SCIENCE CHINA
Life Sciences

• RESEARCH PAPERS •

June 2010 Vol.53 No.6: 690–700
doi: 10.1007/s11427-010-4007-3

A genome-wide survey of maize lipid-related genes: candidate genes mining, digital gene expression profiling and co-location with QTL for maize kernel oil

LI Lin[1], LI Hui[1], LI JiYing[1], XU ShuTu[1], YANG XiaoHong[1], LI JianSheng[1] & YAN JianBing[1,2 *]

[1]*National Maize Improvement Centre of China, China Agricultural University, Beijing 100193, China*;
[2]*International Maize and Wheat Improvement Centre* (*CIMMYT*)*, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico*

Received July 9, 2009; accepted November 20, 2009

Lipids play an important role in plants due to their abundance and their extensive participation in many metabolic processes. Genes involved in lipid metabolism have been extensively studied in *Arabidopsis* and other plant species. In this study, a total of 1003 maize lipid-related genes were cloned and annotated, including 42 genes with experimental validation, 732 genes with full-length cDNA and protein sequences in public databases and 229 newly cloned genes. Ninety-seven maize lipid-related genes with tissue-preferential expression were discovered by *in silico* gene expression profiling based on 1984483 maize Expressed Sequence Tags collected from 182 cDNA libraries. Meanwhile, 70 QTL clusters for maize kernel oil were identified, covering 34.5% of the maize genome. Fifty-nine (84%) QTL clusters co-located with at least one lipid-related gene, and the total number of these genes amounted to 147. Interestingly, thirteen genes with kernel-preferential expression profiles fell within QTL clusters for maize kernel oil content. All the maize lipid-related genes identified here may provide good targets for maize kernel oil QTL cloning and thus help us to better understand the molecular mechanism of maize kernel oil accumulation.

lipid metabolism, *in silico* gene cloning, *in silico* gene expression profiling, quantative trait locus(QTL), maize

Citation: Li L, Li H, Li J Y, *et al*. A genome-wide survey of maize lipid-related genes: candidate genes mining, digital gene expression profiling and co-location with QTL for maize kernel oil. Sci China Life Sci, 2010, 53: 690−700, doi: 10.1007/s11427-010-4115-0

Lipids, which refer to fatty acids (FA) and their natural ester derivatives, are common in plants and participate in a number of biological processes. In plants, lipids serve as the building blocks for structural membrane lipids (such as phospholipids, galactolipids, sphingolipids and sulfolipids) [1,2], the sinks for energy storage (such as TriAcylGlycerol, TAG) [3], the defensive substance (such as wax), the precursor of signal molecules in the wounding and pathogenic response pathways [4,5], and the necessary energy resources for seed germination [6]. Lipid metabolism is a dynamic equilibrium, for which the acly-CoA pool is central. Its ten main biological processes, including FA synthesis [7,8], TAG synthesis and storage [9], FA degradation [12–14], FA elongation and wax synthesis [10,11], phospholipid synthesis [15], lipid signaling [16–19], sphingolipid synthesis [20], galactolipid synthesis [21], sulfolipid synthesis [22] and oxylipin synthesis [23,24], take place in the cytosol, the plastids and the endoplasmic reticulum. Within the plastid, FA synthesis begins with a series of condensation reactions and elongation of acetyl-CoA, forming a pool consisting mostly of C16:0- and C18:0-Acyl-carrier proteins (ACP). Within the cytosol, the acyl-chains are modified through complex interactions of desaturation, elongation and phospholipid/acyl-CoA exchange processes. The dynamic equi-

*Corresponding author (email: yjianbing@cau.edu.cn)

librium system among these pathways ultimately controls the composition of FAs, the yield of TAGs and the rate of *de novo* FA synthesis. As the center of lipid metabolism, the acyl-CoA pool is a sensitive indicator reflecting the flux in acyl-chains [27].

The availability of complete genome sequences for humans and model plants [28,29] provides a good foundation for systematic metabolism research on a genome-wide scale [30,31]. Some metabolic databases have been established, such as Carbohydrate-Active enZYmes-CAZy [32], a database of 320 receptor kinase-like proteins [33] and the P450s family database (http://Arabidopsis-p450.biotec.uiuc.edu/). Beisson *et al*. [34] performed a genome-wide annotation of *Arabidopsis* lipid metabolism; more than 600 *Arabidopsis* lipid-related genes were annotated, which was not only valuable for *Arabidopsis* lipid research, but also provided a good example for research in other species. As one of the most important crops in the world, maize is a model species for genome research. The maize genome sequencing project has been completed (http://www.maizesequence.org/) and the information is stored in a public database. As of March 2009, there were 2018338 maize ESTs in GenBank (http://www.ncbi.nlm.nih.gov/sites/entrez?db=est, search keywords are "*Zea mays* [porgn: _txid4577]"). The number of high throughput genome sequences (HTGS) is 16587 (http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuccore, search keywords are "*Zea mays* [ORGN] AND HTG [KYWD] AND Washington University School of Medicine [ALL]"). Despite this, the research on maize lipid metabolism is still at an initial stage, only 42 maize lipid metabolism genes have been experimentally validated (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene, search keywords are "*Zea mays* [porgn] AND lipid AND (1988/1/1[PUBDATE]: 2009/3/1[PUBDATE])"), and many more related genes are still unknown. Based on comparative genomic analysis, we have cloned *in silico* and annotated as many maize lipid-related genes as possible, with the intention of: 1, annotating maize lipid-related genes with full length information in the public databases; 2, cloning maize lipid-related genes with no full length information; 3, establishing *in silico* gene expression profiles of these genes; and 4, identifying genes co-located with QTL for maize kernel oil content. The information generated here will be valuable for genetic research into the mechanism of maize kernel oil production.

1 Materials and methods

1.1 Collection of lipid-related genes in plants

A total of 635 lipid metabolism-related genes from *Arabidopsis* and 1 from rice were collected in this study. Of these genes, 600 came from the *Arabidopsis* database on lipid metabolism [34] and 36 came from published papers. The 636 genes encode 207 functional enzymes (here, "functional enzymes" refer to classes of genes whose protein products perform similar functions, irrespective of their subcellular locations), including 182 enzymes for substrate conversions, 13 transport proteins, 5 transcription factors, 5 structural proteins and 2 other lipid-associated enzymes. As some gene products might take part in two or more metabolic pathways, they would be counted twice or more. For example, genes in Kennedy pathway participate in synthesis of storage lipids as well as synthesis of phospholipid. The detailed information is shown in Table 1.

1.2 Annotation of lipid-related genes in the maize database

The lipid-related genes collected in plants were used to perform a BLAST [35] search of maize homologs in GenBank (GenBank Release 169, *E*-Value < 1.0E–1) at the amino acid level. The successful protein matches were analyzed by Interproscan [36] to obtain their functional motifs, which were then manually compared with the motifs in their *Arabidopsis* counterparts. If their motifs were identical, the maize and *Arabidopsis* genes would be annotated as homologous genes [34].

1.3 *In silico* cloning of lipid-related genes in maize

The lipid-related genes of *Arabidopsis* were used to search the maize GenBank EST database (GenBank Release 169, *E*-Value < 1.0E–10), and the homologous partial sequences (or genes) obtained were used as seed sequences to search maize genomic sequences databases (including bacterial artificial clone (BAC), genomic survey sequences (GSS) and mRNA database (maize nr database in GenBank)). The aligned search results were analyzed as follows:

(1) For alignments to match a BAC sequence, if the identities or identities + (Gaps/alignment length) × 100% were

Table 1 Summary of *Arabidopsis* lipid metabolism enzymes and genes[a)]

Pathway	FAD	FAEWM	FAS	GS	LS	M	OM	PS	SpS	SuS	TAGSS	Sum
No. of enzymes	18	10	37	8	33	26	16	32	18	2	7	207
No. of genes	52	66	92	12	104	140	74	69	30	2	19	660

a) fatty acid degradation, FAD; fatty acid elongation and wax metabolism, FAEWM; fatty acid synthesis, FAS; galactolipid synthesis, GS; lipid signaling, LS; miscellaneous, M; oxylipin metabolism, OM; phospholipid synthesis, PS; sphingolipid synthesis, SpS; sulfoplipid synthesis, SuS; triacylglycerol synthesis and storage, TAGSS.

greater than or equal to 98% (insertion or deletion among materials), and more than 80% of the query sequence was covered, we assumed that the BAC clone incorporated the query sequence. A 10 kb stretch of sequence was taken from the target region (the length was adjusted according to the length of the homolog) and viewed as the candidate gene sequences. If there were sequencing gaps, we extended the query sequence to the gap and used it as a seed sequence for further elongation using GSS sequences, as described below.

(2) For alignments to match a GSS or EST sequence, we first identified sequences for which the identities or identities + (Gaps/alignment length) × 100% were greater than or equal to 98%. These sequences were then classified into genomic and EST sequences and clustered using Cap3 [37]. This step was repeated until the query sequence (genome or cDNA) could not be further elongated. Finally, the genomic and cDNA sequences were aligned to get information on the gene structure, which was also verified by comparing with the corresponding gene structure in *Arabidopsis*. For maize genes without cDNA sequences, we adopted two strategies to determine the correct gene structure: Genscan [38] or Fgenesh [39] was used to predict gene structure, and the protein sequences of homologous genes in rice and sorghum were analyzed by tBlastn with the maize DNA sequences. The gene structure was then adjusted manually by applying the intron boundary rules (GT/GC-AG).

Finally, the maize genes were translated into proteins and searched for functional domains, which were then compared with their corresponding homologous proteins in *Arabidopsis* to ensure the accuracy of the genes obtained by the *in silico* cloning.

1.4 *In silico* mapping of lipid-related genes

The nucleotide sequences of a gene were used in a Blastn search against the MaizeSequence database (http://www.maizesequence.org) to identify the BAC clone containing the gene. For each BAC clone, the contig where the gene was located and the 2 most nearest public markers were obtained and used for *in silico* mapping. If no corresponding BAC clone could be found or if the corresponding BAC clone was not mapped, the gene location was recorded as unknown.

1.5 Digital expression profiling of maize lipid-related genes

As of March 2009, there were 2018338 EST sequences stored in NCBI, and we downloaded and classified these sequences according to the tissue source of the cDNA libraries to construct a local platform for expression analysis. If the sample size of a cDNA library was less than 500, it was eliminated from the subsequent analysis (see Figure 1).

Stekel *et al*. [40] attributed the relative quantitative difference in ESTs from different cDNA libraries to two reasons: sampling errors in sequencing, which they referred to as the unintended hypothesis, and expression difference among corresponding tissues, which they referred to as the alternative hypothesis. They developed a likelihood function of the two hypotheses, obtained the maximum likelihood and got a parameter, R, to measure the relative expression level of a gene in various tissues. The function is expressed as:

$$R_j = \sum_{i=1}^{m} x_{i,j} \log\left(\frac{x_{i,j}}{N_i f_j}\right),$$

where m is the number of cDNA libraries, $x_{i,j}$ is the number of transcripts of gene j in the i^{th} library, N_i is the total number of cDNA clones sequenced in the i^{th} library, and f_j is the frequency of transcripts of gene j in all the libraries. A gene with $R > 8$ in a particular library was regarded as having a

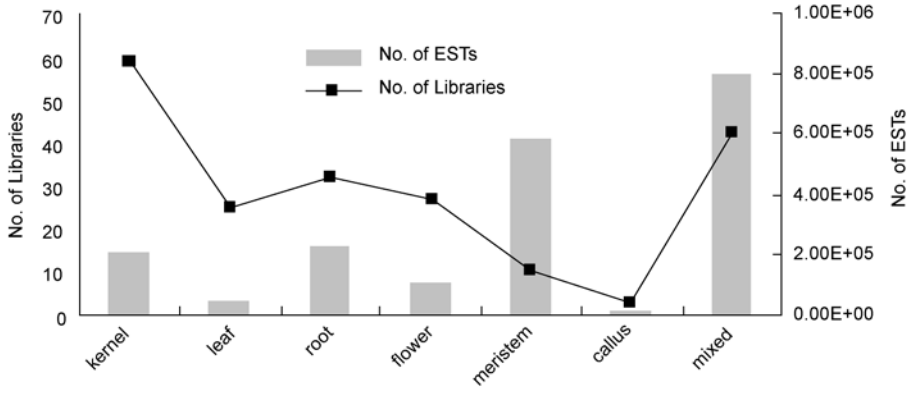

Figure 1 The schematic statistics of maize EST collection. A total of 1984483 ESTs from 182 cDNA libraries have been collected for expression profiles analyses. The ESTs have been classified into six tissues (kernel, leaf, root, flower, meristem and callus) and a mixed library based on the tissue type of the original cDNA library. The kernel library with 59 public cDNA libraries (206005 ESTs) had the most extensive original sources. The mixed library owned the most number of ESTs (799541), originating from 42 cDNA libraries. The callus library has been excluded from the following analysis as it comes from only three original libraries and may not be representative for statistical analysis. Leaf, root, flower and meristem library consist of 25 (48616 ESTs), 32 (223908 ESTs), 27(106667 ESTs) and 10 (584468 ESTs) original cDNA libraries, respectively.

preferential expression in the corresponding library. Otherwise, genes were regarded as showing no preferential expression in the corresponding library.

Varuzza *et al*. [41] developed two significance tests (*P* value and *E* value) for the comparison of digital expression profiles based on the frequency and Bayesian distribution, respectively, which minimizes both type I and type II errors.

This study combined the two methods mentioned above and built expression profiles for lipid-related genes in maize.

1.6 Co-location of kernel oil QTL and maize lipid-related genes

In total, eleven studies that mapped QTLs for maize kernel oil content are summarized in Table 2. The QTLs and the genes from three types of metabolism (FA synthesis, FA degradation and storage lipid synthesis) were integrated into the maize reference map IBM2 2005 Neighbors using QTL-Finder [42]. Co-location analyses were carried out based on the principle that if a lipid-related gene was located within the QTL confidence interval (40 cM, as the resolution of primary QTL mapping is usually 10-15 cM, and the genetic distance of the high-density maize IBM2 2005 Neighbors map is about four times that of the normal maize genetic map [43]), the gene would be assumed to be a candidate gene for the QTL.

2 Results

2.1 Annotation and *in silico* cloning of maize lipid-related genes

Based on sequence and motif similarity to the 636 lipid-related genes identified from other species, we have identified a total of 1003 lipid-related genes in maize (Figure 2). Of these, 42 genes have been experimentally characterized. A number of 732 genes having full-length cDNA sequences have been annotated as lipid-related proteins (134), hypothetical proteins (77) or unknown proteins (521). The remaining 229 genes were newly cloned in this study with 197 genes being supported by ESTs. Among the ten metabolic pathways, the number of genes (172) from FA elongation and wax synthesis was the greatest, while the number of genes (4) from the sulfolipid synthesis pathway was the smallest (Figure 2, Table 3). All the gene information is available at http://www.meta2trait.org.

As shown in Table 3, genes related to maize lipid metabolism were distributed over 10 chromosomes, varying from 57 genes on chromosome 10 to 273 genes on chromosome 1. Apart from genes from the sulfolipid synthesis pathway (SuS) and the sphingolipid synthesis pathway (SpS), genes from the other eight pathways were found across all ten maize chromosomes (Table 3).

2.2 Digital expression profiles of maize lipid-related genes

Digital analysis revealed that a total of 80 enzymes encoded by 97 genes showed tissue-preferential expression patterns. Of the 5 tissues analyzed, root had the greatest number of genes expressed (50 genes encoding 31 enzymes), followed by kernel (45 genes encoding 30 enzymes), meristem (31 genes encoding 18 enzymes) and leaf (1 gene encoding 1 enzyme). No genes showing flower-preferential expression pattern were identified (Table 4). Of the 10 metabolic pathways, no genes from the sulfolipid synthesis pathway (SuS) or the sphingolipid synthesis pathway (SpS) showed tis-

Table 2 Summary of previously identified QTL for maize kernel oil content

Parents	Pop. size	Pop. type	No. of markers	No. of QTL	Methods	References
B73×By804	450	$F_{2:3}$	150	39	CIM	Song *et al*. [44]
B73×By804	298	$F_{2:3}$	183	6	CIM	Zhang *et al*. [45]
B73×By804	223	RIL	228	49	CIM, MIM	Yang *et al*. [46]
IHP76×ILP76	100	$F_{2:3}$	100	17	ANOVA	Goldman *et al*. [47]
IHO90×ILO70	200	$F_{2:3}$	90	41	ANOVA	Alrefai *et al*. [48]
IHO90×ILO70	200	$F_{2:3}$	90	11	ANOVA	Berke *et al*. [49]
IHO90×ILO90	500	$F_1RM_{10}S_2$	479	63	ANOVA, IM	Clark *et al*. [50]
IHP70×ILP70	500	$F_1RM_7S_2$	499	96	IM	Dudley *et al*. [51]
IHO70×ILO70, IHP70×ILP70	500	$F_{2:3}$	499	193	GLM	Dudley *et al*. [52]
IHO90×B73	150	BC_1S_1,TC	110	28	CIM	Wassom *et al*. [53,54]
L-20-01×L-02-03	408	$F_{2:3}$	75	13	CIM	Mangolin *et al*. [55]

BC_1S_1: Backcross with one parental line to which the F1 line was crossed; $F_{2:3}$: Second generation self inter-cross line; $F_1RM_{10}S_2$: The cross was random mated (RM) ten generations and selfed twice to develop lines; $F_1RM_7S_2$: The cross was random mated (RM) seven generations and selfed twice to develop lines; RIL: Recombinant inbred line; TC: Test cross; CIM: Composite interval mapping; MIM: Multiple interval mapping; ANOVA: Analysis of variance; IM: Interval Mapping; GLM: Generalized linear models.

Table 3 Distribution statistics of maize lipid-related genes along chromosomes[a)]

Chr.	FAD	FAEWM	FAS	GS	LS	M	OM	PS	SpS	SuS	TAGSS	Sum
1	15	87	22	6	33	37	19	16	35	1	2	273
2	11	13	17	3	20	29	14	5	11	0	0	123
3	15	11	7	2	15	21	24	13	10	0	0	118
4	6	8	12	2	17	12	10	9	7	0	2	85
5	8	4	8	4	23	20	12	16	8	0	2	105
6	7	11	6	1	16	14	10	7	2	0	2	76
7	12	15	10	3	19	17	5	7	6	0	4	98
8	10	6	11	2	2	12	19	15	7	2	0	86
9	4	10	10	2	14	18	9	2	6	1	1	77
10	5	7	13	1	9	14	3	3	2	0	0	57
Sum	93	172	116	26	168	194	125	93	94	4	13	1098

a) For the abbreviation of metabolisms, see Table 1.

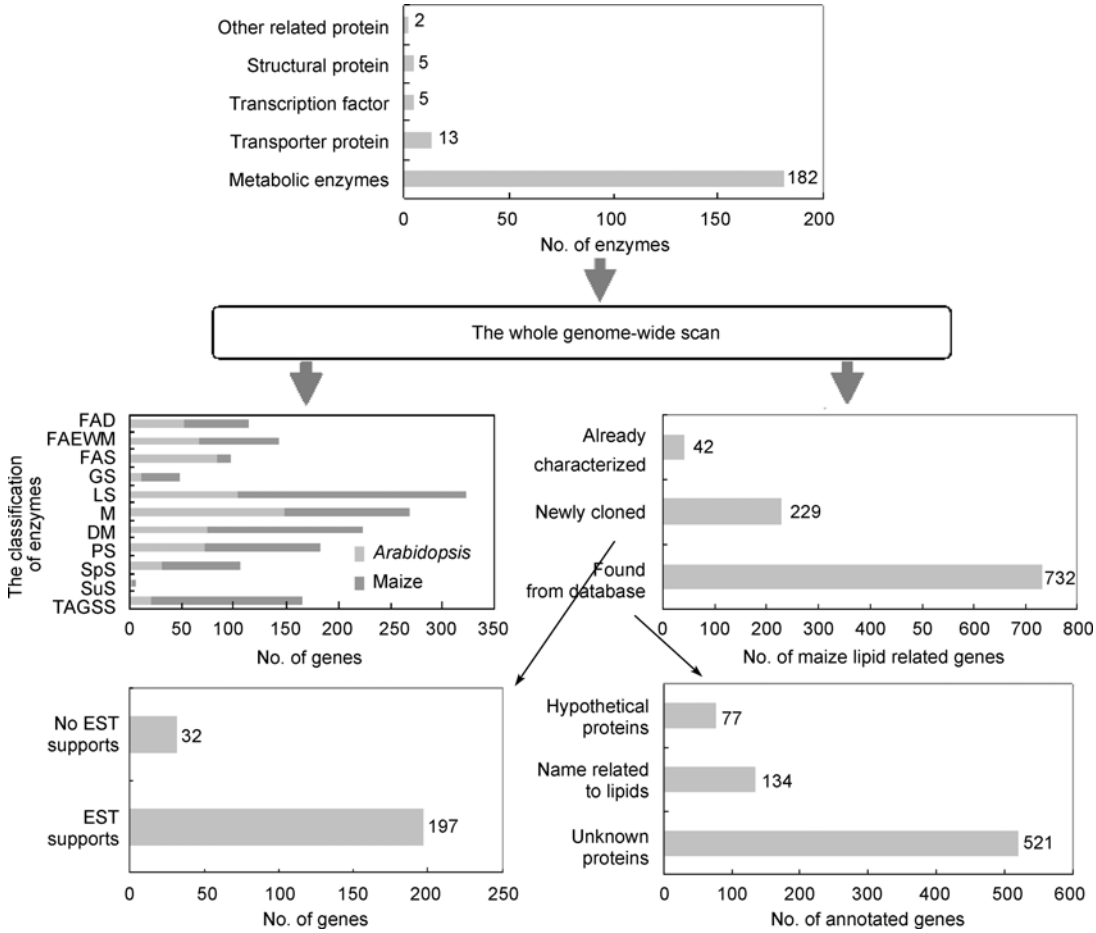

Figure 2 An overview of maize lipid-related gene: construction and global final content. Note: For the abbreviation of metabolisms, see Table 1.

sue-preferential expression. Genes from the other 8 pathways were preferentially expressed at least in one tissue (Table 4, 5).

Table 5 gives a detailed list of the enzymes/genes showing tissue-preferential expression patterns. Some features can be identified. First different enzymes from the same metabolic pathway are highly expressed in different tissues. For example, while FA Multifunctional Protein from the fatty acid degradation (FAD) pathway is highly expressed in kernel, the ketoacyl-CoA thiolase and the enoyl-CoA hydratase are highly expressed in root. Second, even for the same enzyme, different genes are preferentially expressed in different tissues. In the example of ketoacyl-ACP synthase, *ZmCG520* is preferentially expressed in kernel, while

ZmCG1282 is preferentially expressed in root. Third, some genes are expressed highly in more than one tissue; for example, the *ZmCG738* gene from the fatty acid synthesis pathway (FAS) is highly expressed in both kernel and root.

2.3 Candidate genes underlying maize kernel oil QTL

A total of 176 maize kernel oil QTLs have been projected onto the IBM2 2005 Neighbors map based on the consensus flanking markers. These QTLs were unevenly distributed across the maize genome (Figure 3). For example, a cluster of 14 maize kernel oil QTLs was identified in Bins 6.03~6.04, and five QTLs were found in Bin 1.09. The 176 QTLs could be grouped into 70 QTL clusters. These QTL clusters were dispersed along the 10 maize chromosomes, covering 34% of the genome. Chromosome 1 contained the largest number of clusters (12), chromosome 2 had 9 QTL clusters, chromosomes 4, 5 and 6 had 8 QTL clusters each, chromosomes 7 and 9 had 6 QTL clusters each and chromosomes 3, 8 and 10 had 7, 3 and 1 QTL cluster, respectively.

Comparison between the genomic locations of the QTL clusters and the genes from the FA synthesis (FAS), FA degradation (FAD) and TAG synthesis (TAGSS) pathways revealed that 147 of 222 genes (66%) were located within 59 out of 70 QTL clusters (84%). For example, in maize Bin 7.04 where 4 QTLs for oil, palmitic acid and oleic acid were identified, several genes encoding ketoacyl-ACP synthase and acyl-ACP thioesterase were also identified. In total, 147 candidate genes were found within the 70 QTL clusters. Interestingly, all 13 genes of the FA synthesis pathway (FAS) with kernel-preferential expression profiles were co-located with QTL clusters for maize kernel oil content (Table 5, Figure 3).

3 Discussion

3.1 Gene cloning *in silico*

Currently, there are 2 strategies for gene cloning *in silico*. The first uses statistical models, such as dynamic programming or hidden Markov chains, to predict gene structure. Many software programs, including GeneID [56], Fgenesh [39] and Genscan [38], were developed based on this method. The second is based on the alignment of ESTs or protein sequences with the genome sequence. Software of this kind includes GeneParser [57] and GRAIL [58]. The first method depends greatly on the quality of the genome sequence and the models, and it identifies the correct exons about 80% of the time [59]. However, for species with huge genome sizes, poor physical maps or large amounts of repeated sequences the accuracy of this method is low, and short exons cannot be identified using this method. The second method overcomes the short exon problem, but it depends on the accuracy of the sequence alignment and may not give a correct prediction if there are big transposon insertions.

Previous studies show that about 80% of the maize genome is repetitive sequences, so neither of the two methods above can be used alone to correctly identify the gene structures. Therefore, we developed a combinatorial gene cloning system based on mutual verification using comparative genomics, statistical models and sequence similarities. For genes with cDNA or EST sequences, the gene structure can be clearly drawn out by simple alignment. For genes with no cDNA or EST sequences, a putative gene structure can be identified by using some statistical models. All the genes are compared with the *Arabidopsis* genes to ensure the accuracy of the predicted gene structure. Using this method, 1003 maize lipid-related genes were systematically annotated or cloned, of which only 42 genes have been ever validated, such as *DGAT1-2* [60], the Oleosin genes [61] and other genes [61], which are the subjects of patent applications. We also annotated 598 maize lipid-related genes and cloned up to 229 new lipid-related genes in this study. Although the newly cloned genes need further experimental verification, they will surely provide useful information for maize lipid research.

3.2 Maize kernel oil content QTLs and candidate genes

In this study, we have collected a total of 176 QTLs for

Table 4 Number of enzymes encoded by genes with tissue-preferential expression profiles[a)]

	Pathway	FAD	FAEWM	FAS	GS	LS	M	OM	PS	SpS	SuS	TAGSS	Sum
Enzymes / genes with specific expression profiles	Sum of enzymes/genes	18/93	10/172	37/116	8/26	34/168	25/194	16/125	32/93	18/94	2/4	7/13	207/1098
	Kernel	1/2	0/0	9/13	1/1	3/6	8/13	4/5	3/4	1/1	0/0	0/0	30/45
	Root	2/3	0/0	11/15	0/0	2/3	9/22	4/4	2/2	1/1	0/0	0/0	31/50
	Flower	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
	Leaf	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	0/0	0/0	0/0	1/1
	Meristem	1/1	2/3	2/2	1/2	2/4	5/14	1/1	3/3	1/1	0/0	0/0	18/31

a) For the abbreviation of metabolisms, see Table 1; *x*/*y*: *x* and *y* represent the number of enzymes and genes owning tissue-preferential expression profiles, respectively.

Table 5 List of genes showing tissue-preferential expression profiles in different tissues

Tissue	Pathway	Enzyme function	Gene	*R*	*P*	Pathway	Enzyme function	Gene	*R*	*P*
Kernel	FAD	FA Multifunctional Protein	*ZmCG1253*	15	0.00	M	CERT	*ZmCG153*	28	0.00
			ZmCG631	26	0.00			*ZmCG348*	31	0.00
		ACP	*ZmCG738*	26	0.00		Lipid Transfer Protein	*ZmCG733*	42	0.00
			ZmCG76	33	0.00			*ZmCG430*	47	0.00
		α-Ketoacid Decarboxylase	*ZmCG1624*	12	0.00			*ZmCG286*	202	0.00
		Enoyl-ACP Reductase	*ZmCG1643*	51	0.00			*ZmCG841*	12	0.00
			ZmCG378	118	0.00			*ZmCG881*	51	0.00
	FAS	Ketoacyl-ACP Reductase	*ZmCG1279*	19	0.00		Oleosin	*ZmCG638*	27	0.00
			ZmCG768	26	0.00			*ZmCG110*	1324	0.00
		Ketoacyl-ACP Synthase	*ZmCG520*	29	0.00			*ZmCG408*	1162	0.00
		ACP Malonyltransferase	*ZmCG26*	22	0.00			*ZmCG150*	1343	0.00
		PPT1-like Thioesterase	*ZmCG1*	19	0.00		CER2	*ZmCG242*	10	0.00
			ZmCG170	27	0.00		Sphingosine Transfer Protein	*ZmCG1495*	65	0.00
		Pyruvate Dehydrogenase	*ZmCG66*	21	0.00		Allene Oxide Cyclase	*ZmCG42*	37	0.00
			ZmCG107	75	0.00		Hydroperoxide Reductase	*ZmCG1565*	47	0.00
	GS	FAD6	*ZmCG356*	32	0.00	OM	NAD+ Oxidoreductase	*ZmCG709*	30	0.00
		FA Amide Hydrolase	*ZmCG1646*	15	0.00			*ZmCG185*	52	0.00
	LS	Phosphatidylinositol-4-Kinase γ	*ZmCG579*	111	0.00		Oxo-Phytodienoic Acid Reductase (OPR)	*ZmCG161*	18	0.00
			ZmCG1521	84	0.00		Acyltransferase	*ZmCG281*	12	0.00
			ZmCG391	175	0.00		Dihydroxyacetone-Phosphate Reductase	*ZmCG648*	28	0.00
		Phospholipase A2-activating Protein	*ZmCG545*	38	0.00	PS		*ZmCG867*	27	0.00
			ZmCG551	90	0.00		Phosphoethanolamine N-Methyltransferase	*ZmCG563*	42	0.00
	SpS	Acyl-Ceramide Synthase	*ZmCG1604*	16	0.00					
Leaf	M	Lipid-associated Protein	*ZmCG395*	19	0.00					
	FAD	Acyl-CoA Oxidase,	*ZmCG1181*	8	0.00			*ZmCG745*	16	0.00
		FA Omega-Hydroxylase	*ZmCG900*	154	0.00			*ZmCG315*	93	0.00
	FAEWM	Ketoacyl-CoA Synthase	*ZmCG11*	17	0.00			*ZmCG176*	19	0.00
			ZmCG1538	15	0.00		Lipid Transfer Protein	*ZmCG881*	19	0.00
	FAS	Dihydrolipoamide Transacylase	*ZmCG253*	19	0.00			*ZmCG972*	66	0.00
		Homomeric Acetyl-CoA Carboxylase	*ZmCG431*	10	0.00	M		*ZmCG801*	17	0.00
	GS	FAD6	*ZmCG479b*	11	0.00			*ZmCG779*	28	0.00
			ZmCG479	11	0.00		CER2	*ZmCG239b*	16	0.00
AM	LS	Phosphatidylinositol Phosphate Kinase	*ZmCG1417*	12	0.00			*ZmCG239*	16	0.00
			ZmCG1415	12	0.00		Sec14-like Protein,	*ZmCG189*	10	0.00
		Phospholipase A2-activating Protein	*ZmCG551*	13	0.00	OM	Hydroperoxide Reductase	*ZmCG1564*	20	0.00
	M	Lipid Transfer Protein	*ZmCG295*	9	0.00	PS	Diacylglycerol Cholinephosphotransferase	*ZmCG1224*	24	0.00
			ZmCG801b	17	0.00		Phosphatidylserine decarboxylase	*ZmCG1430*	8	0.00
			ZmCG972d	66	0.00		Acyl acceptor Acyltransferase	*ZmCG654*	8	0.00
			ZmCG972c	66	0.00	SpS	Ceramide Sphingobase C4 Hydroxylase	*ZmCG243*	15	0.00
			ZmCG972b	66	0.00					

(*To be continued on the next page*)

(*Continued*)

Tissue	Pathway	Enzyme function	gene	*R*	*P*	Pathway	Enzyme function	gene	*R*	*P*
Root	FAD	Ketoacyl-CoA thiolase	*ZmCG68*	36	0.00	M	Lipid Transfer Protein	*ZmCG375*	29	0.00
			ZmCG1293	10	0.00			*ZmCG798*	32	0.00
		Enoyl-CoA Hydratase	*ZmCG543*	16	0.00			*ZmCG263*	23	0.00
		Acyl Carrier Protein	*ZmCG738*	27	0.00			*ZmCG71*	68	0.00
			ZmCG564	24	0.00			*ZmCG435*	12	0.00
			ZmCG76	22	0.00			*ZmCG972d*	59	0.00
			ZmCG534	28	0.00			*ZmCG972c*	59	0.00
		Acyl-ACP Thioesterase FatA	*ZmCG589*	17	0.00			*ZmCG972b*	59	0.00
	FAS	Dihydrolipoamide Acetyltransferase	*ZmCG1237*	10	0.00			*ZmCG972*	59	0.00
		Dihydrolipoamide Dehydrogenase	*ZmCG1241*	22	0.00			*ZmCG779*	44	0.00
		Enoyl-ACP Reductase	*ZmCG378*	126	0.00			*ZmCG841*	27	0.00
			ZmCG1643	39	0.00			*ZmCG150*	1411	0.00
		Ketoacyl-ACP Reductase	*ZmCG768*	16	0.00			*ZmCG408*	1163	0.00
		Ketoacyl-ACP Synthase	*ZmCG1282*	27	0.00			*ZmCG110*	1335	0.00
		ACP Malonyltransferase	*ZmCG26*	20	0.00		Oleosin	*ZmCG1460*	239	0.00
		PPT1-like Thioesterase	*ZmCG1*	20	0.00			*ZmCG1358*	239	0.00
		Pyruvate Dehydrogenase E1beta subunit	*ZmCG588*	18	0.01			*ZmCG638*	26	0.00
			ZmCG66	17	0.00		Sphingosine Transfer Protein	*ZmCG1495*	70	0.00
	LS	Phosphatidylinositol-4-Kinase γ	*ZmCG579*	112	0.00		Hydroperoxide Reductase	*ZmCG1564*	18	0.00
			ZmCG391	154	0.00	OM	Allene Oxide Cyclase	*ZmCG42*	43	0.00
		Phospholipase A2-activating Protein	*ZmCG551*	36	0.00		Hydroperoxide Reductase	*ZmCG1565*	15	0.00
		CERT	*ZmCG348*	28	0.00		Lipoxygenase	*ZmCG388*	45	0.00
	M	Epoxide Hydrolase	*ZmCG247*	19	0.01	PS	Dihydroxyacetone-Phosphate Reductase	*ZmCG867*	30	0.00
			ZmCG418	10	0.00		Phosphoethanolamine N-Methyltransferase	*ZmCG563*	43	0.00
		Lipid Transfer Protein	*ZmCG286*	127	0.00	SpS	Sphingobase C4 Hydroxylase	*ZmCG316*	15	0.00

Note: For the abbreviation of metabolisms, see Table 1. *R* is the *R* value representing the expression abundance; *P* is the *P* value representing the significance of gene tissue-preferential expression.

maize kernel oil content from 11 studies and projected them onto the IBM2 2005 Neighbors genetic map. Seventy QTL clusters were identified, covering 34% of the maize genome. This is consistent with the finding that about 50 QTLs are involved in the development of high-oil maize [63]. Meanwhile, we have cloned and annotated 1003 maize lipid-related genes. Clarifying the relationship between the QTL clusters and these genes may help us to understand the mechanisms of maize kernel oil accumulation and help us to clone the genes underlying the mapped QTLs. Although ten metabolic pathways are involved in lipid metabolism, only three of them are directly associated with kernel oil content (FAS, FAD and TAGSS). Co-location analysis of the 222 genes involved in these three pathways and the 70 QTL clusters showed that 147 genes (66%) are located within 59 QTL clusters (84%). These two proportions are far greater than the observed portion of genomic regions containing QTL clusters (34%), indicating that the distribution of these genes within the QTL clusters is not due to chance alone. It is noteworthy that all 13 genes from FAS pathway with tissue-preferential expression patterns are located within QTL clusters. The genes co-located with QTL clusters may be good candidates for the corresponding QTL. For example, in bins 6.03-6.04, a total of 14 QTLs were identified in this QTL cluster. The *ZmCG882* gene encoding Acyl-CoA: Diacylglycerol Acyltransferase was also identified in this region, and its function underlying this QTL cluster was confirmed by positional cloning [56]. A total of 147 genes identified here represent ideal candidates for QTL fine mapping and cloning in the near future.

3.3 Utilization of genes identified in this study

Although ten pathways, hundreds of enzymes and a thousand genes are involved in maize lipid metabolism, oil content is still a relatively simple trait compared with more

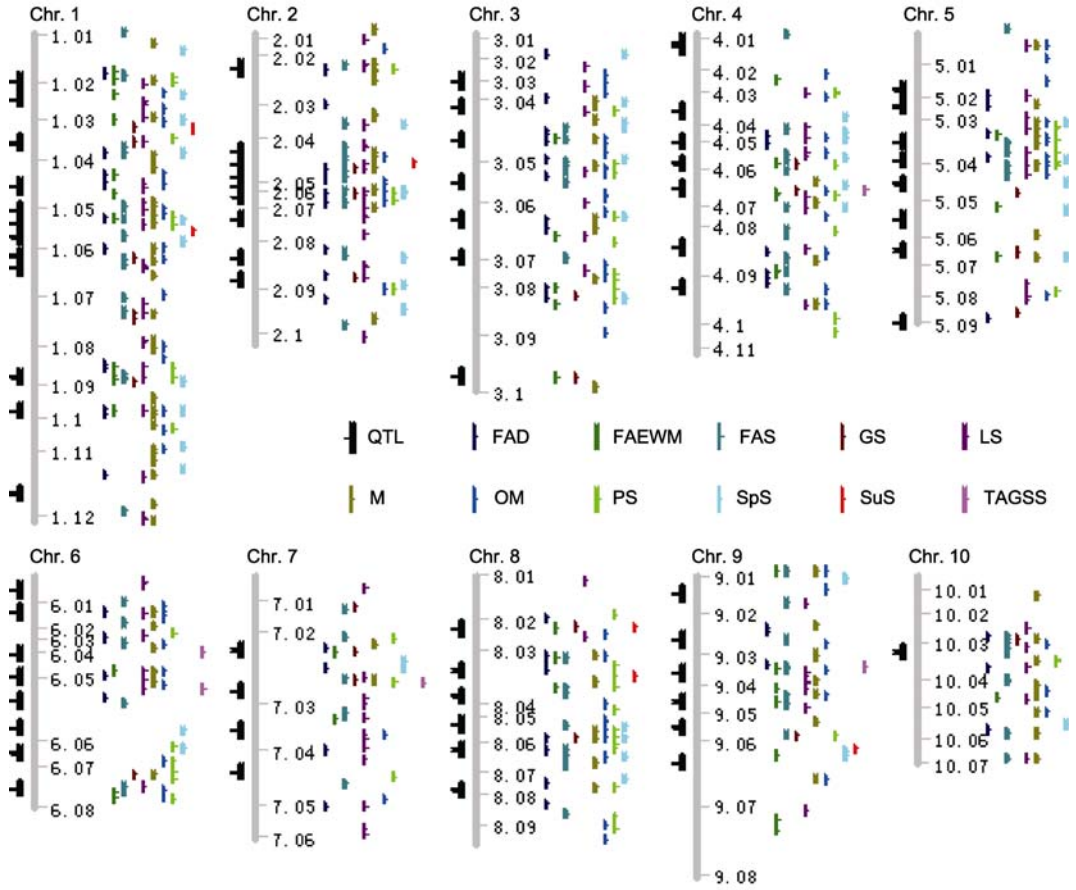

Figure 3 Distribution of maize lipid-related genes and kernel oil QTL clusters along maize chromosomes. For the abbreviation of metabolism, see Table 1. Chr.: Chromosome. x.xx represents Bin. Black objects on the left of each chromosome are the QTL clusters; the ones on the right of each chromosome are genes in maize lipid metabolisms. Different colours represent different metabolic pathways.

complex traits like yield, and the genes regulating it are well-known. Transformation experiments in other species confirmed that regulation of a single gene can result in enhanced oil content. In *Brassica napus* and *Arabidopsis*, genetic engineering of Acyl-ACP thioesterase resulted in a maximum increase of 58% in palmitic acid content [64,65]. Down-regulation of Stearoyl-ACP Δ-9 desaturase in *Brassica napus* improved stearic acid content by 40% [66]. Zheng *et al*. [56] have cloned and conducted genetic transformation of a favorable allele of *DGAT1-2*, which gave a 42% increase in kernel oil content and an increase of up to 107% in oleic acid content. Although there have been great achievements in the genetic engineering of plant lipid metabolism, almost all of them have focused on a single gene (enzyme) or a single metabolite, which certainly could not satisfy the human healthy demands completely. Potrykus and colleagues have transferred multiple genes into rice to create a synthesis pathway for β-carotene, transforming normal rice into golden rice with a high amount of provitamin A, which alleviates vitamin A deficiency [67]. Zhu *et al*. [68] have applied combinatorial nuclear transformation of five carotenogenic genes into a white maize variety. These transformed genes generate plants with extraordinarily high levels of β-carotene and other carotenoids, which are a good example of combinatorial genetic engineering. One important step in combinatorial genetic engineering is the selection of the key genes in the metabolism. This study has annotated and cloned up to 1003 maize lipid-related genes, established *in silico* gene expression profiles of these genes and identified candidate genes underlying maize kernel oil content QTLs. Collectively, the genes we discovered might play important roles in high-oil maize development and are ideal targets for combinatorial genetic engineering. Based on the results from this study, we can name several of these key genes, such as the three upstream genes involved in fatty acid synthesis: the enoyl-ACP reductase gene, the malonyl-CoA:ACP malonyltransferase gene and the ketoacyl-ACP synthase gene. All three genes are preferentially expressed in maize kernels and located within QTL clusters for maize kernel oil content. Thus, over-expression of these three genes may lead to elevated oil content. Another example includes two genes involved in the TAG synthesis and storage pathway, *FAD2* [62] and *DGAT1-2* [56]. Both have been validated as important genes for maize kernel oil con-

tent. In addition, our study found that some Oleosin genes encoding essential structural proteins for oil storage showed kernel-preferential expression and were also located within QTL cluster regions; these genes are also ideal candidates for genetic engineering. Combinatorial engineering of *FAD2*, *DGAT1-2* and Oleosin genes might also result in enhanced oil content. We can also genetically engineer these key genes in combination and then screen different transgenic lines for transgenic materials with the best oil production.

Besides genetic engineering, we can use candidate gene association mapping to identify favorable alleles of the genes in natural populations. This method has been well demonstrated in previous studies. For example, Harjes *et al*. [69] verified the function of *lcyE* in pro-vitamin A synthesis using association analysis and identified important genetic polymorphisms responsible for phenotypic variation. Using a similar strategy, Beló *et al*. [62] confirmed the function of *FAD2* in maize kernel oil content. The genes cloned and annotated, especially those showing a tissue-preferential expression pattern and co-located with QTL clusters, can be regarded as ideal candidates for mining of favorable alleles in natural populations. The genetic variants identified can be further used to develop functional markers used for Maker-Assisted Selection.

Most of the data used in this study was downloaded from websites, and the authors thank very much for the contributors. We also Thank Li Qing, Shi Dianyi, and Ma Hailin for their good advice onto our study and modification of this paper. This research was supported by the National Hi-Tech Research and Development Program of China (*Grant Nos. 2006AA100103, 2006AA10Z183*) *and the College Science Research and Business Plan Project sponsored by the Education Committee of Peking City.*

1 Sprong H, van der Sluijs P, van Meer G. How proteins move lipids and lipids move proteins. Nat Rev Mol Cell Biol, 2001, 2: 504–513
2 Wallis J G, Browse J. Mutants of Arabidopsis reveal many roles for membrane lipids. Prog Lipid Res, 2002, 41: 254–278
3 Broun P, Gettner S, Somerville C. Genetic engineering of plant lipids. Annu Rev Nutr, 1999, 19: 197–216
4 Schneider K, Kienow L, Schmelzer E, *et al*. A new type of peroxisomal acyl-coenzyme a synthetase from *Arabidopsis* thaliana has the catalytic capacity to activate biosynthetic precursors of jasmonic acid. J Biol Chem, 2005, 280: 13962–13972
5 Li C, Schilmiller A L, Liu G, *et al*. Role of β-oxidation in jasmonate biosynthesis and systemic wound signaling in tomato. Plant Cell, 2005, 17: 971–986
6 Angelo A J S, Altschul A M. Lipolysis and the free fatty acid pool in seedlings. Plant Physiol, 1964, 39: 880–883
7 Gueguen V, Macherel D, Jaquinod M, *et al*. Fatty acid and lipoic acid biosynthesis in higher plant mitochondria. J Biol Chem, 2000, 275: 5016–5025
8 Thelen J J, Ohlrogge J B. Metabolic engineering of fatty acid biosynthesis in plants. Metab Eng, 2002, 4: 12–21
9 Beaudoin F, Napier J A. Biosynthesis and compartmentation of triacylglycerol in higher plants. In Daum G. ed. Lipid Metabolism and Membrane Bioiogenesis. Berlin: Springer, 2004. 267–287
10 Kunst L, Samuels A L. Biosynthesis and secretion of plant cuticular wax. Prog Lipid Res, 2003, 42: 51–80
11 Reina-Pinto J J, Yephremov A. Surface lipids and plant defenses. Plant Physiol Biochem, 2009, 47: 540–549
12 Hynes M J, Murray S L, Duncan A, *et al*. Regulatory genes controlling fatty acid catabolism and peroxisomal functions in the filamentous fungus aspergillus nidulans. Eukaryot Cell, 2006, 5: 794–805
13 Graham I A, Eastmond P J. Pathways of straight and branched chain fatty acid catabolism in higher plants. Prog Lipid Res, 2002, 41: 156–181
14 Goepfert S, Poirier Y. β-Oxidation in fatty acid degradation and beyond. Curr Opin Plant Biol, 2007, 10: 245–251
15 Bessoule J J, Moreau P. Lipid Metabolism and Membrane Biogenesis. Berlin: Springer. 2003. 1610–2096
16 Mueller-Roeber B, Pical C. Inositol phospholipid metabolism in arabidopsis characterized and putative isoforms of inositol phospholipid kinase and phosphoinositide-specific phospholipase C. Plant Physiol, 2002, 130: 22–46
17 Chapman K D. Emerging physiological roles for N-acylphosphatidylethanolamine metabolism in plants: signal transduction and membrane protection. Chem Phys Lipids, 2000, 108: 221–230
18 Munnik T. Phosphatidic acid: an emerging plant lipid second messenger. Trends Plant Sci, 2001, 6: 227–233
19 Wang X M. Plant phsopholipases. Plant Mol Biol, 2001, 52: 211–231
20 Futerman A H, Riezman H. The ins and outs of sphingolipid synthesis. Tr Cell Biol, 2005, 15: 312–318
21 Sato N, Moriyama T. Genomic and biochemical analysis of lipid biosynthesis in the unicellular rhodophyte cyanidioschyzon merolae: lack of a plastidic desaturation pathway results in the coupled pathway of galactolipid synthesis. Eukaryot Cell, 2007, 8: 1006–1017
22 Benning C. Biosynthesis and function of the sulfolipid sulfoquinovosyl diacylglycerol. Plant Mol Biol, 1998, 49: 53–75
23 Baysal T, Demirdoven A. Lipoxygenase in fruits and vegetables: a review. Enzyme Microb Technol, 2007, 40: 491–496
24 Feussner I, Wasternack C. The lipoxygenase pathway. Annu Rev Plant Biol, 2002, 53: 275–297
25 Pollard M, Ohlrogge J. Testing models of fatty acid transfer and lipid synthesis in spinach leaf using in vivo oxygen-18 labeling. Plant Physiol, 1999, 121: 1217–1226
26 Germain V, Rylott E L, Larson T R, et al. Requirement for 3-ketoacyl-CoA thiolase-2 in peroxisome development fatty acid beta-oxidation and breakdown of triacylglycerol in lipid bodies of Arabidopsis seedlings. Plant J, 2001, 28: 1–12
27 Stymne S, Stobart A K, Glad G. The role of the acyl-CoA pool in the synthesis of polyunsaturated 18-carbon fatty acids and triacylglycerol production in the microsomes of developing safflower seeds. Lipids and Lipid Metabolism, 1983, 752: 198–208
28 International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature, 2005, 436: 793–800
29 Paterson A H, Bowers J E, Bruggmann R, Dubchak I et al. The sorghum bicolor genome and the diversification of grasses. Nature, 2009, 457: 551–556
30 Borevitz J O, Chory J. Genomics tools for QTL analysis and gene discovery. Curr Opin Plant Biol, 2004, 7: 132–136
31 Varshney R K, Hoisington D A, Tyagi A K. Advances in cereal genomics and applications in crop breeding. Trends Biotechnol, 2006, 24: 490–499
32 Henrissat B, Coutinho P M, Davies G J. A census of carbohydrate-active enzymes in the genome of Arabidopsis thaliana. Plant Mol Biol, 2001, 47: 55–72
33 Ward J M. Identification of novel families of membrane proteins from the model plant *Arabidopsis thaliana*. Bioinformatics, 2001, 17: 560–563
34 Beisson F, Koo A J K, Ruuska S, *et al*. *Arabidopsis* genes involved in acyl lipid metabolism a 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. Plant Physiol, 2003, 132: 681–697
35 Altschul S F, Madden T L, Schaffer A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, 25: 3389–3402
36 Zdobnov E M, Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 2001,

17: 847–849
37 Huang X，Madan A. CAP3: A DNA sequence assembly program. Genome Res, 1999, 9: 868–877
38 Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol, 1997, 268: 78–94
39 Salamov A A, Solovyev V V. Ab initio gene finding in Drosophila genomic DNA. Genome Res, 2000, 10: 516–522
40 Stekel D J, Git Y, Falciani F. The comparison of gene expression from multiple cDNA Libraries. Genome Res, 2000, 10: 2055–2061
41 Varuzza L, Lauretto M S, Brentani H, *et al*. Significance index for digital expression. In: Clamp M, ed. Genome Information Meeting, on CSHL. UK: University of Dundee, 2007.
42 Li L, Yan J B, Lei H Z, *et al*. QTL-Finder: a bioinformatics tool for QTL integration, comparison and discovery of candidate genes across genomes and experiments. In: Zhang Q F, ed. The Abstract of Plant Genomics Conference in China VIII. Plant Genomics Conference in China. Shanghai. 2007. 56–56
43 Lee M, Sharopova N, Beavis W D, *et al*. Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant Mol Biol, 2002, 48: 453–461
44 Song X F, Song T M, Dai J R, *et al*. QTL mapping of kernel oil concentration with high-oil maize by SSR markers. Maydica, 2004, 49: 41–48
45 Zhang J, Lu X Q, Song X F, *et al*. Mapping quantitative trait loci for oil, starch, and protein concentrations in grain with high-oil Maize by SSR Markers. Euphytica, 2008, 162: 335–344
46 Yang X H, Guo Y Q, Yan J B, *et al*. Major and minor QTL and epistasis contribute to fatty acid composition and oil content in high-oil maize. Theor Appl Genet, 2010, 120: 665–678
47 Goldman I, Rocheford T R, Dudley J W. Molecular markers associated with maize kernel oil concentration in an Illinois high protein x Illinois low protein cross. Crop Sci, 1994, 34: 908–915
48 Alrefai R, Berke T G, Rocheford T R. Quantitative trait locus analysis of fatty acid concentration in maize. Genome, 1995, 38: 894–901
49 Berke T, Rocheford T R. Quantitative trait loci for flowering, plant and ear height, and kernel traits in maize. Crop Sci, 1995, 35: 1542–1549
50 Clark D, Dudley J W, Rocheford T R, *et al*. Genetic analysis of corn kernel chemical composition in the random mated 10 generation of the cross of generation 70 of IHO × ILO. Crop Sci, 2006, 46: 807–819
51 Dudley J W, Clark D, Rocheford T R, *et al*. Genetic Analysis of corn kernel chemical composition in the random mated 7 generation of the cross of generations 70 of IHP×ILP. Crop Sci, 2007, 47: 45–57
52 Dudley J W. Epistatic interactions in crosses of Illinois High Oil × Illinois Low Oil and of Illinois High Protein × Illinois Low Protein corn strains. Crop Sci, 2008, 48: 59–68
53 Wassom J J, Wong J C, Martinez E, *et al*. QTL associated with maize kernel oil, protein, and starch concentrations; kernel mass; and grain yield in Illinois High Oil × B73 backcross-derived lines. Crop Sci, 2008, 48: 243–252
54 Wassom J J, Mikkelineni V, Bohn M O, *et al*. QTL for fatty acid composition of maize kernel oil in Illinois High Oil × B73 backcross-derived lines. Crop Sci, 2008, 48: 69–78
55 Mangolin C A, de Souza Jr C L, Garcia A A F, *et al*. Mapping QTLs for kernel oil content in a tropical maize population. Euphytica, 2004, 137: 251–259
56 Guigo R, Knudsen S, Drake N, *et al*. Prediction of gene structure. J Mol Biol, 1992, 226: 141–157
57 Snyder E E, Stormo G D. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. Nucleic Acids Res, 1993, 21: 607–613
58 Uberbacher E C, Mural R J. Locating protein-codeing regions in human DNA sequences by a multiple sensor-neural network approach. Proc Natl Acad Sci USA, 1991, 88: 11261–11265
59 Wang Z, Chen Y, Li Y. A brief review of computational gene prediction methods. Geno Prot Bioinfo, 2004, 2: 216–221
60 Zheng P, Allen W B, Roesler K *et al*. A phenylalanine in DGAT is a key determinant of oil content and composition in maize. Nat Genet, 2008, 40: 367–372
61 Lee K, Huang A H C. Genes encoding oleosins in maize kernel of inbreds Mo17 and B73. Plant Mol Biol, 2004, 26: 1981–1987
62 Beló A, Zheng P Z, Luck S, *et al*. Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Mol Genet Genomics, 2008, 279: 1–10
63 Hill W G. A century of corn selection. Science, 2005, 307: 683–684
64 Voelker T A, Hayes T R, Cranmer A M, et al. Genetic engineering of a quantitative trait: Metabolic and genetic parameters influencing the accumulation of laurate in rapeseed. Plant J, 1996, 9: 229–241
65 Domann P, Voelker T A, Ohlrogge J B. Accumulation of palmitate in *Arabidopsis* mediated by the acyl-acyl carrier protein thioesterase *FATB1*. Plant Physiol, 2000, 123: 637–643
66 Liu Q, Singh S, Green A. Genetic modification of cotton seed oil using inverted-repeat gene-silencing techniques. Biochem Soc Trans, 2000, 28: 927–929
67 Ye X, Al-Babili S, Klöti A, *et al*. Engineering provitamin A (β-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. Science, 2000, 287: 303–305
68 Zhu C F, Naqvi S, Breitenbach J, *et al*. Combinatorial genetic transformation generates a library of metabolic phenotypes for the carotenoid pathway in Maize. Proc Natl Acad Sci USA, 2008, 105: 18232–18237
69 Harjes C E, Rocheford T R, Bai L, *et al*. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science, 2008, 319: 330–333