Genet. Res., Camb. (2011), **93**, *pp.* 65–75.
doi:10.1017/S0016672310000509

MADS-box genes of maize: frequent targets of selection during domestication

QIONG ZHAO[1†‡], ALLISON L. WEBER[1*†], MICHAEL D. MCMULLEN[2], KATHERINE GUILL[2] AND JOHN DOEBLEY[1]
[1] *Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA*
[2] *Plant Genetic Research Unit, USDA-Agricultural Research Service; and Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA*

(*Received 19 July 2010 and in revised form 10 October 2010; first published online 14 December 2010*)

Summary

MADS-box genes encode transcription factors that are key regulators of plant inflorescence and flower development. We examined DNA sequence variation in 32 maize MADS-box genes and 32 randomly chosen maize loci and investigated their involvement in maize domestication and improvement. Using neutrality tests and a test based on coalescent simulation of a bottleneck model, we identified eight MADS-box genes as putative targets of the artificial selection associated with domestication. According to neutrality tests, one additional MADS-box gene appears to have been under selection during modern agricultural improvement of maize. For random loci, two genes were indicated as targets of selection during domestication and four additional genes were indicated to be candidate-selected loci for maize improvement. These results suggest that MADS-box genes were more frequent targets of selection during domestication than genes chosen at random from the genome.

1. Introduction

It has been proposed that the evolution of plant morphology often involves changes in genes coding for transcriptional regulators (Doebley & Lukens, 1998; Cronk, 2001). Several studies have provided examples where changes in the expression or function of transcription factors can give rise to morphological differences in plant architectures, leaf morphology, inflorescence structure and floral configuration (Doebley *et al*., 1997; Hellmann *et al*., 2003; Kim *et al*., 2003; Wang *et al*., 2005).

The MIKC-type (Type II) MADS-box genes encode transcription factors that are key regulators of plant vegetative and reproductive development (Riechmann & Meyerowitz, 1997*a*; Riechmann & Meyerowitz, 1997*b*; Theissen & Saedler, 1999; Theissen *et al*., 2000; Ng & Yanofsky, 2001; Theissen, 2001). Type II MADS-box proteins possess four functional domains, the M (DNA-binding), K (keratin-like), I (intervening) and C (C-terminal) domain. The M domain usually contains ~58 amino acids and is the most conserved region of the MADS protein sequence (Riechmann & Meyerowitz, 1997*b*). The K and I domains are involved in protein–protein dimerization and interaction (Sieburth *et al*., 1995; Fan *et al*., 1997; Egea-Cortines *et al*., 1999). The less conserved C domain is responsible for transactivation, formation of multimeric protein complexes and specificity of protein function (Honma & Goto, 2001; Immink *et al*., 2003). Alterations in the C domain were shown to contribute to diversification and neofunctionalization during floral MADS-box gene evolution (Vandenbussche *et al*., 2003).

Molecular evolution studies showed that the duplication and functional diversification of the MADS-box genes are correlated with the origin of land plants, the establishment of certain floral structures in higher plants and the increasingly diverse and complex flower structures in land plants (Theissen *et al*.,

* Corresponding author. Department of Genetics, North Carolina State University, Thomas Hall 3555, Box 7614, NCSU Campus, Raleigh, NC 27695-7614, USA. Tel: (608)-212-9060. Fax: (919)-515-3355. e-mail: alweber@ncsu.edu
† These authors contributed equally to this work.
‡ Current address: Otsuka Pharmaceuticals, Princeton, NJ 08540, USA.

1996; Purugganan, 1997; Saedler *et al*., 2001; Litt & Irish, 2003; He *et al*., 2004; Kaufmann *et al*., 2005). Since MADS-box genes were critically relevant to the long-term evolution of plant form, the genetic modification at these genes could also provide a source of diversity to be utilized for creating intraspecific morphological variation. However, a few investigations have looked at the evolution pattern of within-species sequence variation at MADS-box genes. A limited number of studies include examinations of *CAL*, *AP3*, *PI*, *SEP1-2* and *SHP1-2* genes in *Arabidopsis* and *Brassica* (Purugganan & Suddith, 1998, 1999; Purugganan *et al*., 2000; Moore *et al*., 2005). Positive selection was detected in the pattern of sequence variation at the *CAL* gene of the domesticated subspecies of *Brassica oleracea* and was used as evidence to suggest that specific *CAL* alleles were selected by early farmers to modify inflorescence structure in *B. oleracea* (Purugganan *et al*., 2000).

Maize (*Zea mays* ssp. *mays*) domestication and improvement processes provide a good system to examine the contribution of MADS-box genes to morphological evolution (Eyre-Walker *et al*., 1998; Vigouroux *et al*., 2002; Wright & Gaut, 2005). The maize gene pool is composed of three components: maize inbreds, maize landraces and teosintes, including *Z. mays* ssp. *parviglumis*, the direct progenitor of maize. Starting with teosinte, native people of the New World constantly selected certain traits to meet different cultural and agricultural needs, and thus produced domesticated maize and diversified this crop into many landraces (Pressoir & Berthaud, 2004). More recently, multiple maize inbred lines important for breeding have been created by selection on landrace populations.

The purpose of this study was to examine the MADS-box genes for their involvement in maize domestication and improvement. We compared the pattern of genetic polymorphism of 32 MADS-box genes to that in loci randomly chosen from the maize genome. The possibility of MADS-box and control genes being putative targets of selection was evaluated by neutrality tests and a test based on a bottleneck model of domestication.

2. Materials and methods

(i) *Plant materials and sequence data*

Our sample of 32 MADS-box genes included 30 previously described genes (Theissen *et al*., 1996; Münster *et al*., 2002; Vigouroux *et al*., 2002) (Fig. 1). Additional MADS-box genes were found by querying the Entrez and the Maize Assembled Genomic Island (MAGI, version of April 2004) translated nucleotide databases using TBLASTN (http://www.ncbi.nlm.nih.gov/BLAST). This strategy identified two unique type-II MADS-box genes, AY109828 and CA483635.

DNA sequences were obtained for the 32 MADS-box genes by PCR amplifying and sequencing a 300 to 1500 bp DNA fragment in a common set of 28 different maize inbred lines, 16 maize landraces and 21 teosinte accessions (Supplementary Table 1). In MADS-box genes with strong evidence of selection, multiple amplicons within the gene were tested in order to determine the extent of selection but were not included in analyses comparing the original MADS-box-sequenced amplicons to those from the control genes. The 28 maize inbreds represent much of the genetic diversity among important public lines currently available for breeding (Liu *et al*., 2003). The 16 maize landraces represent the genetic diversity present in maize before modern breeding efforts (Tenaillon *et al*., 2001). Sixteen of the 21 different teosintes were chosen on the basis of geographic criteria and cover the entire natural distribution of *Z. mays* ssp. *parviglumis*. Single alleles for each specific gene were isolated from *Tripsacum*, a sister genus of *Zea*, or *Zea diploperennis*, when a *Tripsacum* sequence was not obtainable (Supplementary Table 1). Sampled fragments mainly encode I, K and C domains of the MADS proteins. Previously generated DNA sequences of *zag11* in maize landraces and the outgroup were also used for analysis (Vigouroux *et al*., 2002).

Thirty-two randomly chosen genes served as controls for comparison to the MADS-box genes. These genes represent a subset of 774, loci which had previously been sequenced in maize inbreds and teosinte individuals (Wright *et al*., 2005). From the 774 loci, we randomly chose a subset of 32 loci that were $\geqslant$500 bp in sequence length and had been successfully sequenced in at least eight maize inbreds and at least eight teosintes. We subsequently sequenced these 32 genes in 16 maize landraces and an outgroup. A smaller set of teosintes (16 accessions of *Z. mays* ssp. *parviglumis*) and maize inbreds (14 lines) were sampled for the 32 control genes in comparison with the MADS-box genes. In order to eliminate false-positive results due to sampling differences, we only included sequence data from 16 *Z. mays* ssp. *parviglumis* individuals and 14 maize inbred lines when comparing the nucleotide polymorphism data from the MADS-box and control genes.

For maize landraces, maize inbreds and the teosinte individuals, we were able to directly sequence PCR products using a standard protocol (Applied Biosystems, Foster City, CA) from homozygous or haploid DNA sources. Our DNA sources for *Z. diploperennis* and *Tripsacum* DNAs are potentially heterozygous, and thus PCR products from these sources were cloned into the TOPO-TA vector (pCR 2.1-TOPO kit, Invitrogen, Carlsbad, CA) and multiple clones were sequenced to identify a single allele and

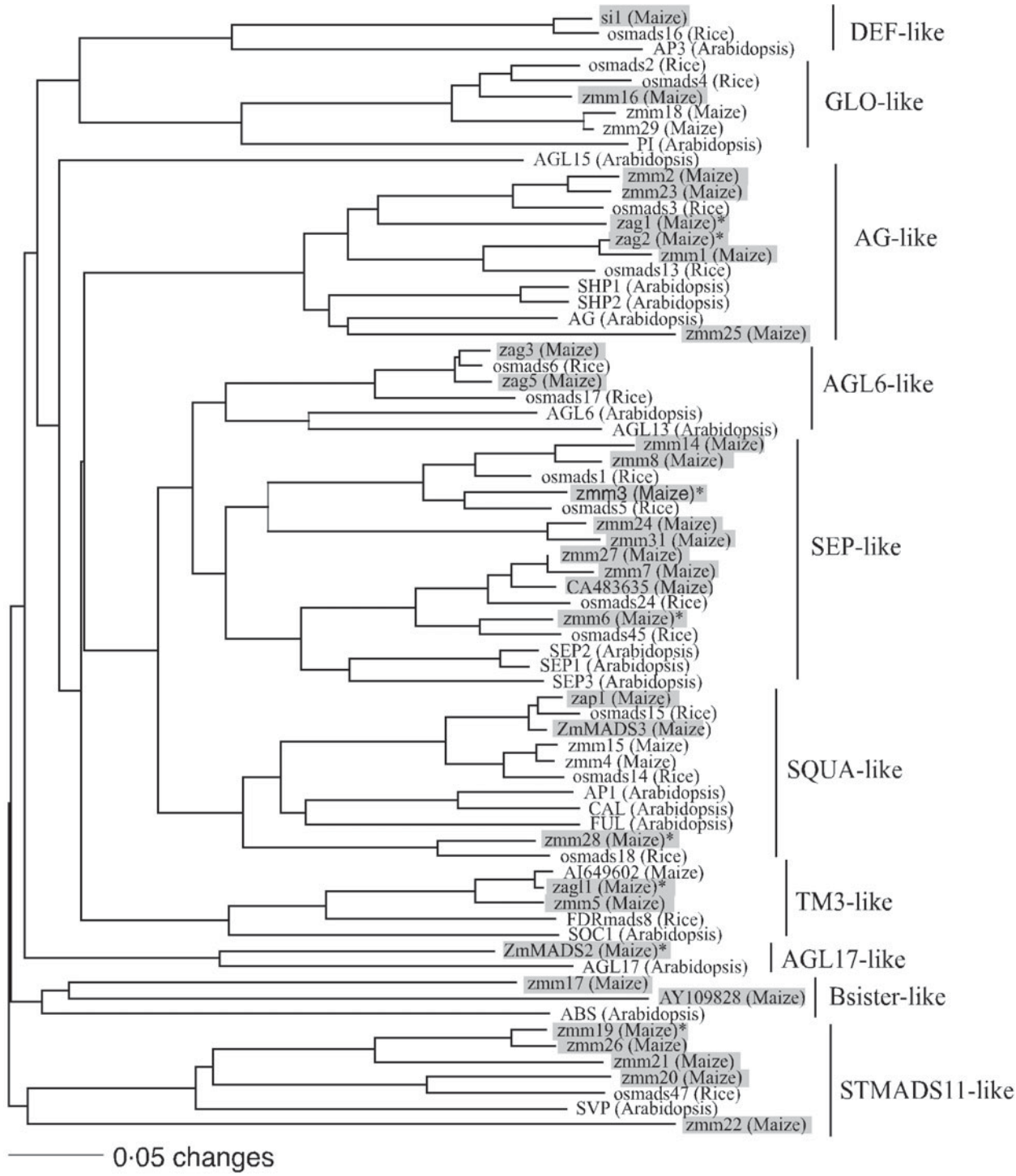

Fig. 1. The phylogenetic relationship of maize MADS-box genes with homologous genes in rice and *Arabidopsis*. The tree was constructed using amino acid sequences of the MIK domains and the neighbour-joining (NJ) method with the distance option of mean character difference in the PAUP* 4·0b10 (Swofford, 2003). Subfamilies were named according to Münster *et al*. (2002). Genes shaded with boxes are the maize MADS-box genes surveyed in this study. The eight MADS-box genes that our analysis implicated as being under selection during domestication are indicated with an asterisk.

correct *Taq* errors introduced during PCR. The forward and reverse DNA sequences were assembled for each individual using the Sequencher software (Gene Codes, Ann Arbor, MI). Individual sequences from the maize inbreds, maize landraces, teosinte and an outgroup were then manually aligned using SE-Al version 2.0 a11 (Rambaut, 1996). Unique single-base-pair variants (singletons) were double checked by manually inspecting the corresponding raw chromatogram peaks.

Table 1. *Sequence statistics of 32 control genes and 32 MADS-box genes*

	Control genes				MADS-box Genes			
	N[a]	L[b]	S[c]	θ[d]	N[a]	L[b]	S[c]	θ[d]
Maize inbreds	11·81	609·3	14·1	0·0077	26·0	653·9	21·0	0·0088
Maize landraces	15·2	610·8	18·1	0·0091	15·5	653·1	18·25	0·0088
Teosinte	12·4	603·9	25·8	0·0142	11·1	648·2	29·7	0·0164

[a] Average number of sequences in the alignment.
[b] Average length of alignments, excluding gaps.
[c] Average number of segregating sites (SNPs) in the alignments.
[d] Average amount of nucleotide polymorphism (Watterson's estimator of population mutation parameter).

(ii) *Tests for neutrality*

Molecular population genetic statistics were generated using DnaSP Version 4.0 (Rozas *et al*., 2003). Two estimators of the population mutation rate, nucleotide diversity (π) (Tajima, 1983) and nucleotide polymorphism (θ) (Watterson, 1975) were calculated based on all sites. Three neutrality tests, Tajima's *D* test (Tajima, 1989), Fay and Wu's *H* test (Fay & Wu, 2000) and the Hudson–Kreitman–Aguadé (HKA) test (Hudson *et al*., 1987), were performed to test for selection. A multi-locus HKA test (http://genfaculty.rutgers.edu/hey/software#HKA) was performed for testing the overall fitness of the observed nucleotide polymorphism and divergence of the 32 MADS-box genes or the 32 control genes to a neutral equilibrium model. The pair-wise HKA test, as implemented in DnaSP (Rozas *et al*., 2003), was performed to test for selection at each individual locus. Eleven neutral loci (*adh1*, *an1*, *asg75*, *bz2*, *csu1138*, *csu1171*, *csu381*, *csu1132*, *fus6*, *glb1* and *umc128*) (Eyre-Walker *et al*., 1998; Hilton & Gaut, 1998; Tenaillon *et al*., 2001) were used for HKA tests involving maize landraces. A smaller set of neutral loci (*adh1*, *glb1*, *bz2*, *csu1132* and *csu1171*) was available and used for HKA tests involving maize inbreds and teosinte (Tenaillon *et al*., 2004). The overall χ^2 value for each pair-wise HKA test was calculated by summing up the χ^2 values across different neutral loci.

(iii) *Coalescent-simulation-based approach to testing for selection*

For each MADS-box gene a coalescent-simulation-based (CS) test was performed to determine whether the gene was a potential target of selection during domestication. We used a modified version of the standard coalescence procedure (Hudson *et al*., 1987) that incorporated the domestication bottleneck as previously described (Eyre-Walker *et al*., 1998). All parameters in the model were assigned to previously established values (Eyre-Walker *et al*., 1998; Tenaillon *et al*., 2004). The severity of the bottleneck (k) was defined as a function of the population size during the bottleneck (N_b) and the duration of the bottleneck (d) such that $k = N_b/d$. Using sequence data from 30 neutral control genes, the best multilocus estimate of k was found to be 1·8 using methods previously described (Tenaillon *et al*., 2004). To estimate k, we used the number of segregating sites (S) as the summary statistic and explored d values of 500, 1000 and 1500 generations. Finally, k values ranging from 0·5 to 5 (in increments of 0·1) were explored.

We used the coalescence model described above to test for selection in 32 MADS-box genes. This model was implemented using a program provided by Innan & Kim (2004). For each of the 32 MADS-box genes, 10 000 simulations were conducted. The number of segregating sites S_{simul} was calculated for each of the 10 000 simulated sequence sets. A gene was considered to be a potential target of selection during domestication, if the observed $S_{maizelandraces}$ was $<97{\cdot}5\%$ of the S_{simul} values.

3. Results

(i) *Nucleotide diversity in maize and teosinte*

First, we compared sequence diversity between the 32 MADS-box genes and the 32 genes chosen at random from the genome. The number of maize landraces and teosintes assayed and the average sequence length sampled were similar in the MADS-box and control genes (Table 1). The proportion of sequence diversity maintained in the maize landraces (or inbreds) to that in teosinte ($r = \theta_{maize}/\theta_{teosinte}$) was calculated for each gene. When averaged across the control genes, maize landraces retained 64·1% of the genetic diversity found in teosinte. The MADS-box genes retained less sequence diversity (53·4%) than the control genes. This is evident in a plot of nucleotide diversity in maize landraces by that in teosinte where the MADS-box genes have values closer to the x-axis as compared to the control genes (Fig. 2). The values

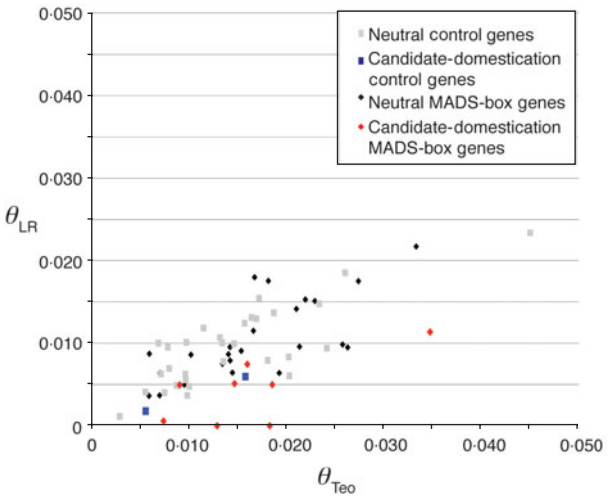

Fig. 2. Nucleotide diversity maintained in maize landraces. Nucleotide polymorphism (θ) (Watterson, 1975) in maize landraces (y-axis) plotted against that observed in teosinte (x-axis). Values for the 30 neutral control genes are represented by grey squares; the two candidate-domestication neutral genes excluded when tuning the bottleneck model are represented by blue squares. Values for the 24 neutral MADS-box genes are represented by black diamonds and the values for the eight candidate-domestication MADS-box genes are represented by the red diamonds.

observed in the control genes correspond to previous studies that reported values for maize landraces or maize inbreds ranging from 57 to 80% (Tenaillon *et al*., 2004; Wright *et al*., 2005). This provides evidence that the majority of control genes are neutral (did not undergo selection) and therefore serve as a good control set to compare with the MADS-box genes.

(ii) *Statistical tests for neutrality*

A multi-locus HKA test was conducted separately for maize inbreds, landraces and teosintes to assess the overall fit of MADS-box genes or control genes to the neutral model. For MADS-box genes, the null hypothesis of neutrality was rejected in both maize inbreds ($P<0{\cdot}001$) and landraces ($P<0{\cdot}001$), but not in teosinte ($P=0{\cdot}71$). However, for the control genes, the test was only marginally significant for maize inbreds ($P=0{\cdot}04$), but not for the maize landraces ($P=0{\cdot}858$) or teosinte ($P=0{\cdot}818$). These results demonstrate that the MADS-box genes lost more genetic diversity during domestication than a sample of control genes from the genome.

We performed pair-wise HKA tests for selection at individual loci. A gene was considered as a candidate selection gene for domestication if it had significant results for the HKA test in maize landraces but not in teosinte. A gene was considered as a candidate selection gene for improvement if it had a significant HKA test result in maize inbreds but not in maize landraces or teosinte. For nine MADS-box genes (*zag1*, *zag2*, *zagl1*, *zmm3*, *zmm6*, *zmm19*, *zmm20*, *zmm28* and *ZmMADS2*) the pair-wise HKA test was significant in maize landraces but not in teosinte (Table 2). All these genes demonstrated a reduced level of polymorphism except for *zmm20*, where more polymorphism was observed among the maize landraces than expected under the neutral evolution model. It is difficult to interpret the HKA test results for *zmm20*. There are ten haploptyes inferred from 16 landraces; however, there is no strong evidence suggesting that these ten haplotypes underwent diversifying or balancing selection. Only one MADS-box gene, *zmm22*, yielded a significant pair-wise HKA test in maize inbreds but not in maize landraces (Table 2). Collectively, based on the results of pair-wise HKA tests, signatures of domestication selection were evident at *zag1*, *zag2*, *zagl1*, *zmm3*, *zmm6*, *zmm19*, *zmm28* and *ZmMADS2* (Fig. 2), and evidence of improvement selection was found at *zmm22*. Within the control genes, we identified AY111689 as a candidate domestication gene and four additional genes, AY105750, AY108201, AY111546 and AY112456 as candidate improvement genes (Table 3).

(iii) *Tests for selection under the domestication bottleneck model*

The bottleneck model tuned by the 30 neutral control genes was used in order to test for selection for 32 MADS-box genes (Supplementary Results). Control genes AY111689 and AY111546 were excluded because of their domestication gene candidacy (Fig. 2) as indicated by the pair-wise HKA test (Table 3) or results obtained from Tajima's D and Fay and Wu's H test (Supplementary Table 2). We simulated the sequence evolution for each MADS-box gene under the bottleneck model with bottleneck severity (k) equal to 1·8 (when $d=1000$) with 10 000 replicates. If the observed number of segregating sites, S, in maize landraces was smaller than 97·5% of its simulated values, then the bottleneck effect alone could not explain the observed severe reduction in S and a past selection event may have occurred. For three of the 32 MADS-box genes (*zag2*, *zagl1* and *zmm6*), the observed S was significantly smaller than expected from the bottleneck effect alone (Table 2). Therefore, coalescent simulation of the bottleneck model provided evidence of selection during domestication at these three loci.

The bottleneck model also provides an opportunity to test whether as a group the MADS-box genes are enriched for domestication genes relative to the group of control genes. If MADS-box genes were targeted by selection more frequently than genes chosen at random from the genome, the bottleneck severity estimated using the 32 MADS-box genes is

Table 2. *Results of the pair-wise HKA and CS tests for 32 MADS-box genes*

Gene	Panzea Marker[a]	*P*-values from pair-wise HKA test				*P*-values from CS test	
		Maize inbreds	Maize landraces	Teosinte	Selection status	Maize versus teosinte	Selection status
AY109828	PZD00004	0·785	0·404	0·786	–	0·903	–
CA483635	PZD00005	0·996	0·357	0·918	–	0·453	–
silky1	PZD00072	0·916	0·443	0·795	–	0·959	–
zag1	PZD00011	<0·001	<0·001	0·090	Domestication	0·343	–
zag2	PZD00016	0·659	0·000	0·971	Domestication	<0·001	Domestication
zag3	PZD00017	0·935	0·254	0·972	–	0·197	–
zag5	PZD00018	0·960	0·789	0·893	–	0·881	–
zagl1	PZD00021	<0·001	<0·001	0·955	Domestication	<0·001	Domestication
zap1	PZD00022	0·278	0·059	0·267	–	0·891	–
zmm1	PZD00033	0·985	0·965	0·980	–	0·648	–
zmm2	PZD00034	0·783	0·438	0·985	–	0·115	–
zmm3	PZD00048	0·260	0·010	0·975	Domestication	0·070	–
zmm5	PZD00050	0·957	0·930	0·996	–	0·208	–
zmm6	PZD00053	<0·001	<0·001	0·796	Domestication	0·006	Domestication
zmm7	PZD00054	0·950	0·991	0·874	–	0·657	–
zmm8	PZD00055	0·990	0·961	0·996	–	0·829	–
zmm14	PZD00026	0·990	1·000	0·916	–	0·658	–
zmm16	PZD00027	0·399	0·237	0·572	–	0·787	–
zmm17	PZD00028	0·936	0·988	0·823	–	0·622	–
zmm19	PZD00030	<0·001	<0·001	0·055	Domestication	0·524	–
zmm20	PZD00035	0·139	0·002[b]	0·464	–	0·050[b]	–
zmm21	PZD00036	0·990	0·955	0·994	–	0·978	–
zmm22	PZD00037	<0·001	0·674	0·439	Improvement	0·518	–
zmm23	PZD00038	0·847	0·590	0·967	–	0·093	–
zmm24	PZD00039	0·947	0·953	0·994	–	0·440	–
zmm25	PZD00040	0·655	0·283	0·827	–	0·573	–
zmm26	PZD00041	0·096	0·090	0·057	–	0·973	–
zmm27	PZD00042	0·584	0·959	0·947	–	0·080	–
zmm28	PZD00044	0·919	0·030	0·421	Domestication	0·295	–
zmm31	PZD00046	0·951	0·985	0·979	–	0·999	–
ZmMADS2	PZD00056	0·904	0·010	0·158	Domestication	0·648	–
ZmMADS3	PZD00057	0·840	0·982	0·900	–	0·904	–

[a] These sequences are publically available at www.panzea.org and have been deposited into Genbank (HM992947-HM994864).
[b] The neutral evolution model was rejected because more polymorphisms were observed among the maize landraces than expected under the model.

expected to be much smaller than that for the 32 control genes. When $d=1000$ and S was used to fit the bottleneck intensity, the approximate maximum likelihood (ML) estimate of k is equal to 1·0 for 32 MADS-box genes versus 1·6 for 32 control genes. The likelihoods for 32 control genes under $k=1{\cdot}6$ and $k=1{\cdot}0$ are $1{\cdot}60\times10^{-19}$ and $3{\cdot}48\times10^{-21}$, respectively, and this difference is statistically significant as indicated by the likelihood ratio (LR) test ($-2\ln\mathrm{LR}=7{\cdot}66$, $P=0{\cdot}006$). Thus, the control genes fit better with the less severe bottleneck. Reciprocally, the likelihoods for the 32 MADS-box genes under $k=1{\cdot}6$ and $k=1$ are $1{\cdot}27\times10^{-24}$ and $5{\cdot}98\times10^{-23}$, respectively, and again this difference is significant ($-2\ln\mathrm{LR}=7{\cdot}70$, $P=0{\cdot}006$). This result indicates that the MADS-box genes fit better with the more severe bottleneck. Therefore, the 32 MADS-box genes experienced a significantly more severe bottleneck than the 32 control genes.

To test whether our 'neutral' MADS-box genes fit the 'neutral' bottleneck model, we evaluated estimates of k over 24 potentially neutral MADS-box genes (as assessed by the pair-wise HKA and CS tests) using $\pm20\%$ S_{maize} as a fitting criterion when $d=1000$. The approximate maximum likelihood was located at $k=1{\cdot}5$ with a confidence interval ranging from $\sim1{\cdot}1$ to $\sim2{\cdot}2$. The likelihoods for 24 MADS-box genes under $k=1{\cdot}8$ and $k=1{\cdot}5$ are $4{\cdot}40\times10^{-13}$ and $6{\cdot}63\times10^{-13}$, respectively, which are not statistically different ($-2\ln\mathrm{LR}=0{\cdot}41$, $P=0{\cdot}52$). Hence, the 'neutral' MADS-box genes overall fit well with the 'neutral' bottleneck model tuned by the 30 neutral control genes. In addition, the consistency of k estimates between neutral MADS-box and neutral

Table 3. *Results of the pair-wise HKA and CS tests for 32 control genes*

		P-values from pair-wise HKA test				*P*-values from CS test	
Gene	Panzea marker	Maize inbreds	Maize landraces	Teosinte	Selection status	Maize versus teosinte	Selection status
AY104395	PZA02980	0·094	0·969	0·958	–	0·621	–
AY105273	PZA02792	0·487	0·131	0·237	–	0·914	–
AY105750	PZA00188	0·006	0·845	0·212	Improvement	0·123	–
AY106506	PZA00029	0·109	0·57	0·626	–	0·804	–
AY106816	PZA00638	0·536	0·494	0·829	–	0·671	–
AY106876	PZA00618	0·967	0·941	0·919	–	0·119	–
AY107192	PZA00522	0·992	0·963	0·987	–	0·897	–
AY107248	PZA00508	0·984	0·953	0·98	–	0·467	–
AY107317	PZA00496	0·971	0·237	0·936	–	0·251	–
AY107667	PZA00440	0·892	0·446	0·893	–	0·857	–
AY107692	PZA00436	0·905	0·748	0·773	–	0·271	–
AY107967	PZA00393	0·730	<0·001	0·037	Teosinte	0·710	–
AY107970	PZA00392	0·784	0·979	0·881	–	0·410	–
AY108077	PZA00382	0·952	0·955	0·567	–	0·535	–
AY108087	PZA00381	0·686	0·803	0·605	–	0·784	–
AY108201	PZA00365	0·009	0·828	0·186	Improvement	0·237	–
AY108540	PZA00300	0·95	0·966	0·634	–	0·141	–
AY108738	AY108738	0·834	0·957	0·978	–	0·719	–
AY110882	PZA00256	0·066	0·975	0·877	–	0·658	–
AY110897	PZA00100	0·768	0·567	0·591	–	0·997	–
AY110958	PZA00097	0·915	0·983	0·976	–	0·521	–
AY110964	PZA00096	0·970	0·983	0·956	–	0·268	–
AY111431	PZA00731	0·684	0·058	0·245	–	0·767	–
AY111546	PZA00719	0·023	0·840	0·993	Improvement	0·234	–
AY111689	PZA00710	0·644	0·002	0·492	Domestication	0·222	–
AY111711	PZA00706	1·000	0·949	0·992	–	0·992	–
AY111763	PZA00704	0·041	<0·001	0·002	Teosinte	0·903	–
AY111798	PZA00694	0·362	0·993	0·847	–	0·101	–
AY112207	PZA00570	0·868	0·573	0·940	–	0·404	–
AY112358	PZA00553	0·650	0·986	0·044	Teosinte	0·012	Domestication
AY112456	PZA00543	<0·001	0·097	0·449	Improvement	0·497	–
AY112544	PZA00538	0·946	0·974	0·834	–	0·630	–

control genes suggests that $k = \sim 1{\cdot}5$–$1{\cdot}8$ may well represent a general effect of bottleneck on maize sequence variation.

(iv) *Selection sweeps in candidate domestication MADS-box genes*

In order to investigate the extent of the selection sweeps in seven of the candidate domestication MADS-box genes (*zag1*, *zag2*, *zagl1*, *zmm3*, *zmm6*, *zmm19* and *zmm28*), we sampled DNA polymorphism among maize landraces and teosinte in additional coding and 5′ regulatory regions (Fig. 3). Evidence of selection was found in the additional sampled coding sequences from five (*zag1*, *zag2*, *zagl1*, *zmm6* and *zmm28*) of the candidate domestication genes (Fig. 3, Table 4). A single haplotype was fixed among all maize landraces sampled in the 3′ coding regions of both *zag2* and *zagl1*, suggesting the presence of the selected site(s) within these regions. The evidence of selection was limited to the initially sequenced regions in the other two candidate domestication genes, *zmm3* and *zmm19*. In contrast to the coding regions, the examined 5′ regulatory regions did not show any evidence of selection associated with domestication in any of the seven genes.

4. Discussion

We tested 32 MADS-box genes to determine if they were under selection during maize domestication or improvement. Neutrality and coalescent simulation-based (CS) tests identify eight of these genes as putative domestication genes and one as a putative improvement gene. In order to assess if MADS-box genes were more frequent targets of selection than expected by chance, we also tested 32 randomly chosen genes from the genome for signatures of selection. This comparison indicates that the MADS-box genes are more enriched for selected genes than would be expected by chance.

Table 4. *Results of the pair-wise HKA and CS tests for additional regions sequenced in candidate domestication MADS-box genes*

Gene	Panzea marker	*P*-values from pair-wise HKA test			*P*-values from CS test	
		Maize landraces	Teosinte	Selection status	Maize versus teosinte	Selection status
zag1	PZD00012	<0·001	0·437	Domestication	0·014	Domestication
zag2	PZD00013	<0·001	<0·001	Teosinte	0·499	–
	PZD00014	<0·001	0·023	Teosinte	0·41	–
	PZD00015	<0·001	0·969	Domestication	0·002	Domestication
zagl1	PZD00081	0·007	<0·001	Teosinte	0·757	–
	PZD00020	0·000	0·729	Domestication	0·002	Domestication
zmm3	PZD00047	0·738	0·995	–	0·639	–
	PZD00049	0·202	0·49	–	0·885	–
zmm6	PZD00051	0·974	0·937	–	0·586	–
	PZD00052	0·02	0·404	Domestication	0·939	–
zmm19	PZD00029	0·948	0·357	–	<0·001[a]	–
	PZD00031	0·822	0·123	–	0·783	–
zmm28	PZD00043	0·138	0·586	–	0·634	–
	PZD00045	0·039	0·443	Domestication	0·927	–

[a] The neutral evolution model was rejected because more polymorphisms were observed among the maize landraces than expected under the model.

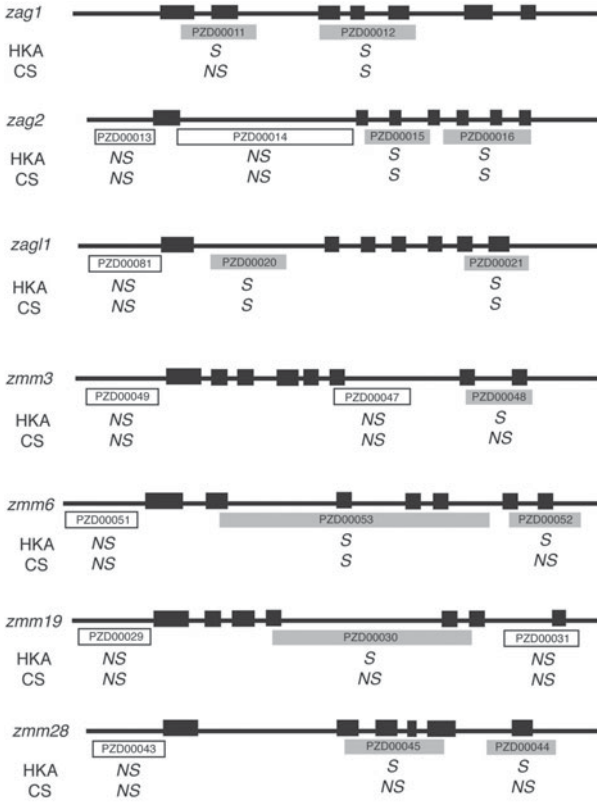

Fig. 3. Extended survey of the selection sweeps in seven MADS-box candidate-domestication genes. Solid black boxes represent exons and lines represent UTR regions or introns. White and grey boxes represent sequenced regions and are labelled with the corresponding PANZEA marker number (http://www.panzea.org). The HKA or coalescent simulation (CS) test gave no evidence of selection (*NS*) for amplicons depicted as white boxes. Amplicons depicted as grey boxes had evidence of selection (*S*) in at least one of the two tests.

(i) *Search for targets of selection*

Three of the eight MADS-box genes were identified as putative targets of selection during domestication by both the pair-wise HKA and the CS test. The other five putative domestication MADS-box genes were identified by the pair-wise HKA test alone. There are many possible reasons why the results from the pair-wise HKA and the CS test were not entirely consistent. First, the CS test is heavily influenced by the model parameters used. In our analysis, we estimated the population recombination rate using data from teosinte. This was most likely an underestimate due to the fact that only a small proportion of the recombination events that occurred can be detected based on population genetic data (Hudson & Kaplan, 1985; Stumpf & McVean, 2003). This would lead to wide distributions for summary statistics (*e.g.*, S, π), making the test more conservative.

Second, selection associated with domestication could have acted on existing alleles with moderate frequency in the teosinte populations as opposed to newly arising mutations. In such cases, selection does not necessarily leave an apparent signature on the patterns of nucleotide variation in the regions closely linked to the selected site. If the initial frequency of the beneficial allele, p, is $<0{\cdot}2$, the signature of artificial selection can be captured with a reasonably high probability, but the chance of detecting selection is very low when $p>0{\cdot}5$ (Innan & Kim, 2004). Therefore, various tests could yield inconsistent results due to the weak trace left by the selection event. For example, when p is small, selection is likely to be detected by the HKA test; however, when p is

moderate and polymorphism is not significantly reduced, those tests looking at the allele frequency spectrum (e.g. Tajima's *D* and Fay and Wu's *H* tests) will have more power (Innan & Kim, 2004).

Third, the chance to detect positive selection is also a function of the strength of selection and the amount of recombination between the selected and neutral sites (Braverman *et al*., 1995; Przeworski, 2002; Przeworski, 2003; Wright & Gaut, 2005). Borderline evidence of selection may be found at loci that were under weak selection. Recombination could have broken down the linkage between the selected site and the regions we have surveyed and hence only marginally or nearly significant results were found in the sampled regions.

(ii) *MADS-box genes as frequent targets of selection during maize domestication*

There are striking differences in inflorescence and plant architecture between maize and teosinte as a result of selection during domestication. MADS-box transcription factors are known to affect both inflorescence and plant architecture in various plant species (Gu *et al*., 1998; De Bodt *et al*., 2003). Accordingly, we hypothesized that MADS-box genes had contributed to the morphological change of maize during domestication and would be more enriched for domestication genes than expected by chance.

Several lines of evidence support the hypothesis that MADS-box genes were more frequently targeted by selection during domestication than a comparable set of genes chosen at random. First, the multi-locus HKA test conducted using the maize landraces detected evidence of selection among the MADS-box genes but not for the control genes. This same test did not detect any evidence of selection for either group of genes in teosinte. Second, the bottleneck intensity based on the 32 MADS-box genes is significantly more severe than the bottleneck intensity estimated over the 32 control genes. The parameter of the bottleneck intensity, k, is $\sim 1{\cdot}0$ for 32 MADS-box genes versus $\sim 1{\cdot}6$ for 32 control genes. The likelihood ratio test indicated that this difference of k was significant ($-2\ln \mathrm{LR} = 7{\cdot}66$, $P = 0{\cdot}006$). Third, the proportion of MADS-box genes identified as domestication gene candidates by the pair-wise HKA test was higher than that observed for the control genes. Eight out of the 32 (25%) MADS-box genes were classified into the 'domestication' class as opposed to only one out of the 32 (3·1%) control genes. Moreover, if we compare 25% to an empirical estimate of the proportion of selected genes during maize domestication and improvement (~2–4%) (Wright *et al*., 2005), then 25% is significantly higher than the upper bound of the estimate (~4%) (Binomial test, $P < 0{\cdot}001$). In summary, the higher proportion of domestication genes in the MADS-box gene family, together with results from other independent tests, provides evidence for the MADS-box genes being more frequent targets of selection during domestication than expected by chance.

Our results argue that MADS-box genes have served an important role in the morphological change that was selected for during maize domestication. Additional experimentation and analysis in other crops will be necessary to see if this phenomenon is limited to maize. Preliminary evidence suggests that MADS-box genes will prove to be an important source of domestication genes and varietal differences in other systems (Purugganan *et al*., 2000; Smith & King, 2000; Vrebalov *et al*., 2002; Yan *et al*., 2003). This study also provides further evidence that transcription factors are over-represented among domestication genes (Doebely, 2006). This observation suggests that further sequencing and analysis of transcription factor families could result in the identification of other domestication genes and subsequently, clarify our understanding of the domestication process.

This work was funded by National Science Foundation grants DBI-0321467 and DBI-0820619, National Institutes of Health grant GM-58816, U.S. Department of Agriculture Hatch grant WIS04772 and research funds provided by the USDA-ARS to M.D.M.

References

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**, 783–796.

Cronk, Q. C. (2001). Plant evolution and development in a post-genomic context. *Nature Reviews Genetics* **2**, 607–619.

De Bodt, S., Raes, J., Van de Peer, Y. & Theissen, G. (2003). And then there were many: MADS goes genomic. *Trends in Plant Science* **8**, 475–483.

Doebely, J. (2006). Unfallen grains: How ancient farmers turned weeds into crops. *Science* **312**, 1318–1319.

Doebley, J. & Lukens, L. (1998). Transcriptional regulators and the evolution of plant form. *Plant Cell* **10**, 1075–1082.

Doebley, J., Stec, A. & Hubbard, L. (1997). The evolution of apical dominance in maize. *Nature* **386**, 485–488.

Egea-Cortines, M., Saedler, H. & Sommer, H. (1999). Ternary complex formation between the MADS-box proteins SQUAMOSA, DEFICIENS and GLOBOSA is involved in the control of floral architecture in *Antirrhinum majus*. *EMBO Journal* **18**, 5370–5379.

Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L. & Gaut, B. S. (1998). Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences of the USA* **95**, 4441–4446.

Fan, H. Y., Hu, Y., Tudor, M. & Ma, H. (1997). Specific interactions between the K domains of AG and AGLs, members of the MADS domain family of DNA binding proteins. *Plant Journal* **12**, 999–1010.

Fay, J. C. & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413.

Gu, Q., Ferrándiz, C., Yanofsky, M. F. & Martienssen, R. (1998). The *FRUITFULL* MADS-box gene mediates cell differentiation during *Arabidopsis* fruit development. *Development* **125**, 1509–1517.

He, C., Münster, T. & Saedler, H. (2004). On the origin of floral morphological novelties. *FEBS Letters* **567**, 147–151.

Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S. & Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. *American Journal of Human Genetics* **72**, 1527–1535.

Hilton, H. & Gaut, B. S. (1998). Speciation and domestication in maize and its wild relatives: evidence from the *globulin-1* gene. *Genetics* **150**, 863–872.

Honma, T. & Goto, K. (2001). Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. *Nature* **409**, 525–529.

Hudson, R. R. & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Hudson, R. R., Kreitman, M. & Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Immink, R. G., Ferrario, S., Busscher-Lange, J., Kooiker, M., Busscher, M. & Angenent, G. C. (2003). Analysis of the petunia MADS-box transcription factor family. *Molecular Genetics and Genomics* **268**, 598–606.

Innan, H. & Kim, Y. (2004). Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences of the USA* **101**, 10667–10672.

Kaufmann, K., Melzer, R. & Theissen, G. (2005). MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. *Gene* **347**, 183–198.

Kim, M., McCormick, S., Timmermans, M. & Sinha, N. (2003). The expression domain of PHANTASTICA determines leaflet placement in compound leaves. *Nature* **424**, 438–443.

Litt, A. & Irish, V. F. (2003). Duplication and diversification in the APETALA1/FRUITFULL floral homeotic gene lineage: implications for the evolution of floral development. *Genetics* **165**, 821–833.

Liu, K., Goodman, M. M., Muse, S., Smith, J. S. C., Buckler, E. S. & Doebley, J. (2003). Genetic structure diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* **165**, 2117–2128.

Moore, R. C., Grant, S. R. & Purugganan, M. D. (2005). Molecular population genetics of redundant floral-regulatory genes in *Arabidopsis thaliana*. *Molecular Biology and Evolution* **22**, 91–103.

Münster, T., Deleu, W., Wingen, L. U., Ouzunova, M., Cacharron, J., Faigl, W., Werth, S., Kim, J. T. T., Saedler, H. & Theissen, G. (2002). Maize MADS-box genes galore. *Maydica* **47**, 287–301.

Ng, M. & Yanofsky, M. F. (2001). Function and evolution of the plant MADS-box gene family. *Nature Reviews Genetics* **2**, 186–195.

Pressoir, G. & Berthaud, J. (2004). Population structure and strong divergent selection shape phenotypic diversification in maize landraces. *Heredity* **92**, 95–101.

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189.

Przeworski, M. (2003). Estimating the time since the fixation of a beneficial allele. *Genetics* **164**, 1667–1676.

Purugganan, M. D. (1997). The MADS-box floral homeotic gene lineages predate the origin of seed plants: phylogenetic and molecular clock estimates. *Journal of Molecular Evolution* **45**, 392–396.

Purugganan, M. D., Boyles, A. L. & Suddith, J. I. (2000). Variation and selection at the *CAULIFLO*WER floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*. *Genetics* **155**, 855–862.

Purugganan, M. D. & Suddith, J. I. (1998). Molecular population genetics of the Arabidopsis *CAULIFLOWER* regulatory gene: Nonneutral evolution and naturally occurring variation in floral homeotic function. *Proceedings of the National Academy of Sciences of the USA* **95**, 8130–8134.

Purugganan, M. D. & Suddith, J. I. (1999). Molecular population genetics of floral homeotic loci. Departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. *Genetics* **151**, 839–848.

Rambaut, A. (1996). Se-Al: Sequence Alignment Editor. Available from: http://tree.bio.ed.ac.uk/software/seal.

Riechmann, J. L. & Meyerowitz, E. M. (1997*a*). Determination of floral organ identity by Arabidopsis MADS domain homeotic proteins AP1, AP3, PI, and AG is independent of their DNA-binding specificity. *Molecular Biology of the Cell* **8**, 1243–1259.

Riechmann, J. L. & Meyerowitz, E. M. (1997*b*). MADS domain proteins in plant development. *Biological Chemistry* **378**, 1079–1101.

Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497.

Saedler, H., Becker, A., Winter, K. U., Kirchner, C. & Theissen, G. (2001). MADS-box genes are involved in floral development and evolution. *Acta Biochimica Polonica* **48**, 351–358.

Sieburth, L. E., Running, M. P. & Meyerowitz, E. M. (1995). Genetic separation of third and fourth whorl functions of AGAMOUS. *Plant Cell* **7**, 1249–1258.

Smith, L. & King, G. (2000). The distribution of *BoCal-a* alleles in *Brassica oleracea* is consistent with a genetic model for curd development and domestication of the cauliflower. *Molecular Breeding* **6**, 603–613.

Stumpf, M. P. & McVean, G. A. (2003). Estimating recombination rates from population-genetic data. *Nature Reviews Genetics* **4**, 959–968.

Swofford, D. L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite population. *Genetics* **105**, 437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F. & Gaut, B. S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays L.*). *Proceedings of the National Academy of the Sciences of the USA* **98**, 9161–9166.

Tenaillon, M. I., U'Ren, J., Tenaillon, O. & Gaut, B. S. (2004). Selection versus demography: a multilocus investigation of the domestication process in maize. *Molecular Biology and Evolution* **21**, 1214–1225.

Theissen, G. (2001). Development of floral organ identity: stories from the MADS house. *Current Opinion in Plant Biology* **4**, 75–85.

Theissen, G., Becker, A., Di Rosa, A., Kanno, A., Kim, J. T., Münster, T., Winter, K. U. & Saedler, H. (2000). A

short history of MADS-box genes in plants. *Plant Molecular Biology* **42**, 115–149.

Theissen, G., Kim, J. T. & Saedler, H. (1996). Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. *Journal of Molecular Evolution* **43**, 484–516.

Theissen, G. & Saedler, H. (1999). The golden decade of molecular floral development (1990–1999): a cheerful obituary. *Developmental Genetics* **25**, 181–193.

Vandenbussche, M., Theissen, G., Van de Peer, Y. & Gerats, T. (2003). Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. *Nucleic Acids Research* **31**, 4401–4409.

Vigouroux, Y., McMullen, M., Hittinger, C. T., 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 the National Academy of Sciences of the USA* **99**, 9650–9655.

Vrebalov, J., Ruezinsky, D., Padmanabhan, V., White, R., Medrano, D., Drake, R., Schuch, W. & Giovannoni, J. (2002). A MADS-Box gene necessary for fruit ripening at the tomato *Ripening-Inhibitor* (*Rin*) locus. *Science* **296**, 343–346.

Wang, H., Nussbaum-Wagler, T., Li, B., Zhao, Q., Vigouroux, Y., Faller, M., Bomblies, K., Lukens, L. & Doebley, J. F. (2005). The origin of the naked grains of maize. *Nature* **436**, 714–719.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D. & Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314.

Wright, S. I. & Gaut, B. S. (2005). Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution* **22**, 506–519.

Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T. & Dubcovsky, J. (2003). Positional cloning of the wheat vernalization gene *VRN1*. *Proceedings of the National Academy of Sciences of the USA* **100**, 6263–6268.