based on the average frequencies of the eight possible combinations of presence and absence of AFLP fragments, analysed in individual genotypes and obtained from 250 AFLPs (about 10 % of the number of markers used for the bulk analysis). The data, shown in Fig. 5, and the validation with re-sampling of an independent set of markers, indicate that the frequencies were highly correlated between the observed and estimated results, even if the tendency for $F_{ST}$ was to underestimate slightly the level of divergence at high values (Fig. 6).

(*b*) Would the error rate made by inferring allelic frequencies using the bulks be small enough in combination with the method for the detection of outlier loci? The answer to this question can be obtained by a comparison between the number of markers identified as being putatively under selection based on the observed and inferred frequencies. When the results of the analysis conducted between individual genotypes and bulks (using the same set of 250 markers), both for $PS_{0.05}$ and $PS_{0.01}$, are compared, the inferred frequencies from the bulks underestimate the number of loci under selection by a similar amount (8 %

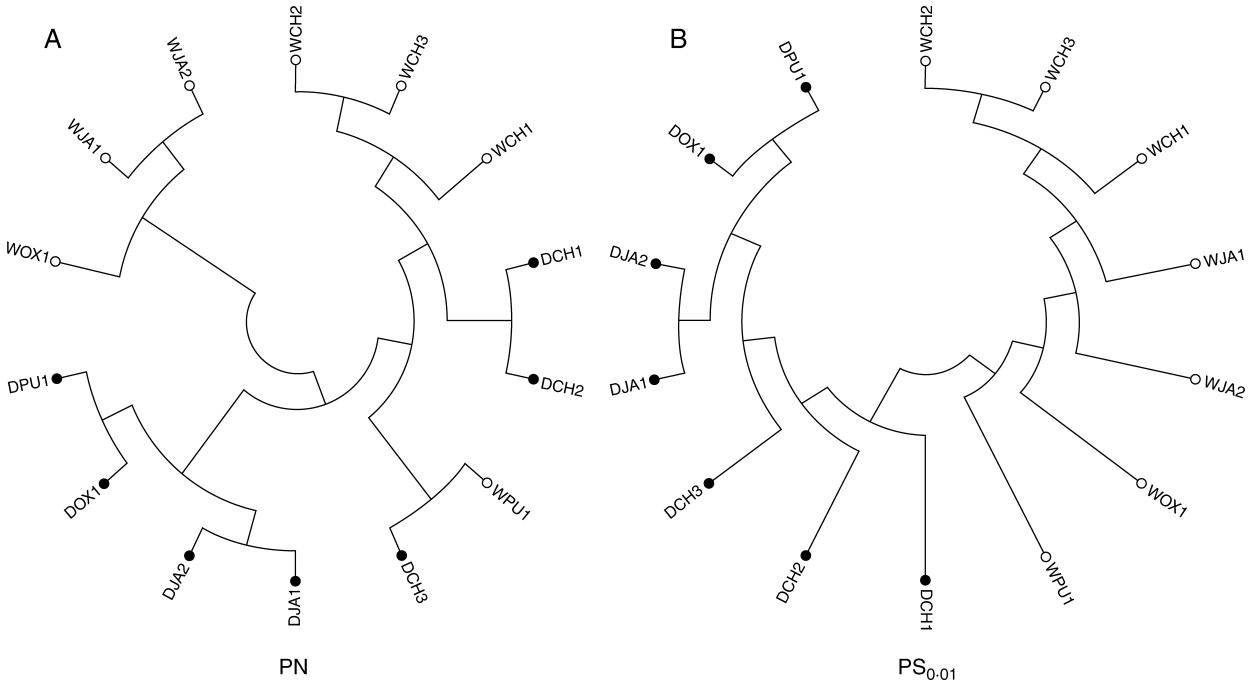


FIG. 8. Neighbour-joining consensus tree (condensed at <50 % bootstrap support) among bulks using 1000 bootstrap re-samplings for PN and $PS_{0·01}$ markers.

for $PS_{0·05}$ and 13 % for $PS_{0·01}$). This is probably due to the underestimation of $F_{ST}$ using inferred frequencies from the bulk data shown in Fig. 6. Thus, even in this case, the only apparent effect of using inferred frequencies from the bulks was an underestimation of the outliers. Using the estimated error (13 %) to correct the fraction of the $PS_{0·01}$ markers, 18 % $PS_{0·01}$ would have been obtained, instead of 16 %.

(*c*) Was the population divergence method used to detect the loci under selection appropriate for the present system? The population divergence methods implemented in FDIST2 have been widely used in several reports (Flint *et al.*, 1999; Vasemägi *et al.*, 2005; Bonin *et al.*, 2006; Hoffman *et al.*, 2006; Mealor and Hild, 2006; Pariset *et al.*, 2006) and have been shown to be quite robust by simulations among various demographic scenarios (Beaumont and Balding, 2004). However, as also indicated by Storz (2005), the risk of detecting false positives can be large because bottlenecks can produce effects similar to selection. Indeed, domestication induced a strong bottleneck in the domesticated gene pool (Gepts, 2004) of common bean (Gepts *et al.*, 1986; Sonnante *et al.*, 1994). For this reason, a stringent approach to the detection of putatively neutral $F_{ST}$ was adopted by eliminating only the markers significantly departing from neutral expectation at $P < 0·01$ at every iterative process implemented in FDIST2. Moreover, as has been suggested (Storz, 2005), additional evidence that the outliers detected were indeed loci that may have been affected by selection was sought. The results were validated by comparing the map location of the outlier AFLPs with the map location of the major QTLs and genes that affect the traits of the domestication syndrome (Koinange *et al.*, 1996). Indeed, the frequency of markers putatively under selection $PS_{0·01}$ was significantly higher for loci linked to genes and QTLs involved in the domestication process, compared with the other parts of the genome, and $PS_{0·01}$ showed a significantly smaller distance from domestication QTLs compared with PN. Moreover, the $PS_{0·01}$ markers that were classified as not involved in the domestication process were actually all located in areas of the genome where additional QTLs related to domestication were most recently identified (number of pods and seeds per plant, seed weight, yield and days to flowering) (Blair *et al.*, 2006). Even if a similar trend was seen for $PS_{0·05}$, the test for association between $PS_{0·05}$ and domestication loci was not found to be statistically significant. These results, and the fact that all but one of the D outliers were seen to be $PS_{0·01}$, suggest that a threshold of $P < 0·01$ would be more appropriate to avoid the detection of false positives, at least in the present case.

Additional indirect evidence can be obtained from the phylogenetic analysis. Indeed, this analysis for both the individual genotypes and the bulks showed that using the PN markers, the wild and domesticated genotypes from the Chiapas group fell into a single clade, compared with the genotypes from the other wild and domesticated genotypes from other Mexican states. However, when $PS_{0·01}$ (or $PS_{0·05}$) markers were used, the domesticated genotypes separated with high bootstrap support from all of the wild genotypes regardless of their geographic origins. As was seen in the common bean by Papa and Gepts (2003) and Papa *et al.* (2005), for nearly neutral loci unlinked to selected genes, introgression can occur because it is not counteracted by the effects of selection. Indeed, if the selected loci were not being tagged, no such differences in reconstructing the phylogeny for the PS and PN markers would have been observed. From this perspective, it is of particular interest to see that for putatively neutral

loci, wild and domesticated tend to group together in agreement with their geographical origin. Indeed, this result is what would be expected for putative neutral loci if gene flow is acting in such populations. Overall, it can be considered that the present approach is robust and conservative in the detection of the effects of selection particularly when a 1 % significance threshold is applied.

*Size of the genome under the effect of selection*

The main finding of this study is that a large fraction of the genome of the common bean appears to be under the effect of selection during domestication. Using an $F_{ST}$ approach, it has been shown that assuming a random distribution of the markers analysed, at least 16 % ($P < 0{\cdot}01$) of the genome (approx. 100 Mbp) appears to have been affected by the domestication process. Based on the discussion above, the results appear to be more an underestimation rather than an overestimation of the effects of selection. However, it has been taken into account that the AFLP markers are probably not evenly distributed within the genome. This may constitute a bias in the estimation.

The size of the genome affected by selection (16–18 %) appears to be very large. This is likely due to the combined action of selection and recombination at neutral loci, including hitchhiking (Maynard Smith and Haigh, 1974; Kaplan *et al.*, 1989) and background selection (Charlesworth *et al.*, 1993), rather than direct selection at target loci. In rice, Lu *et al.* (2006) recently hypothesized a strong effect of hitchhiking by comparing the complete genome sequence of two domesticated subspecies. On the other hand, the studies that have been conducted in maize using simple sequence repeats (SSRs) mainly derived from genes (Vigouroux *et al.*, 2002, 2005) or SNPs on gene fragments (Wright *et al.*, 2005; Yamasaki *et al.*, 2005) have shown that from 2 to 5 % of the genes of maize appear to be subject to the effects of selection during domestication and subsequent plant breeding. One possible explanation for these differences could be related to the divergent reproductive systems, which affect the effective recombination rate in autogamous vs. allogamous species. The higher the effective recombination rate (without considering the effects of population structure and epistatic selection), the lower the decay of linkage disequilibrium (LD) over time, and thus the effects of selection at neutral linked loci (i.e. hitchhiking) may extend over large genomic intervals when the effective recombination rate is small. In maize, LD decays within a few kilobases (Remington *et al.*, 2001; Tenaillon, 2001; Palaisa *et al.*, 2003). In rice, a predominantly self-pollinated species, the size of the selective sweep around the *waxy* locus, a gene selected during domestication, is around 250 kb (Olsen *et al.*, 2006), and a similarly slow decay has been observed in the selfer *Arabidopsis* (Nordborg *et al.*, 2002), even if in some autogamous species such as barley the decay of LD is much faster (Lin *et al.*, 2002; Morrell *et al.*, 2005; Caldwell *et al.*, 2006).

However, these strong differences could also originate from or be enhanced by an underestimation of the effects of domestication in maize, as was suggested by Wright *et al.* (2005) due to the lack of polymorphism of gene fragments used in their analysis. Alternatively, many loci under the effects of selection are not located in the transcribed portion of the genes. Similarly, these differences could arise because the effects of selection have been overestimated in relation to the possible non-uniform distribution of the markers in the common bean genome, as they could be located in part in areas of reduced recombination, where the effects of hitchhiking would be enhanced.

These results have implications for the use of wild germplasm for plant breeding in the common bean, and probably in other autogamous species as well. According to the estimations of the fraction of the genome that will be subject to the effects of selection, the domestication process appears to have affected not only target genes, but also a large portion of the genome around these genes. The regions of the genome surrounding the major domestication genes appear to be particularly interesting to tag the introgression from wild relatives into modern cultivars. Because of the combined actions of selection and recombination, these 'domestication islands' have probably experienced a higher level of isolation between the wild and the domesticated forms in comparison with the rest of the genome. Indeed, farmers and breeders selecting for domesticated alleles have probably also selected against many other tightly linked genes. As shown recently in *P. vulgaris* (Papa *et al.*, 2005), the domestication regions of the genome appear to harbour much higher levels of genetic variation in the wild populations, in comparison with the domesticated populations. Hence, tagging the domestication loci would be useful in two ways: the identification of markers that are tightly linked to undesirable genes (e.g. shattering); and the possibility to tag the surrounding chromosomal regions that would be most likely to harbour the highest and historically less exploited diversity of the wild germplasm.

*Diversity*

Papa *et al.* (2005) showed a higher loss of diversity and divergence between wild and domesticated populations at loci involved in the process of domestication. Here, this result is confirmed by the reverse approach, which consists of identifying markers showing the signature of selection and then comparing those with putatively neutral markers. In contrast to Papa *et al.* (2005), some differences in the pattern of genetic diversity are seen across markers within wild and domesticated populations. In particular, a decrease in diversity for both wild and domesticated forms from PN to PS loci is seen, while Papa *et al.* (2005) showed an increase in diversity from UN to D loci in the wild and the absence of any significant difference among UN and D markers in the domesticated gene pool. This discrepancy could be due to the fact that loci closer to the target genes are indeed being tagged, compared with Papa *et al.* (2005). The lower diversity observed in the wild and domesticated populations for PS markers suggests the occurrence of hitchhiking or background selection at these loci. However, these differences are probably due to the different aims

and thus sampling methods of genotypes between the two studies.

### *QTL detection and BPA screening method*

The present results indicate that the use of population genomics is very useful to detect QTLs for adaptive variants which cannot be observed using conventional QTL mapping, because of a lack of polymorphism in the segregating populations or because the adaptive trait does not show an obvious phenotype. An efficient strategy to scan the genome for adaptive variants using pooled DNA has been shown here; a large number of markers could be efficiently pre-screened that could also be used to develop PCR-based markers, for use after validation in marker-assisted plant breeding. This approach, which resembles BSA and could be referred to as bulk population analysis (BPA), is relatively inexpensive and appears particularly suitable for species where limited genomic information is available from public databases, as compared with other crops.

## ACKNOWLEDGEMENTS

We thank Jacques David for his valuable suggestions. This research was supported by the Italian Government (MIUR) grant no. # 2005071310, Project PRIN 2005. Funding to pay the Open Access publication charges for this article was provided by the OECD.

## LITERATURE CITED

**Arnheim N, Strange C, Erlich H. 1985.** Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. *Proceedings of National Academy of Sciences, USA* **82**: 6970–6974.

**Arumuganathan K, Earle ED. 1991.** Nuclear DNA content of some important plant species. *Plant Molecular Biology* **9**: 208–218.

**Beaumont MA. 2005.** Adaptation and speciation: what can F(st) tellus? *Trends in Ecology and Evolution* **20**: 435–440.

**Beaumont MA, Balding DJ. 2004.** Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**: 969–980.

**Beaumont MA, Nichols RA. 1996.** Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B: Biological Sciences* **263**: 1619–1626.

**Bennett MD, Bhandol P, Leitch IJ. 2000.** Nuclear DNA amounts in Angiosperms and their modern uses – 807 new estimates. *Annals of Botany* **86**: 859–909.

**Blair MW, Iriarte G, Beebe S. 2006.** QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean × wild common bean (*Phaseolus vulgaris* L.) cross. *Theoretical and Applied Genetics* **112**: 1149–1163.

**Bonin A, Taberlet P, Miaud C, Pompanon F. 2006.** Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution* **23**: 773–783.

**Broughton WJ, Hernández G, Blair M, Beebe S, Gepts P, Vanderleyden J. 2003.** Beans (*Phaseolus* spp.) – model food legumes. *Plant and Soil* **252**: 55–128.

**Butcher LM, Meaburn E, Dale PS, Sham P, Schalkwyk LC, Craig IW, Plomin R. 2005.** Association analysis of mild mental impairment using DNA pooling to screen 432 brain-expressed single nucleotide polymorphisms. *Molecular Psychiatry* **10**: 384–392.

**Caldwell KS, Russell J, Langridge P, Powell W. 2006.** Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* **172**: 557–567.

**Cavalli-Sforza L. 1966.** Population structure and human evolution. *Proceedings of the Royal Society B: Biological Sciences* **164**: 362–379.

**Charlesworth B, Morgan MT, Charlesworth D. 1993.** The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.

**Crow JF, Kimura M. 1970.** *Introduction to population genetics theory*. New York: Harper & Row.

**Darvasi A, Weinreb A, Minke V, Weller JI, Seller M. 1993.** Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134**: 943–951.

**Flint J, Bond J, Rees DC, Boyce AJ, Roberts-Thomson JM, Excoffier L, *et al*. 1999.** Minisatellite mutational processes reduce $F_{ST}$ estimates. *Human Genetics* **105**: 567–576.

**Freyre R, Skroch PW, Geffroy V, Adam-Blondon A-F, Shirmohamadali A, Johnson WC, *et al*. 1998.** Towards an integrated linkage map of common bean. 4. Development of a core map and alignment of RFLP maps. *Theoretical and Applied Genetics* **97**: 847–856.

**Gepts P. 1998.** Origin and evolution of common bean: past events and recent trends. *Horticultural Science* **33**: 1124–1130.

**Gepts P. 2004.** Crop domestication as a long-term selection experiment. *Plant Breeding Reviews* **24**: 1–44.

**Gepts P, Papa R. 2002.** Evolution during domestication. In: *Encyclopedia of life sciences*. London: Nature Publishing Group, 1–7.

**Gepts P, Osborn TC, Rashka K, Bliss FA. 1986.** Phaseolin-protein variability in wild forms and landraces of the common bean (*Phaseolus vulgaris*): evidence for multiple centers of domestication. *Economy Botany* **40**: 451–468.

**Gur A, Zamir D. 2004.** Unused natural variation can lift yield barriers in plant breeding. *PLoS Biology* **2**: e245.

**Hinds DA, Seymour AB, Durham LK, Banerjee P, Ballinger DG, Milos PM, *et al*. 2004.** Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. *Human Genomics* **1**: 421–434.

**Hoffman EA, Schueler FW, Jones AG, Blouin MS. 2006.** An analysis of selection on a colour polymorphism in the northern leopard frog. *Molecular Ecology* **15**: 2627–2641.

**Ibarra-Pérez F, Ehdaie B, Waines G. 1997.** Estimation of outcrossing rate in common bean. *Crop Science* **37**: 60–65.

**Jarvis DI, Hodgkin T. 1999.** Wild relatives and crop cultivars: detecting natural introgression and farmer selection of new genetic combinations in agroecosystems. *Molecular Ecology* **256**: 159–173.

**Kaplan NL, Hudson RR, Langley CH. 1989.** The 'hitch-hiking effect' revisited. *Genetics* **123**: 887–899.

**Kelly JD. 2004.** Advances in common bean improvement: some case histories with broader applications. *Acta Horticulturae* **637**: 99–122.

**Koinange EMK, Singh SP, Gepts P. 1996.** Genetic control of the domestication syndrome in common bean. *Crop Science* **36**: 1037–1045.

**Kumar S, Tamura K, Nei M. 2004.** MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics* **5**: 150–163.

**Lewontin RC, Krakauer J. 1973.** Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. *Genetics* **74**: 175–195.

**Lin JZ, Morrell PL, Clegg MT. 2002.** The influence of linkage and inbreeding on patterns of nucleotide sequence diversity at duplicate alcohol dehydrogenase loci in wild barley (*Hordeum vulgare* ssp. *spontaneum*). *Genetics* **162**: 2007–2015.

**Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI. 2006.** The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics* **22**: 126–131.

**Maynard Smith J, Haigh J. 1974.** The hitch-hiking effect of a favourable gene. *Genetical Research* **23**: 23–35.

**McCouch S. 2004.** Diversifying selection in plant breeding. *PLoS Biology* **2**: e347.

**Mealor BA, Hild AL. 2006.** Potential selection in native grass populations by exotic invasion. *Molecular Ecology* **15**: 2291–2300.

**Michelmore RW, Paran I, Kesseli RV. 1995.** Identification of markers linked to disease-resistance genes by bulked segregant analysis: a

rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of National Academy of Sciences*, *USA* **88**: 9828–9832.

**Morrell PL, Toleno DM, Lundy KF, Clegg MT. 2005.** Low levels of linkage disequilibrium in wild barley (*Hordeum vulgar*e ssp. *spontaneum*) despite high rates of self-fertilization. *Proceedings of the National Academy of Sciences*, *USA* **112**: 2442–2447.

**Nei M. 1978.** Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**: 583–590.

**Nielsen R. 2005.** Molecular signatures of natural selection. *Annual Review of Genetics* **39**: 197–218.

**Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, *et al*. 2002.** The extent of linkage disequilibrium in *Arabidopsis thaliana. Nature Genetics* **30**: 190–193.

**Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD. 2006.** Selection under domestication: evidence for a sweep in the rice *Waxy* genomic region. *Genetics* **173**: 975–983.

**Palaisa KA, Morgante M, Williams M, Rafalski A. 2003.** Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *Plant Cell* **15**: 1795–1806.

**Papa R, Gepts P. 2003.** Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. *Theoretical and Applied Genetics* **106**: 239–250.

**Papa R, Acosta J, Delgado-Salinas A, Gepts P. 2005.** A genome-wide analysis of differentiation between wild and domesticated *Phaseolus vulgaris* from Mesoamerica. *Theoretical and Applied Genetics* **111**: 1147–1158.

**Pariset L, Cappuccio I, Joost S, D'Andrea M, Marletta D, Ajmone Marsan P, Valentini A, the ECONOGENE Consortium. 2006.** Characterization of single nucleotide polymorphisms in sheep and their variation as evidence of selection. *Animal Genetics* **37**: 290–292.

**Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, *et al*. 2001.** Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of National Academy of Sciences*, *USA* **98**: 11479–11484.

**Singh SP, Molina A, Gepts P. 1995.** Potential of wild common bean for seed yield improvement of cultivars in the tropics. *Canadian Journal of Plant Science* **75**: 807–813.

**Sonnante G, Stockton T, Nodari RO, Becerra Velásquez VL, Gepts P. 1994.** Evolution of genetic diversity during the domestication of common-bean (*Phaseolus vulgaris* L.). *Theoretical and Applied Genetics* **89**: 629–635.

**Storz JF. 2005.** Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* **14**: 671–688.

**Tanksley SD, McCouch SR. 1997.** Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* **277**: 1063–1066.

**Tenaillon MI. 2001.** Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proceedings of National Academy of Sciences*, *USA* **98**: 9161–9166.

**Vasemägi A, Nilsson J, Primmer CR. 2005.** Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). *Molecular Biology and Evolution* **22**: 1067–1076.

**Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J. 2002.** Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proceedings of National Academy of Sciences*, *USA* **99**: 9650–9655.

**Vigouroux Y, Mitchell S, Matsuoka Y, Hamblin M, Kresovich S, Stephen J, *et al*. 2005.** An analysis of genetic diversity across the maize genome using microsatellites. *Genetics* **169**: 1617–1630.

**Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, Hornes M, *et al*. 1995.** AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**: 4407–4414.

**Weir BS. 1996.** *Genetic data analysis II*, 2nd edn. Sunderland, MA: Sinauer Associates, Inc.

**Weir BS, Cockerham C. 1984.** Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

**Wright SI, Vroh Bi I, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. 2005.** The effects of artificial selection on the maize genome. *Science* **308**: 1310–1314.

**Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, Doebley JF, Gaut BS, McMullen MD. 2005.** A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *The Plant Cell* **17**: 2859–2872.