Theor Appl Genet (2012) 125:897–907
DOI 10.1007/s00122-012-1881-z

ORIGINAL PAPER

Evidence of selection on fatty acid biosynthetic genes during the evolution of cultivated sunflower

Mark A. Chapman · John M. Burke

Received: 22 January 2012 / Accepted: 19 April 2012 / Published online: 12 May 2012

Abstract The identification of genes underlying the phenotypic transitions that took place during crop evolution, as well as the genomic extent of resultant selective sweeps, is of great interest to both evolutionary biologists and applied plant scientists. In this study, we report the results of a molecular evolutionary analysis of 11 genes that underlie fatty acid biosynthesis and metabolism in wild and cultivated sunflower (*Helianthus annuus*). Seven of these 11 genes showed evidence of selection at the nucleotide level, with 1 (*FAD7*) having experienced selection prior to domestication, 2 (*FAD2-3* and *FAD3*) having experienced selection during domestication, and 4 (*FAB1*, *FAD2-1*, *FAD6*, and *FATB*) having experienced selection during the subsequent period of improvement. Sequencing of a subset of these genes from an extended panel of sunflower cultivars revealed little additional variation, and an analysis of the genomic region surrounding one of these genes (*FAD2-1*) revealed the occurrence of an extensive selective sweep affecting a region spanning at least ca. 100 kb. Given that previous population genetic analyses have revealed a relatively rapid decay of linkage disequilibrium in sunflower, this finding indicates the occurrence of strong selection and a rapid sweep.

Communicated by M. Frisch.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1881-z) contains supplementary material, which is available to authorized users.

M. A. Chapman · J. M. Burke (✉)
Department of Plant Biology, University of Georgia,
Miller Plant Sciences Bldg., Athens, GA 30602, USA
e-mail: jmburke@uga.edu

M. A. Chapman
e-mail: mark.chapman@plants.ox.ac.uk

M. A. Chapman
Department of Plant Sciences, University of Oxford,
Oxford OX1 3RB, UK

Introduction

Going back at least as far as Darwin (1859, 1868), it has been argued that the evolution of crop plants provides a useful model for studying phenotypic evolution in response to strong directional selection (see also Heiser 1988; Pickersgill 2010; Zohary 2004). At a genetic level, such selection can also have a profound impact on patterns and levels of genetic diversity across the genome, with the end result typically being a dramatic loss of variation in genes targeted by selection, as well as at linked loci (e.g., Clark et al. 2004; Olsen et al. 2006; Palaisa et al. 2004). In crops, these so-called "selective sweeps" are typically superimposed on a genome-wide loss of genetic diversity resulting from the genetic bottleneck associated with domestication (e.g., Eyre-Walker et al. 1998; Wright et al. 2005). The genomic extent of these sweeps is largely determined by the strength of selection and local recombination rates (e.g., Braverman et al. 1995; Durrett and Schweinsberg 2004; Kaplan et al. 1989), with stronger selection and/or reduced recombination resulting in more extensive sweeps. While the theory underlying this process has been well understood for over 30 years (e.g., Maynard-Smith and Haigh 1974), only recently have detailed analyses of the extent of such sweeps been carried out.

In cereal crops such as rice, maize, and sorghum, it has been shown that selection at a single locus can result in sweeps spanning relatively large genomic regions. For example, in temperate japonica varieties of rice (*Oryza sativa* L.), selection for glutinous grains has resulted in a

sweep spanning ca. 250 kb around the *Waxy* (*Wx*) gene (Olsen et al. 2006). In maize (*Zea mays* spp. *mays*), selection for increased apical dominance resulted in a selective sweep reaching 60–90 kb upstream of the *tb1* gene (Clark et al. 2004), and selection favoring yellow kernel color, presumably due to its increased nutritional value, has resulted in a sweep spanning ca. 800 kb surrounding the *Y1* locus (Palaisa et al. 2004). Similarly, selection on one or more genes has produced a selective sweep spanning 1.1 Mb on maize chromosome 10 (Tian et al. 2009). Finally, in sorghum, Casa et al. (2006) documented an apparent selective sweep spanning 99 kb along chromosome 1, with patterns of diversity and haplotype structure suggesting that selection may have targeted a protein phosphatase gene in this region. Interestingly, in all three species, these selective sweeps were more extensive than might have been expected based on background levels of linkage disequilibrium (LD), suggesting the occurrence of strong selection and rapid sweeps (Kim and Nielsen 2004).

In the present study, we investigate the genomic effects of selection on the fatty acid (FA) biosynthetic pathway during the evolution of cultivated sunflower (*Helianthus annuus* L.). Cultivated sunflower, which is one of the world's most important oilseed crops as well as a major source of confectionery seeds and ornamental flowers, was domesticated from wild *H. annuus* in eastern North America over 4,000 years ago (Heiser 1978; Rieseberg and Seiler 1990). Although sunflower was originally domesticated as a source of edible seeds, pigments, and medicine (Heiser 1951), a subset of the cultivated sunflower gene pool has more recently (i.e., within the past 100 years) been subjected to intense selection for increased seed oil content and altered oil composition (Burke et al. 2005; Putt 1997). It thus seems likely that, at least in a certain segment of the sunflower gene pool, genes involved in FA biosynthesis have been targeted by selection. Most recently, lines with oil that is rich in oleic acid have been developed, and the causal mutation for this shift in oil composition has been identified (i.e., a tandem duplication of the oleate desaturase *FAD2-1* that results in a loss of expression; Schuppert et al. 2006).

The genetic basis of oil biosynthesis in plants is generally well known (e.g., Gibson et al. 1994; Harwood 2005; Ohlrogge and Jaworski 1997). The most common fatty acids have chain lengths of 16 or 18 with up to three double bonds, and the chemical structure of these fatty acids is often expressed as a numerical shorthand (e.g., 16:0, 18:1, etc.). Sunflower oil is naturally rich in linoleic acid (18:2) and, as noted above, a number of high oleic (18:1) lines have been developed. The balance of the sunflower fatty acid profile is largely composed of palmitic acid (16:0) and stearic acid (18:0), which typically combine to account for ca. 10–12 % of the fatty acids present (Burke et al. 2005; Schuppert et al. 2006). Five minor fatty acids (myristic, linolenic, arachidic, behenic, and lignoceric) collectively account for <2 % of the oil composition. In plant cells, palmitic acid (16:0) is the first FA produced during fatty acid synthesis, and is initially bound to an acyl carrier protein (ACP) within the chloroplast. While still in the plastid, the synthase *FAB1* can convert 16:0-ACP to 18:0-ACP and the desaturase *FAB2* can convert 18:0-ACP to 18:1-ACP (Shanklin and Cahoon 1998; Wallis and Browse 2002). Thioesterases (*FATA* and *FATB*) are responsible for releasing the fatty acids from the ACP, thereby allowing them to be exported. Further, desaturation can occur in the chloroplast by *FAD6*, *FAD7*, and/or *FAD8*, or in the endoplasmic reticulum (ER) by *FAD2* and/or *FAD3*. The primary lipid substrates in the plastid and the ER are monogalactosyldiacylglycerol (MGDG) and phosphatidylcholine, respectively (Wallis and Browse 2002).

Previous population genetic analyses have suggested that LD decays relatively rapidly—within several kilobase pairs—in cultivated sunflower (Liu and Burke 2006; Kolkman et al. 2007), though the extent of LD in genomic regions targeted by selection remains unknown. In a prior study, we utilized a genome scan approach to identify a number of candidates for genes that were under selection during sunflower domestication and improvement (Chapman et al. 2008). In the current investigation, we used a candidate gene approach, isolating 11 homologs of the FA genes described above, including three different copies of *FAD2*, and performing DNA sequencing and molecular evolutionary analyses aimed at identifying genes that experienced selection during sunflower evolution. Through the use of a stratified sampling strategy involving wild sunflower, primitive landraces, and improved cultivars, we were further able to investigate the timing of selection (i.e., initial domestication vs. subsequent improvement). For one of the genes that showed strong evidence of recent selection, we used sequence information from a bacterial artificial chromosome (BAC) clone to facilitate an extended investigation of the distribution of sequence variation in the surrounding genomic region and to determine the extent of the sweep.

Materials and methods

Locus selection and sequencing

Polymerase chain reaction (PCR) primers for 11 sunflower FA biosynthetic genes were designed using primer3 (Rozen and Skaletsky 2000) based on either published or unpublished sequences from Genbank or expressed sequence tags (ESTs) from *Helianthus* that showed orthology to *Arabidopsis* homologs of genes of interest (Table 1). These

Table 1 Loci sequenced from sunflower *Helianthus annuus*. Watterson's θ for the wild, primitive, and improved sunflower populations, and the level of significance of the ML-HKA tests for non-neutral evolution are given

Locus	Source[a]	Length (bp)	θ Wild	θ Primitive	θ Improved	θ Improved2[g]	HKA[h] wild	HKA primitive	HKA improved
c25	Chapman et al. (2008)	1011	0.0239	0.0150	0.0138				
c1111	Chapman et al. (2008)	530	0.0094	0.0006	0.0019				
c1351	Chapman et al. (2008)	499	0.0138	0.0190	0.0111				
c2016	Chapman et al. (2008)	474	0.0301	0.0189	0.0162				
c2307	Chapman et al. (2008)	444	0.0050	0.0055	0.0062				
c5369	Chapman et al. (2008)	389	0.0111	0.0097	0.0052				
c5456	Chapman et al. (2008)	440	0.0185	0.0101	0.0054				
MeanNEUT			**0.0160**	**0.0111**	**0.0086**				
FAB1	Schuppert et al. (unpublished); Blast searches against Compositae Genome Project EST database	456	0.0079	0.0093	0.0016	0.0011			*
FAB2[b]	Blast searches against Compositae Genome Project EST database	988[d]	0.0089	0.0047	0.0030				
FAD2-1[c]	Schuppert et al. (2006); Martinez-Rivas et al. (2001)	1041	0.0156	0.0054	0.0000	0.0000			**
FAD2-2[c]	Schuppert et al. (unpublished); Martinez-Rivas et al. (2001)	986	0.0171	0.0081	0.0067				
FAD2-3[c]	Schuppert et al. (unpublished); Martinez-Rivas et al. (2001)	1048	0.0029	0.0010	0.0000			*	**
FAD3	Blast searches against Compositae Genome Project EST database	471	0.0059	0.0008	0.0000			*	**
FAD6	Blast searches against Compositae Genome Project EST database	705	0.0114	0.0096	0.0000				**
FAD7	Blast searches against Compositae Genome Project EST database	1132	0.0040	0.0031	0.0041		*	*	*
FAD8	Blast searches against Compositae Genome Project EST database	640[e]	0.0279	0.0377	0.0249				
FATA	Schuppert et al. (unpublished)	1030[f]	0.0165	0.0184	0.0108				
FATB	Schuppert et al. (unpublished)	1251	0.0092	0.0074	0.0000	0.0050			*
MeanFA			**0.0116**	**0.0096**	**0.0046**				

[a] Sequence information was obtained from the following references

[b] Two copies of FAB2 exist in sunflower; however, we were unable to reliably amplify one of these

[c] Three paralogs of FAD2 exist in sunflower

[d] Excluding the intervening intron for which alignment was unreliable

[e] Excluding a 844-bp indel and region for which alignment was unreliable

[f] Excluding a 374-bp insertion only present in the outgroup

[g] For three loci, extended sequencing of up to 12 more improved cultivars was carried out

[h] Results of ML-HKA test for selection: * significant at the 0.05 level; ** significant at the 0.01 level

genes were PCR-amplified from genomic DNA and sequenced from a panel of wild *H. annuus* individuals, primitive landraces, and improved lines, plus *H. petiolaris*, which was included as an outgroup (Supplemental Table 1) using standard protocols (e.g., Chapman et al. 2008). The wild, primitive, and improved classes are hereafter referred to as populations. The improved population included both confectionery and oilseed lines; as such, we did not restrict our analysis to a specific market class within the improved sunflower gene pool. In addition, seven genes that previously showed no evidence of selection (Chapman et al. 2008; Table 1) were included as putatively neutral control genes for the evolutionary analyses. Each PCR contained 10 ng of template DNA, 30 mM Tricine pH 8.4-KOH, 50 mM KCl, 2 mM $MgCl_2$, 100 mM of each dNTP, 0.1 mM of each primer, and 1 unit of *Taq* DNA polymerase. Primer sequences are listed in Supplemental Table 2, and PCR cycling conditions, preparation of templates for sequencing, and cycle sequencing parameters followed previously established protocols (Chapman et al. 2008). Following examination of the sequencing electropherograms, PCR products that showed evidence of length heterozygosity due to insertions/deletions (indels) were reamplified, gel-purified using a QIAquick Gel Extraction Kit (Qiagen, Valencia, CA), and cloned into pGEM-T vectors (Promega, Madison, WI). Vectors were transformed into competent *Escherichia coli*, grown overnight, and PCR-screened for the presence of an insert. Four positive colonies were then sequenced for each gene/individual as detailed above, except that vector primers (T7 and SP6) were used.

To explore the extent to which our results apply to the cultivated sunflower gene pool as a whole, three of the genes that showed evidence of selection (see "Results") were sequenced from an expanded diversity panel composed of 12 additional improved sunflower lines (Supplemental Table 1). These lines, known as the 'Core 12,' were selected from an initial collection of >400 cultivars to capture as much of the allelic diversity present in the cultivated sunflower gene pool as possible (Mandel et al. 2011).

Tests for selection

Watterson's θ (a measure of genetic diversity; Watterson 1975) was estimated using DnaSP ver. 5.10 (Librado and Rozas 2009) for each of the FA genes separately in the wild, primitive, and improved populations. Selection tests were then carried out as described previously (Chapman et al. 2008) using the same seven putatively neutral genes and the HKA (Hudson et al. 1987) test in a maximum likelihood framework (Wright and Charlesworth 2004). Briefly, the data were partitioned into wild versus outgroup, primitive versus outgroup, and improved versus outgroup comparisons. For each comparison, each of the FA biosynthetic genes was tested against the seven neutral genes. To do this, a strictly neutral model was first run, followed by a model in which the one candidate gene was deemed under selection while the other seven genes were assumed to be evolving neutrally. All models were evaluated based on 100,000 replicates, and the level of significance was calculated as twice the difference in likelihoods of the two models (i.e., neutral vs. selected) for a given comparison and compared to a χ^2 distribution with 1 degree of freedom.

Analysis of genetic polymorphism surrounding a selection candidate

The region surrounding one of the genes confirmed as having been under selection (*FAD2-1*; see "Results") was analyzed for DNA polymorphism. A BAC containing this gene was isolated using radioactively labeled overgo probes from the sunflower BAC library (available from Clemson University Genomics Institute; http://www.genome.clemson.edu/). This BAC (P396I22; Genbank Number FJ269356) was sequenced at the Joint Genome Institute using a Sanger shotgun approach with automatic and manual finishing (Staton et al., submitted). The resulting BAC sequence was subjected to analysis using the programs LTR_Finder (Xu and Wang 2007) to identify putative transposons, and BLASTn (Altschul et al. 1997) and FGENESH (Salamov and Solovyev 2000) to identify putative genes by comparing to *Arabidopsis* and *Vitis* genome sequences, as well as to ESTs from members of the Asteraceae (Staton et al., submitted).

Three putative genes (in addition to *FAD2-1*) were identified in the BAC sequence (Table 2), and portions of these genes were sequenced from the same panel of wild, primitive, and improved sunflowers as above. Primer sequences are listed is Supplemental Table 2. A region ca. 2 kb downstream of the *FAD2-1* was included in this analysis, because variation had previously been reported (Schuppert et al. 2006). This region comprises two exons of the *U-box-like* gene, although it had not been previously annotated as a coding region (Schuppert et al. 2006). Taken together, these gene sequences allowed us to investigate sequence diversity across a ca. 48-kb region surrounding *FAD2-1*. To further extend our knowledge of sequence variation surrounding this gene, primers flanking two putatively non-coding, single-copy regions of the BAC located further from *FAD2-1* (Fig. 2) were designed and used to sequence the same panel of sunflowers (Table 2), thereby extending the size of total region analyzed to ca. 94 kb. Sequences from these six additional regions were all analyzed and tested for selection as outlined above for the FA genes.

Table 2 The seven regions in the BAC that were sequenced

Locus	Position[a]	Function	Length (bp)[b]	θ Wild	θ Primitive	θ Improved
F2bacL1	527–1092	Non-coding	566	0.1141	0.0025	0.0005
F2bacL2	26859–27395	Non-coding	537	0.0254	0.0166	0.0000
COG	49048–49713	COG5125-like protein	666	0.0351	0.0230	0.0000
FAD2-1	79111–80151	Fatty acid desaturase	1041	0.0156	0.0054	0.0000
U-box1	82205–83013	U-box domain-containing protein[c]	810	0.0184	0.0082	0.0041
U-box2	87329–86722	U-box domain-containing protein	608	0.0097	0.0016	0.0000
RPA1	94186–94523	Replication protein A1	338	0.0084	0.0038	0.0000

Portions of the four genes in the BAC were sequenced (including two regions of the U-box gene) as well as two non-coding regions. The length of the region sequenced as well as values of Watterson's θ in the wild, primitive, and improved populations are given

[a] Based on BAC sequence in Genbank (accession number FJ269356)

[b] Including gaps

[c] This was annotated as an intergenic region in the Genbank file; however, we predict it to be a portion of coding region of the U-box domain-containing protein (see text for details)

Phylogenetic analysis of sunflower domestication

The sequences of (1) all 11 FA genes, and (2) all loci from throughout the BAC were concatenated for each of the sunflower individuals. The concatenations totaled ca. 9.5 and 4 kb per individual, respectively. Alleles within an individual were then collapsed and heterozygous bases were coded using standard IUPAC code. Indels were coded as nucleotides such that heterozygous individuals could be coded using IUPAC codes, though indels with more than two allelic states across the dataset (e.g., microsatellite-like regions) were excluded from the analysis. Maximum likelihood (ML) phylogenetic analysis was then carried out using PHYML (Guindon and Gascuel 2003) with default parameters, and the resulting trees were visualized in TreeView (Page 1996).

Results

Eleven homologs of nine genes involved in FA metabolism in *Arabidopsis* were identified and sequenced from sunflower, including three *FAD2* paralogs. All sequences are deposited in Genbank under accession numbers JQ974402-JQ974815. Table 1 lists basic information about each gene, and Fig. 1 shows the levels of polymorphism (Watterson's θ) for each gene/population along with the results of the tests for selection. Polymorphism was generally highest in the wild population, lower in the primitive population, and lowest in the improved population. On average, the FA biosynthetic genes exhibited somewhat lower levels of polymorphism in all three populations as compared to the neutral genes, though none of the differences were significant overall. Five of the FA biosynthetic genes (*FAD2-1*, *FAD2-3*, *FAD3*, *FAD6*, and *FATB*) were completely monomorphic in the improved population despite exhibiting variation in both the wild and primitive populations.

Of the 11 genes analyzed, 7 showed evidence of selection. *FAD7* exhibited a significant departure from neutrality ($0.01 < P < 0.05$) in all three tests (i.e., in the wild, primitive, and improved vs. outgroup comparisons); hence, this gene was likely targeted by natural selection prior to domestication. In contrast, *FAD2-3* and *FAD3* showed evidence of selection in both the primitive–outgroup and improved–outgroup comparisons, indicating that selection occurred during sunflower domestication (i.e., since the split between wild and primitive populations). Finally, *FAB1*, *FAD2-1*, *FAD6*, and *FATB* showed selection in the improved–outgroup comparison only; hence, selection in these cases must have occurred more recently (i.e., during the subsequent period of improvement).

Of the three genes that were subsequently sequenced in the expanded set of improved lines (i.e., the Core 12), *FAB1* exhibited two haplotypes in the initial sequencing set of improved lines, whereas both *FAD2-1* and *FATB* were monomorphic in the original lines. In the expanded sequencing panel, an additional haplotype was discovered in both *FAB1* (present in three individuals) and *FATB* (present in one individual). For *FAD2-1*, no additional haplotypes were found. Thus, *FAD2-1* appears to be completely devoid of variation in the improved sunflower gene pool (at least in the portion sequenced) despite this gene exhibiting comparable levels of variation in the wild gene pool as compared to the neutral genes (Fig. 1; Table 1).

Because *FAD2-1* showed a complete loss of sequence variation, we analyzed the surrounding genomic region for polymorphism. The BAC that served as the basis for this analysis (Genbank FJ269356) was 107,161 bp in length and

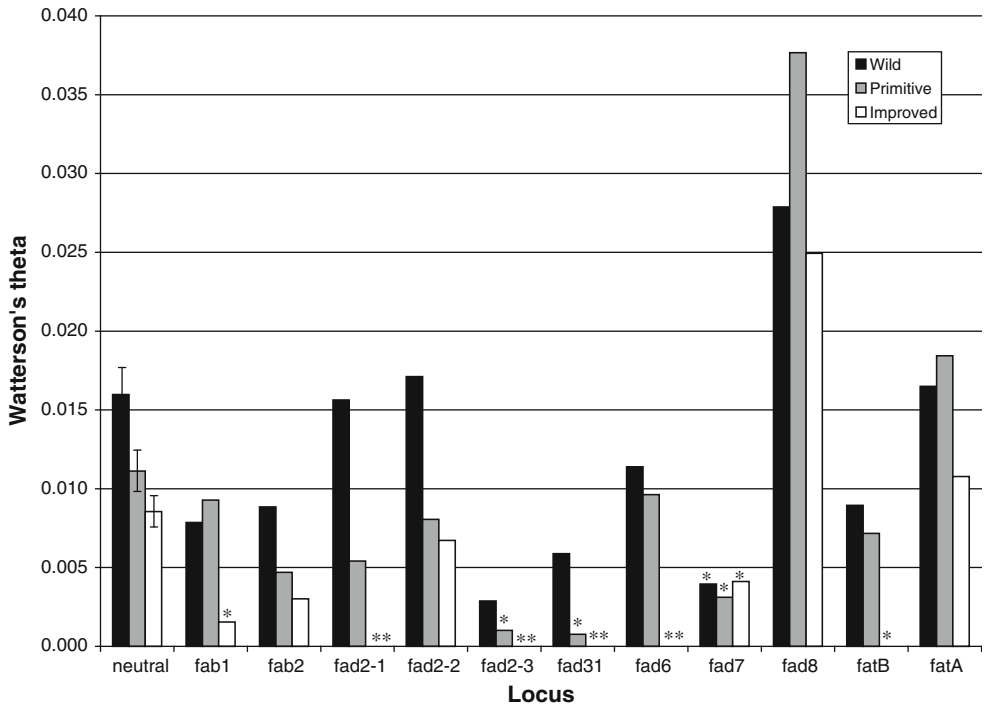

Fig. 1 Sequence polymorphism (measured as Watterson's θ) for seven neutral genes (with *error bars*) and the 11 fatty acid loci. Significant evidence for selection based on HKA tests is indicated above the relevant populations/loci. $*P \leq 0.05$; $**P \leq 0.01$

contained three genes in addition to *FAD2-1*. These additional genes code for a COG5125-like protein, a U-box domain-containing protein, and an RPA1-like protein. Six regions, including portions of these three genes (two portions of the *U-box* gene) and two non-coding, single-copy sequences, were sequenced in the original panel of wild, primitive, and improved sunflowers. Levels of DNA polymorphism for each of these loci are listed in Table 2. These sequences were almost completely monomorphic in the improved population, whereas polymorphism was similar between these BAC-derived loci and the neutral loci for both the primitive and wild populations (Tables 1, 2). The sole exception was RHA280, which had one SNP at *F2bacL1* and also harbored a divergent haplotype in the *U-box1* region. Tests of selection for all of these BAC-derived loci were, like *FAD2-1*, significant for the improved–outgroup comparison. The second U-box-like region also showed evidence of selection in the primitive–outgroup comparison. As such, it appears that selection in this region has resulted in a selective sweep spanning a minimum of ca. 94 kb.

The maximum likelihood phylogenetic analysis of the concatenated fatty acid genes (9,589 bp; Fig. 3a) revealed that primitive and improved sunflower form a monophyletic group with the exception of the Hopi landrace. For the combined BAC loci (i.e., the ca. 100 kb surrounding *FAD2-1*; total 4,037 bp; Fig. 3b), ML analysis revealed that primitive and improved sunflower form a monophyletic group with the exception of the Havasupai sunflower.

Discussion

Theory predicts that strong directional selection will reduce genetic diversity at both the target locus and at linked, neutral loci (e.g., Kaplan et al. 1989; Maynard-Smith and Haigh 1974); reviewed in (Nielsen 2005). While population genetic bottlenecks, such as those that occur during domestication (e.g., Eyre-Walker et al. 1998), are known to produce a genome-wide reduction in diversity, selected loci are expected to exhibit an even greater reduction in diversity (Burger et al. 2008). The locus-specific nature of this diversity loss makes it possible to identify selectively important regions of the genome (Burke et al. 2007). While we previously used a genome scan approach to identify loci that showed evidence of selection during sunflower domestication and/or improvement (Chapman et al. 2008), the present study reflects a more targeted approach in which we investigated the genes comprising an important biochemical pathway underlying a key aspect of the evolution of cultivated sunflower.

Reduced genetic diversity and evidence for selection in genes associated with FA metabolism

As has previously been found in sunflower (Chapman et al. 2008; Liu and Burke 2006; Tang and Knapp 2003) and other crops (e.g., Caicedo et al. 2007; Casa et al. 2005; Vigouroux et al. 2002; Zhu et al. 2007), genetic diversity

was highest in the wild population and lowest in the improved cultivars (Fig. 1), presumably due to the occurrence of one or more genetic bottlenecks associated with the domestication and subsequent improvement of sunflower. Moreover, for the three genes that were re-sequenced from the expanded panel of cultivars, very little additional variation was found. As compared to the neutral genes, the 11 FA genes had slightly lower genetic diversity in all three populations (Table 1), and just over half of the FA biosynthetic genes (7 of 11) showed evidence of selection. While we did not sequence the full coding region of each of these loci, the presence of polymorphism in the wild (and primitive) sunflower lines indicates that these are not simply invariant genomic regions; rather, these regions exhibit unexpectedly low levels of diversity, presumably due to the effects of selection.

Because our sampling of sunflowers included wild, primitive, and improved individuals, we can further investigate the timing of selection. One gene, *FAD7*, showed a significant departure from neutrality in all three populations; hence, we can only say that it experienced selection sometime since the split between *H. annuus* and *H. petiolaris*. For the other six genes, however, we can infer the timing based on the results of the selection tests. Two genes (*FAD2-3* and *FAD3*) showed significant evidence for selection in both the primitive–outgroup and improved–outgroup comparisons, and thus appear to have experienced selection during the initial domestication of sunflower. The remaining four (*FAB1*, *FAD2-1*, *FAD6*, and *FATB*) were significant in the improved–outgroup comparison only, and thus appear to have experienced selection during the post-domestication era, presumably during the transformation of sunflower from a primitive domesticated species into a commercially viable crop plant.

The two domestication genes (*FAD2-3* and *FAD3*) both code for desaturases that act in the endoplasmic reticulum, and which are thought to be active in a wide variety of tissues (Martinez-Rivas et al. 2001). Interestingly, these two genes act consecutively, converting oleic acid (18:1) to linoleic acid (18:2) and linoleic acid (18:2) to linolenic acid (18:3), respectively. Because sunflower was initially domesticated for its edible seeds (Heiser 1951; Putt 1997), as opposed to having been explicitly selected for oil content and/or composition, the selection acting on these genes may have been related to palatability. Alternatively, the degree of oil saturation/desaturation is known to influence the conditions under which seeds will germinate (Linder 2000), so it is may be that one or both of these genes experienced selection during domestication due to possible indirect effects on germinability.

The molecular signature of selection during sunflower improvement is evident for four genes. These genes are *FAB1*, encoding a synthase that converts 16:0-ACP to 18:0-ACP, *FATB* encoding a thioesterase responsible for releasing FAs from ACP allowing them to be exported, and two desaturase genes (*FAD2-1* and *FAD6*) that convert oleic (18:1) to linoleic (18:2) acid. *FAB1* is important in controlling the amount of 18 carbon (18C) FAs relative to 16C FAs (e.g., Pidkowich et al. 2007); as such, selection on *FAB1* may have played a role in increasing the total amount of 18:0-ACP available as a substrate for the production of oleic and linoleic acid. Similarly, differential regulation of *FATB*, a gene responsible for freeing up fatty acids for export from the plastids, and which is known to be expressed during sunflower seed formation (Bonaventure et al. 2003; Martinez-Force et al. 2000), could increase the total amount of FA substrate in the seeds. Given that the improved population contained both oilseed and confectionery lines, it appears that these selective events were not limited to the oilseed segment of the gene pool. Rather, these genes may have experienced selection prior to the divergence between the modern oilseed and confectionery types. This conclusion is underscored by the results from the sequencing of the Core 12, which is likewise composed of a mix of oilseed and confectionery lines.

It is interesting that *FAD2-1* and *FAD6* both encode oleoyl-specific desaturases, but the former is expressed in the endoplasmic reticulum (ER) in a seed-specific manner (Martinez-Rivas et al. 2001), whereas the latter is expressed solely in the chloroplast and in a wider variety of tissues (Browse et al. 1989). Given the known effect of this gene on elevated oleic acid content in high oleic lines (Schuppert et al. 2006), evidence for selection on *FAD2-1* is perfectly understandable. It is less clear why the analogous plastidial desaturase (*FAD6*) would have been under selection, though it has been suggested that flux through the plastid and ER FA modification pathways can compensate for each other and that, although the ER pathway is the major contributor to seed FAs, some flux through the plastid pathway also plays a role (Browse and Somerville 1991). Therefore, selection on the oleic/linoleic step in the sunflower oil biosynthetic pathway may have resulted in parallel selection on both *FAD2-1* and *FAD6*, even though expression of the latter is primarily restricted to the chloroplast.

The genomic extent of the *FAD2-1* selective sweep

Our investigation of patterns of polymorphism in the region surrounding *FAD2-1* revealed evidence of an extensive selective sweep—far more extensive than expected based on what was previously known about the structure of LD across the sunflower genome (Liu and Burke 2006; Kolkman et al. 2007). Very little DNA polymorphism was detected in the improved population in a nearly 100-kb genomic window surrounding this gene. In contrast, the wild and primitive

populations exhibited levels of variation across this region similar to those exhibited by the neutral loci (Tables 1, 2), and the tests of selection in the improved–outgroup comparison were significant across this entire region. As such, we can conclude that the sweep targeting this region spanned at least 94 kb (Fig. 2), though the full size of this swept region remains unknown due to a lack of genomic resources for exploring diversity beyond the boundaries of the BAC. With the forthcoming sequence of the sunflower genome (Kane et al. 2011), however, it should soon be possible to explore such issues in much greater detail. Given this finding of dramatically elevated LD in selectively important genomic regions, it is important to note that candidate genes showing evidence of selection may not have been targeted by selection themselves; rather, they may have hitchhiked along with the actual targets of selection. In the case of *FAD2-1*, our conclusion are backed up by functional evidence that this gene plays an important role in determining seed oil composition (Schuppert et al. 2006) and the fact that the other genes in the swept region have no known role in FA biosynthesis. In other cases, it remains possible that selection at a nearby gene has produced the observed pattern.

Five of the six improved sunflower individuals shared the same haplotype throughout the entire region. The sixth accession (RHA280, a confectionery cultivar) contained a single SNP in locus F2bacL1, the most distant locus relative to *FAD2-1*, and a single SNP and a 568-bp deletion (relative to the other improved sunflowers) in the *U-box1* locus, ca. 2 kb downstream of *FAD2-1* (Fig. 2). This deletion spans all of exon 8, such that the 'short' allele (present in RHA280) misses an entire exon from the coding region. Despite this anomaly, the tests for selection across this entire 100-kb region are significant. Interestingly, the deleted region is bracketed by a pair of inverted repeats with 90 % homology over 112 bp. While inverted repeats are a hallmark of Class II transposons (Kumar and Bennetzen 1999), neither these repeats nor the intervening sequence show homology to known transposons. Similar inverted repeats can, however, be found in other sunflower genes (e.g., ZVG77, a putative β-amylase and ZVG32, a putative NAD-dependent sorbitol dehydrogenase; Kolkman et al. 2007).

It is noteworthy that one of the Native American landraces (Seneca) and all of the wild sunflowers surveyed herein exhibited the 'short' allele of *U-box1*, whereas the remaining five primitive lines shared the 'long' allele with the five improved cultivars, not including RHA280. A previous PCR screen of 16 wild sunflowers (Schuppert et al. 2006) revealed that the long allele was present at low frequency in the wild and was present in all 16 oilseed cultivars tested, suggesting that fixation of the long allele occurred during the evolution of cultivated sunflower. We note that the *FAD2-1* coding region is identical in sequence across all improved lines (i.e., RHA280 is identical to the other cultivars despite this difference downstream of the *FAD2-1* coding region). As such, it may be that exon 8 of the *U-box* locus was independently lost in RHA280 (or a related line) subsequent to the selective sweep affecting that region.

The origin of cultivated sunflower

It has been suggested that the analysis of domestication (and improvement) genes in the wild progenitors of crops

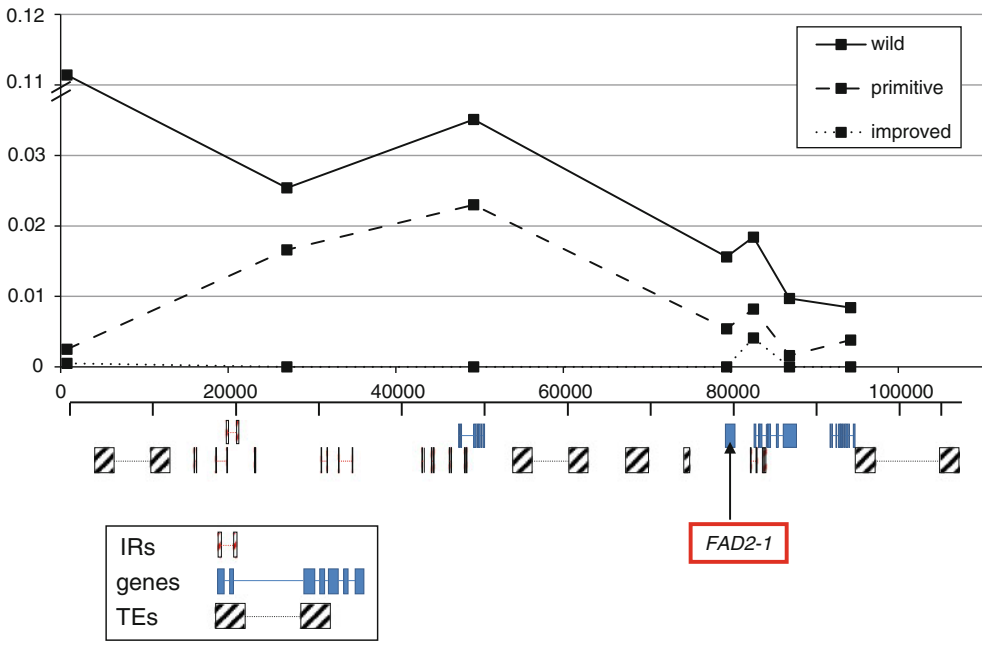

Fig. 2 Sequence polymorphism (measured as Watterson's θ) for seven loci across a ca. 100-kb window surrounding *FAD2-1*. Below the graph are indicated the positions of genes, inverted repeats (IRs), and putative transposable elements (TEs)

could be the best way to discern where crops originated (Gross and Olsen 2010). Although our sampling of wild sunflower was limited to only eight individuals, we did include wild samples from throughout the geographic range. In the phylogenetic analysis of both the concatenated fatty acid genes and the BAC loci from the swept *FAD2-1* region (Fig. 3), the bulk of the primitive and improved sunflower lines formed a monophyletic group, as one might expect based on prior evidence that cultivated sunflower is a product of a single domestication (Blackman et al. 2011; Harter et al. 2004; Wills and Burke 2006). The two exceptions were the Hopi landrace, which fell outside the primary group in the analysis of the concatenated FA genes, and the Havasupai landrace, which fell outside the primary group in the analysis of the BAC-derived sequences surrounding *FAD2-1*. Notably, the Hopi and Havasupai sunflowers have previously been shown to be quite divergent from the balance of the cultivated sunflower gene pool (e.g., Chapman et al. 2008; Harter et al. 2004; Tang and Knapp 2003), and these lines may have not been major contributors to the modern sunflower gene pool. The analyses presented here also add to the evidence that sunflower was domesticated in east-central North America (Harter et al. 2004; Heiser 1978), as the most closely related wild populations came from this general region.

Conclusions

As noted above, prior analyses of the extent of selective sweeps in cereal crops have revealed the occurrence of sweeps affecting large genomic regions in the order of 100–1,000 kb (Casa et al. 2006; Clark et al. 2004; Olsen et al. 2006; Palaisa et al. 2004; Tian et al. 2009). Our results extend this general pattern of extensive selective sweeps, even in species with otherwise relatively low levels of background linkage disequilibrium (Liu and Burke 2006; Tanksley and McCouch 1997; Yamasaki et al. 2005) to non-cereal crops.

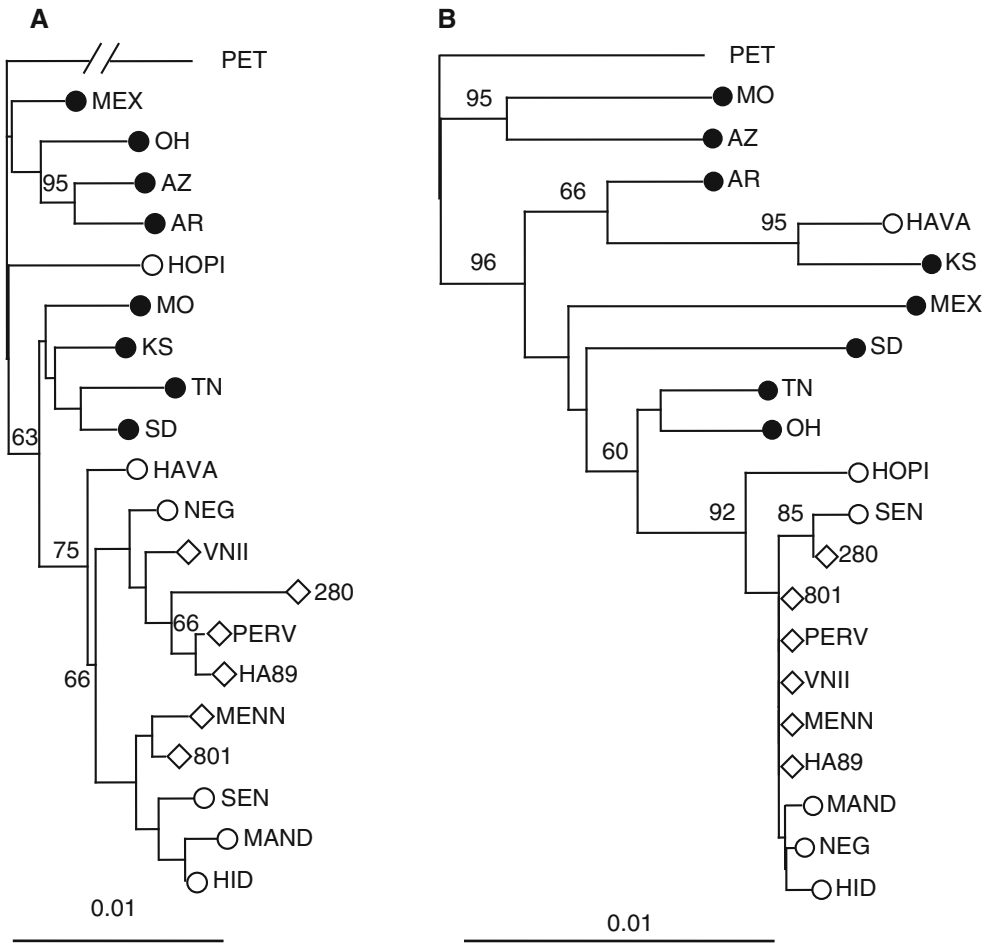

Fig. 3 Maximum likelihood phylogenetic trees for the wild (*closed circles*), primitive (*open circles*), and improved (*open diamonds*) sunflower individuals, rooted with *H. petiolaris* (PET). **a** Based on concatenation of fatty acid genes, **b** based on concatenation of loci throughout the BAC. Individual acronyms are given in Supplementary Table 1. Bootstrap values >50 % are indicated

The dramatic reduction in diversity in and around genes targeted by selection also has important implications for ongoing efforts aimed at the continued improvement of sunflower (Tanksley and McCouch 1997; Yamasaki et al. 2005). Response to selection on oil-related traits in sunflower is likely to be hampered by the absence of segregating variation at a number of the loci analyzed here. Five of the 11 candidate genes were completely devoid of variation across the improved lines, and even when a larger number of lines were sequenced, little additional variation was uncovered. Moreover, 9 of the 11 loci had less genetic variation in the primitive lines than the average neutral gene. These findings suggest that future efforts aimed at the continued improvement of sunflower as an oilseed crop will require the exploitation of wild germplasm as a source of novel alleles.

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Blackman BK, Scascitelli M, Kane NC, Luton HH, Rasmussen DA, Bye RA, Lentz DL, Rieseberg LH (2011) Sunflower domestication alleles support single domestication center in eastern North America. Proc Nat Acad Sci USA 108:14360–14365

Bonaventure G, Salas JJ, Pollard MR, Ohlrogge JB (2003) Disruption of the FATB gene in *Arabidopsis* demonstrates an essential role of saturated fatty acids in plant growth. Plant Cell 15:1020–1033

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency-spectrum of DNA polymorphisms. Genetics 140:783–796

Browse J, Kunst L, Anderson S, Hugly S, Somerville C (1989) A mutant of *Arabidopsis* deficient in the chloroplast 16-1/18-1 desaturase. Plant Physiol 90:522–529

Browse J, Somerville C (1991) Glycerolipid synthesis—biochemistry and regulation. Ann Rev Plant Physiol Plant Mol Biol 42:467–506

Burger JC, Chapman MA, Burke JM (2008) Molecular insights into the evolution of crop plants. Am J Bot 95:113–122

Burke JM, Burger JC, Chapman MA (2007) Crop evolution: from genetics to genomics. Curr Opin Genet Dev 17:525–532

Burke JM, Knapp SJ, Rieseberg LH (2005) Genetic consequences of selection during the evolution of cultivated sunflower. Genetics 171:1933–1940

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR, Bustamante CD, Purugganan MD (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet 3:1745–1756

Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, Aquadro CF, Kresovich S (2005) Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. Theor Appl Genet 111:23–30

Casa AM, Mitchell SE, Jensen JD, Hamblin MT, Paterson AH, Aguadro CF, Kresovich S (2006) Evidence for a selective sweep on chromosome 1 of cultivated sorghum. Crop Sci 46:S27–S40

Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ, Burke JM (2008) A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). Plant Cell 20:2931–2945

Clark RM, Linton E, Messing J, Doebley JF (2004) Pattern of diversity in the genomic region near the maize domestication gene *tb1*. Proc Nat Acad Sci USA 101:700–707

Darwin C (1859) On the origin of species by means of natural selection. John Murray, London

Darwin CR (1868) The Variation of Animals and Plants under Domestication. John Murray, London

Durrett R, Schweinsberg J (2004) Approximating selective sweeps. Theor Popul Biol 66:129–138

Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS (1998) Investigation of the bottleneck leading to the domestication of maize. Proc Nat Acad Sci USA 95:4441–4446

Gibson S, Falcone DL, Browse J, Somerville C (1994) Use of transgenic plants and mutants to study the regulation and function of lipid composition. Plant Cell Environ 17:627–637

Gross BL, Olsen KM (2010) Genetic perspectives on crop domestication. Trends Plant Sci 15:529–537

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Harter AV, Gardner KA, Falush D, Lentz DL, Bye RA, Rieseberg LH (2004) Origin of extant domesticated sunflowers in eastern North America. Nature 430:201–205

Harwood JL (2005) Fatty acid biosynthesis. In: Murphy DJ (ed) Plant lipids: biology, utilization and manipulation. Blackwell Publishing, Oxford

Heiser CB Jr (1951) The sunflower among North American Indians. Proc Am Philos Soc 95:432–448

Heiser CB Jr (1978) Taxonomy of *Helianthus* and origin of domesticated sunflower. In: Carter JF (ed) Sunflower science and technology. American Society of Agronomy, Madison, pp 31–53

Heiser CB Jr (1988) Aspects of unconscious selection and the evolution of domesticated plants. Euphytica 37:77–81

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

Kane NC, Gill N, King MG, Bowers JE, Berges H, Gouzy J, Bachlava E, Langlade NB, Lai Z, Stewart M, Burke JM, Vincourt P, Knapp SJ, Rieseberg LH (2011) Progress towards a reference genome for sunflower. Botany-Botanique 89:429–437

Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. Genetics 123:887–899

Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513–1524

Kolkman JM, Berry ST, Leon AJ, Slabaugh MB, Tang S, Gao WX, Shintani DK, Burke JM, Knapp SJ (2007) Single nucleotide polymorphisms and linkage disequilibrium in sunflower. Genetics 177:457–468

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33:479–532

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452

Linder CR (2000) Adaptive evolution of seed oils in plants: accounting for the biogeographic distribution of saturated and unsaturated fatty acids in seed oils. Am Nat 156:442–458

Liu AZ, Burke JM (2006) Patterns of nucleotide diversity in wild and cultivated sunflower. Genetics 173:321–330

Mandel JR, Dechaine JM, Marek LF, Burke JM (2011) Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus* L. Theor Appl Genet 123:693–704

Martinez-Force E, Cantisan S, Serrano-Vega MJ, Garces R (2000) Acyl–acyl carrier protein thioesterase activity from sunflower (*Helianthus annuus* L.) seeds. Planta 211:673–678

Martinez-Rivas JM, Sperling P, Luhs W, Heinz E (2001) Spatial and temporal regulation of three different microsomal oleate desaturase genes (FAD2) from normal-type and high-oleic varieties of sunflower (*Helianthus annuus* L.). Mol Breed 8:159–168

Maynard-Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. Genet Res 23:23–35

Nielsen R (2005) Molecular signatures of natural selection. Ann Rev Genet, pp 197–218

Ohlrogge JB, Jaworski JG (1997) Regulation of fatty acid synthesis. Ann Rev Plant Physiol Plant Mol Biol 48:109–136

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

Page RDM (1996) TREEVIEW: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12:357–358

Palaisa K, Morgante M, Tingey S, Rafalski A (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. Proc Nat Acad Sci USA 101:9885–9890

Pickersgill B (2010) Domestication of plants revisited - Darwin to the present day. Bot J Linn Soc 162:S37–S46

Pidkowich MS, Nguyen HT, Heilmann I, Ischebeck T, Shanklin J (2007) Modulating seed beta-ketoacyl-acyl carrier protein synthase II level converts the composition of a temperate seed oil to that of a palm-like tropical oil. Proc Nat Acad Sci USA 104:4742–4747

Putt ED (1997) Early history of sunflower. In: Schneiter AA (ed) Sunflower Technology and Production. American Society of Agronomy, Madison

Rieseberg LH, Seiler GJ (1990) Molecular evidence and the origin and development of the domesticated sunflower (*Helianthus annuus*, Asteraceae). Econ Bot 44(Supplement 3):79–91

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology, http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi Humana Press, Totowa, NJ, pp 365-386

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 10:516–522

Schuppert GF, Tang SX, Slabaugh MB, Knapp SJ (2006) The sunflower high-oleic mutant Ol carries variable tandem repeats of FAD2-1, a seed-specific oleoyl-phosphatidyl choline desaturase. Mol Breeding 17:241–256

Shanklin J, Cahoon EB (1998) Desaturation and related modifications of fatty acids. Annual Review of Plant Physiology and Plant Molecular Biology 49:611–641

Tang S, Knapp SJ (2003) Microsatellites uncover extraordinary diversity in native American land races and wild populations of cultivated sunflowers. Theoretical and Applied Genetics 106: 990–1003

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

Tian F, Stevens NM, Buckler ES (2009) Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proc Nat Acad Sci USA 106:9979–9986

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. Proc Nat Acad Sci USA 99:9650–9655

Wallis and Browse (2002) Mutants of *Arabidopsis* reveal many roles for membrane lipids. Prog Lipid Res 41:254–278

Watterson GA (1975) On the number of segregating sites in genetic models without recombination. Theor Popul Biol 7:256–276

Wills DM, Burke JM (2006) Chloroplast DNA variation confirms a single origin of domesticated sunflower (*Helianthus annuus* L.). J Hered 97:403–408

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

Wright SI, Charlesworth B (2004) The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168:1071–1076

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–W268

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. Plant Cell 17:2859–2872

Zhu QH, Zheng XM, Luo JC, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: Severe bottleneck during domestication of rice. Mol Biol Evol 24:875–888

Zohary D (2004) Unconscious selection and the evolution of domesticated plants. Econ Bot 58:5–10