Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels

Hui Li[1,6], Zhiyu Peng[2,6], Xiaohong Yang[1,6], Weidong Wang[1,6], Junjie Fu[3,6], Jianhua Wang[1,6], Yingjia Han[1], Yuchao Chai[1], Tingting Guo[1], Ning Yang[4], Jie Liu[4], Marilyn L Warburton[5], Yanbing Cheng[2], Xiaomin Hao[1], Pan Zhang[1], Jinyang Zhao[2], Yunjun Liu[3], Guoying Wang[3], Jiansheng Li[1] & Jianbing Yan[4]

Maize kernel oil is a valuable source of nutrition. Here we extensively examine the genetic architecture of maize oil biosynthesis in a genome-wide association study using 1.03 million SNPs characterized in 368 maize inbred lines, including 'high-oil' lines. We identified 74 loci significantly associated with kernel oil concentration and fatty acid composition ($P < 1.8 \times 10^{-6}$), which we subsequently examined using expression quantitative trait loci (QTL) mapping, linkage mapping and coexpression analysis. More than half of the identified loci localized in mapped QTL intervals, and one-third of the candidate genes were annotated as enzymes in the oil metabolic pathway. The 26 loci associated with oil concentration could explain up to 83% of the phenotypic variation using a simple additive model. Our results provide insights into the genetic basis of oil biosynthesis in maize kernels and may facilitate marker-based breeding for oil quantity and quality.

Maize oil is high in energy and in polyunsaturated fatty acids, which makes maize with high oil concentration ('high-oil' maize) a popular resource for food, feed and bioenergy. Thus, the ability to manipulate oil quantity and quality has become a key target for plant breeding and biotechnology-assisted improvement. The oil stored in most plant seeds is composed of triacylglycerols. Studies with the model plant *Arabidopsis thaliana* have generated an extensive understanding of storage-oil biosynthetic pathways, the genes involved and their regulation[1–4].

The long-term selection of high-oil maize populations has led to the development of unique genetic resources, including the Illinois high-oil (IHO) population, the Alexho single-kernel synthetic population and the Beijing high-oil population[5–7]. These have provided opportunities for dissecting the genetic architecture of oil biosynthesis in maize kernels. The continuing phenotypic response to selection over many generations for high kernel oil concentration and correlated traits provides convincing evidence for the involvement of many genes, each having a small effect[8]. On the basis of linkage analysis using high-oil inbred lines developed from these high-oil populations, several QTLs involved in the biosynthesis of maize kernel oil have been identified[8,9]. Recently, the nested association mapping (NAM) population of 5,000 lines and high-density markers has been used to identify 22 QTLs affecting oil concentration[10]. However, despite a good understanding of the plant oil biosynthetic pathway and many of the relevant genes, the molecular basis of natural variation in oil biosynthesis has not been fully elucidated in maize owing to the limited number of parental lines used in the previous studies[8–10].

Genome-wide association studies (GWAS) provide the opportunity to methodically analyze the genetic architecture of complex traits in maize and benefit from the high diversity and rapid linkage disequilibrium (LD) decay in this species[11]. Millions of polymorphisms would be required to ensure complete coverage for a GWAS in maize, considering the small size of conserved LD blocks[12]. Here we used massively parallel RNA sequencing (RNA-seq) to obtain abundant and informative SNPs from expressed regions of the genome and to simultaneously monitor the expression of each of the analyzed loci in the context of its biological function[13]. We analyzed these data, along with 56,110 genomic SNPs from the Illumina MaizeSNP50 BeadChip[14], in an association study of oil concentration and composition. We used two diverse association panels, both of which contained a wide range of phenotypic variation for the traits under study. Some of the identified oil-associated SNPs we validated by expression analysis and/or linkage analysis in biparental populations after PCR amplicon resequencing.

RESULTS

Phenotypic variation

We observed abundant variation in oil-related traits in the association panel of 508 diverse inbred lines, which included 35 high-oil lines[15]. Variation ranged from a 2.3-fold difference in palmitic acid (C16:0) composition to an 8-fold difference in stearic acid (C18:0) composition (**Supplementary Table 1**). Five of the ten measured fatty acids accounted for 98.4% of the oil concentration; these included palmitic

[1]National Maize Improvement Center of China, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing, China. [2]Beijing Genomics Institute-Shenzhen, Shenzhen, China. [3]Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing, China. [4]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China. [5]Corn Host Plant Resistance Research Unit, US Department of Agriculture–Agricultural Research Service, Starkville, Mississippi, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to J.Y. (yjianbing@mail.hzau.edu.cn), J.L. (lijiansheng@cau.edu.cn) or G.W. (gywang@caas.net.cn).

Received 21 May; accepted 6 November; published online 16 December 2012; doi:10.1038/ng.2484

(C16:0, 15.7%), stearic (C18:0, 2.1%), oleic (C18:1, 28.0%), linoleic (C18:2, 51.2%) and linolenic (C18:3, 1.4%) acids (**Supplementary Fig. 1**). Many of the traits were highly correlated, often because they are physiologically correlated (**Supplementary Table 2**). Compared with regular lines, the high-oil lines had higher oil concentration but similar oil composition (**Supplementary Fig. 2**). We observed

Table 1 SNPs and candidate genes significantly associated with oil concentration

Candidate gene[a]	Chromosome	Position[b]	SNP	Allele[c]	MAF	*P* value[d]	QTL[e]	QTL direction[f]	eQTL[g]	Annotation[h]
GRMZM2G080524	1	16370466	M1c16370466	T,C	0.07	7.7×10^{-8}	10		NS	Epoxide hydrolase, *EH*[51,52]
GRMZM2G410515	1	17643572	M1c17643572	T,A	0.06	6.9×10^{-8}	10		4.5×10^{-13}	Phytoene desaturase
GRMZM2G115615	1	170961674	PZE-101132612	C,A	0.06	4.1×10^{-8}	48,49	→	N.S	Tetratrico peptide repeat
GRMZM2G110298	1	248149904	M1c248149904	T,C	0.05	1.3×10^{-6}	10,48	↑	3.5×10^{-12}	Acyl carrier protein, *ACP*[30,53,54]
GRMZM2G134432	2	54358837	PZE-102073982	A,G	0.13	3.0×10^{-7}	9,48,50	↑	NA	Phosphatidylinositol 3 kinase, *PI3Ks.a*[30,55]
GRMZM2G079236	2	149517635	M2c149517635	T,G	0.05	1.3×10^{-7}	10,48	↑	NS	Long-chain Acyl-CoA synthetase, *LACS*[30,56,57]
GRMZM2G176542	3	166664152	M3c166664152	C,G	0.08	4.6×10^{-7}	49		NS	Triglyceride lipases, *TAGL*[30,58–60]
GRMZM2G118423	3	167431166	M3c167431166	C,T	0.08	1.1×10^{-6}	49		NA	Oxidoreductase activity, Cytochrome P450, *CYPOR*[61,62]
GRMZM2G083195	3	178136002	M3c178136002	G,C	0.06	8.5×10^{-7}	[49](D/C)+	→→	NS	Glycerol-phosphate acyltransferase, *GPAT*[30,63–65]
GRMZM2G133675	4	6601732	M4c6601732	A,G	0.06	1.8×10^{-7}	N		2.3×10^{-13}	Regulation of transcription
GRMZM5G847159	4	32810884	M4c32810884	A,G	0.10	9.4×10^{-7}	50	↑	1.7×10^{-10}	Oxidoreductase activity, cytochrome P450, *CYPOR*[61,62]
GRMZM2G122767	4	165680223	M4c165680223	G,C	0.14	8.5×10^{-7}	48	↑	NS	ATP binding
GRMZM2G125268	4	165969105	M4c165969105	G,A	0.05	1.0×10^{-6}	48	↑	4.2×10^{-14}	Aldehyde dehydrogenases
GRMZM2G092321	4	228013669	M4c228013669	C,T	0.12	1.2×10^{-6}	N		2.1×10^{-23}	Unknown
GRMZM2G041060	4	236185943	M4c236185943	G,C	0.06	1.2×10^{-9}	N		NS	Unknown
GRMZM2G065194	5	15700222	M5c15700222	A,G	0.10	6.7×10^{-7}	10,49		1.3×10^{-13}	Short-chain dehydrogenases/ reductases
GRMZM2G439195	5	15800012	M5c15800012	G,C	0.05	3.8×10^{-13}	10,49		7.5×10^{-11}	Maize nicotianamine synthase
GRMZM2G035779	5	25549428	M5c25549428	C,T	0.12	1.5×10^{-6}	49		2.6×10^{-7}	Hydrolase activity
GRMZM2G169089	6	104859429	M6c104859429	C,T	0.15	3.9×10^{-15}	[9,10,23](C)+	↑→	NS	Diglyceride acyltransferase, *DGAT1-2* (refs. 22,30, 66–69)
GRMZM2G092550	7	109329336	M7c109329336	C,A	0.07	1.1×10^{-6}	10		NS	Phosphatidylinositol 3 kinase, *PI3Ks.b*[30,55]
GRMZM2G136072	8	21615641	M8c21615641	G,T	0.07	2.7×10^{-12}	[50](D/C)+	↑→→	9.2×10^{-12}	Oxidoreductase activity
GRMZM2G003022	8	38521846	M8c38521846	G,T	0.06	3.4×10^{-14}	[50](D)+	↑→	NS	COPII-coated vesicles, *COPII*[30,70]
GRMZM2G052855	8	100960678	M8c100960678	G,A	0.05	6.1×10^{-13}	[48](D)+	↑→	2.0×10^{-23}	Unknown
GRMZM2G169240	10	16487751	M10c16487751	A,C	0.05	1.6×10^{-7}	N		NS	Fatty acid desaturase-1, *FAD2* (refs. 22,30, 71–74)
GRMZM2G162972	10	26483664	M10c26483664	T,C	0.18	7.7×10^{-7}	9	↑	1.4×10^{-8}	Unknown
GRMZM5G828253	10	116894888	PZE-110061746	G,A	0.08	1.3×10^{-8}	10		NA	Cytochrome P450, *CYPOR*[61,62]

[a] A plausible biological candidate gene in the locus or the nearest annotated gene to the lead SNP. [b] Position in base pairs for the lead SNP according to version 5b.60 of the maize reference sequence (MaizeSequence, see URLs). [c] Major allele, minor allele; underlined bases are the favorable alleles. [d] *P* value of the oil concentration only. [e] The candidate gene located in one of the QTL intervals as reported previously or in the By804/B73 recombinant inbred line population (B) or in one or both of the $F_{2:3}/F_{2:4}$ populations K22/Dan340 (D) and CI7/K22 (C). N, candidates not located in any QTL interval; +, candidates located in one or more QTL intervals in the B, D and C populations. [f] The allele effect direction in B, D and C populations. ↑, high parent; ↓, low parent; →, not segregated. [g] *P* values for the SNP located within 200 kb of the candidate gene that was most significantly associated with the expression level of this gene. NS, not significant ($P > 1.8 \times 10^{-6}$); NA, not available (no expression data for this candidate gene). [h] Each candidate gene is annotated according to InterProScan (see URLs).

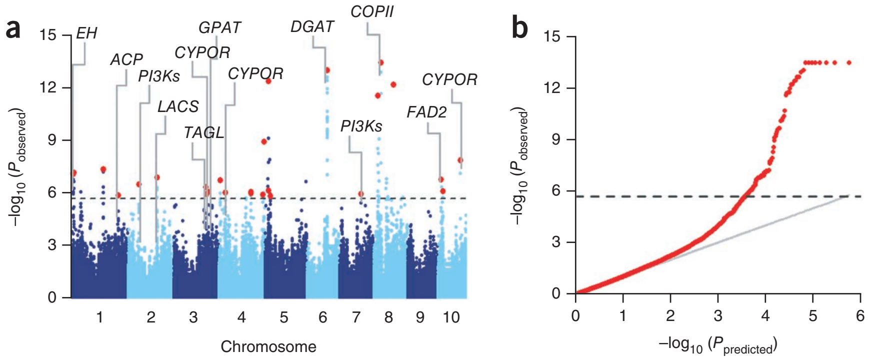

Figure 1 Manhattan and quantile-quantile plots resulting from the GWAS results for oil concentration in maize kernels. (**a**) Manhattan plot for oil concentration. The dashed horizontal line depicts the Bonferroni-adjusted significance threshold (1.8×10^{-6}). Twenty-six unique SNPs are indicated with red dots, and the corresponding genes in the lipid metabolic pathway are shown. Full gene names are listed in **Table 1**. (**b**) Quantile-quantile plot for oil concentration.

broad-sense heritability of >90% for oil concentration and each of the ten compositional traits, based on phenotypic data measured in four environments (**Supplementary Table 1**).

Loci associated with oil-related traits

From more than 3.6 million SNPs identified in 28,769 annotated genes (J. Yan, J. Wang and G. Wang, unpublished data), we selected 1.03 million high-quality SNPs for this study. About 560,000 polymorphisms with minor allele frequency (MAF) ≥0.05 were selected for a GWAS by combining the two genotyping platforms (RNA-seq and SNP array). Association analysis with these polymorphisms identified 63 loci associated with oil concentration and/or at least one of the derived compositional traits at $P < 1.8 \times 10^{-6}$ (**Table 1** and **Supplementary Table 3**). As shown in the quantile-quantile and Manhattan plots for oil concentration (**Fig. 1** and **Supplementary Fig. 3**) and other traits (**Supplementary Fig. 4**), we found notable positive associations after using the mixed linear model to account for population structure and familial relatedness[16,17]. The predicted genes at 21 loci were implicated in lipid metabolism in *Arabidopsis* or other species (**Table 1**, **Supplementary Fig. 5** and **Supplementary Table 3**). The proteins encoded by the remaining 42 genes were classified as transcription factors, enzymes involved in biological pathways including oxidation-reduction reactions and protein metabolism, and transport complexes. The function of approximately one-third of the identified genes is currently unknown (**Fig. 2**). It is possible that these genes are not directly involved in the relevant pathway and that the polymorphic markers within them are actually linked to the causal polymorphisms in a nearby gene.

Although 26 loci were highly significantly associated with oil concentration ($P < 1.8 \times 10^{-6}$), fewer than 7 loci were significantly associated with each compositional trait ($P < 1.8 \times 10^{-6}$). It made biological sense that more genes affected oil concentration but fewer affected oil composition because oil concentration is a product of all the compositional traits combined. To evaluate whether additional associations could be found, we performed conditional association analyses for each of the measured traits using 63 identified loci as covariates in a new GWAS. We identified 11 additional significantly associated loci ($P < 1.8 \times 10^{-6}$), including 3 genes known to be involved in the oil biosynthesis pathway, for 5 oil compositional traits, but no additional loci were identified for oil concentration (**Supplementary Table 4**). This brings the total number of loci with identified associations to 74 (**Fig. 2**). Among these 74 detected loci, 26 associated with oil concentration explained up to 83% of the phenotypic variation, indicating that the additive effect is important in oil synthesis and accumulation. All potential candidate genes within 100 kb (50 kb upstream and downstream of the lead SNP) of the 74 loci are listed in **Supplementary Table 5**.

Validation of the strong association signals ruled out the possibility that they were simply a product of population structure. Again, we performed an association analysis between all 74 lead SNPs and the lead traits, this time excluding the high-oil lines. In this analysis, only 13 of the 26 SNPs affecting oil concentration were still significant at $P < 0.01$; however, the MAF of most SNPs fell to less than 0.05 after exclusion of the high-oil lines, reducing the detection power (**Supplementary Table 6**). For oil component traits, the association significance and MAF of SNPs did not change significantly ($P = 2.4 \times 10^{-4}$ to 2.5×10^{-17}; **Supplementary Table 7**). These results indicate that variation in oil concentration is mainly due to changes in the frequency of non-fixed alleles, whereas component traits have not been the target of substantial selection pressure.

Because of the high marker density and the prohibitive computing time, we chose only one SNP from each gene at random and tested them in pairwise combinations for epistatic interactions. We detected no significant epistatic interactions ($P < 1.0 \times 10^{-4}$), in agreement with previous studies of oil concentration[8,10] and of other quantitative traits studied in the NAM population, including flowering time[18], leaf architecture[19] and disease resistance[20,21].

GWAS, QTL mapping and eQTL mapping

We found considerable overlap between the genes identified via GWAS and previously reported QTLs for oil concentration and composition. Of the 74 identified loci, more than half (43/74) were located in QTLs found previously or mapped in the present study (**Table 1** and **Supplementary Tables 3**, **4** and **8**). Nine of the loci identified in the present association study overlapped the confidence intervals of the 22 QTLs affecting oil concentration reported in the NAM population[10] (**Table 1**). Of the 74 loci identified via GWAS, 27 (16 with the same direction) were located within the confidence intervals of the mapped QTLs for the same traits in at least 1 of our 3 independent mapping populations, providing additional support for our GWAS results (**Table 1** and **Supplementary Tables 3**, **4** and **8**).

One example of overlap includes the locus containing the *FAD2* gene (encoding oleate desaturase) on chromosome 4, which was significantly associated with oleic acid composition via association and independent linkage analysis in our study (**Fig. 3a**) and in a previous one[22]. Another overlap involved the major QTL affecting oil concentration, which is caused by the *DGAT1-2* gene (encoding diacylglycerol acyltransferase)[23]; this gene is one of the most significant loci identified in the present study as well (**Fig. 3b**). We found that a 3-bp insertion and/or deletion (indel) ($P = 2.9 \times 10^{-11}$, $n = 508$) in *DGAT1-2* at the functional site identified previously[23] is the most significant polymorphism. A third example of overlap is the major QTL affecting palmitic acid content, which we mapped to chromosome 9 (ref. 9). The QTL had been cloned and identified as gene *FATB*, and an 11-bp insertion in the last exon of *FATB* decreases the palmitic acid composition, leading to an improvement in the ratio of saturated to unsaturated

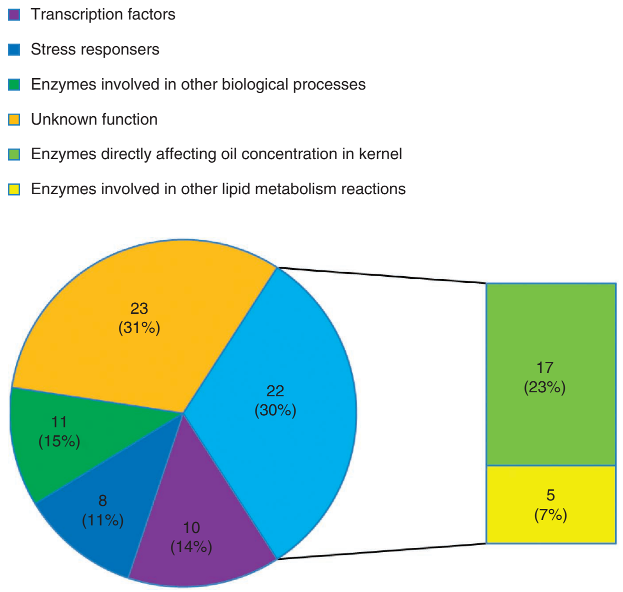

Figure 2 Functional category annotations for 74 candidate genes and their respective percentages identified via GWAS as significantly associated with oil concentration and composition in maize kernels.

fatty acids[24]. In the present study, *FATB* displayed the strongest association signal with palmitic acid composition and affected other compositional traits as well (**Fig. 3c** and **Supplementary Table 3**).

Genetic mechanisms that regulate phenotypic variation can act at genomic, transcriptional and post-transcriptional levels. The differences in expression may account for a substantial proportion of variation in the traits, especially for quantitative traits. We tested the correlation between the polymorphisms identified in the DNA sequence and the mRNA expression levels of the loci identified by GWAS and QTL analysis to look for possible genes regulating oil synthesis and accumulation at the expression level. Expression data were available for 67 of the 74 loci from the RNA-seq data set of 28,769 annotated genes sequenced from kernels collected 15 d after pollination from 368 genotypes. At $P < 1.8 \times 10^{-6}$, 41 of the 67 loci defined clear expression QTLs and exhibited a statistical correlation between DNA sequence polymorphisms and expression levels (**Table 1**, **Supplementary Fig. 6** and **Supplementary Tables 3** and **4**). Notably, expression levels of 14 of the 41 genes were also correlated directly with the phenotypic variation of the target traits, and an additional 18 genes were correlated with a related trait, all at $P < 0.01$ (**Supplementary Table 9**). This strongly suggests that at least some of the genes affect the phenotypic variation via transcriptional regulation.

We constructed a coexpression network to identify the relationships between genes associated with oil metabolism (**Supplementary Fig. 7**). We found that the gene *GRMZM2G132468*, which encodes a putative Ca^{2+}-dependent lipid-binding protein containing a C2 domain, is the central node in this oil transcriptional network (**Supplementary Fig. 7**). C2 domain–containing proteins are involved not only in signal transduction but also in vesicle trafficking and other cellular processes in animals and plants[25]. In the presence of Ca^{2+}, many C2 domains bind the phospholipid membrane[25]. According to the coexpression analysis, *GRMZM2G132468* appears to fall upstream of several key genes predicted to be associated with oil-related traits, including *LACS*, *DGAT1-2*, *TAGL* and *COPII* (**Supplementary Fig. 7**). It is likely that *GRMZM2G132468* is important in regulating downstream biological pathways, including the oil metabolic pathway.

Variation identified by resequencing

Approximately one-third of the annotated candidate genes identified in this study belong to the lipid metabolic pathway (**Fig. 2**). To further

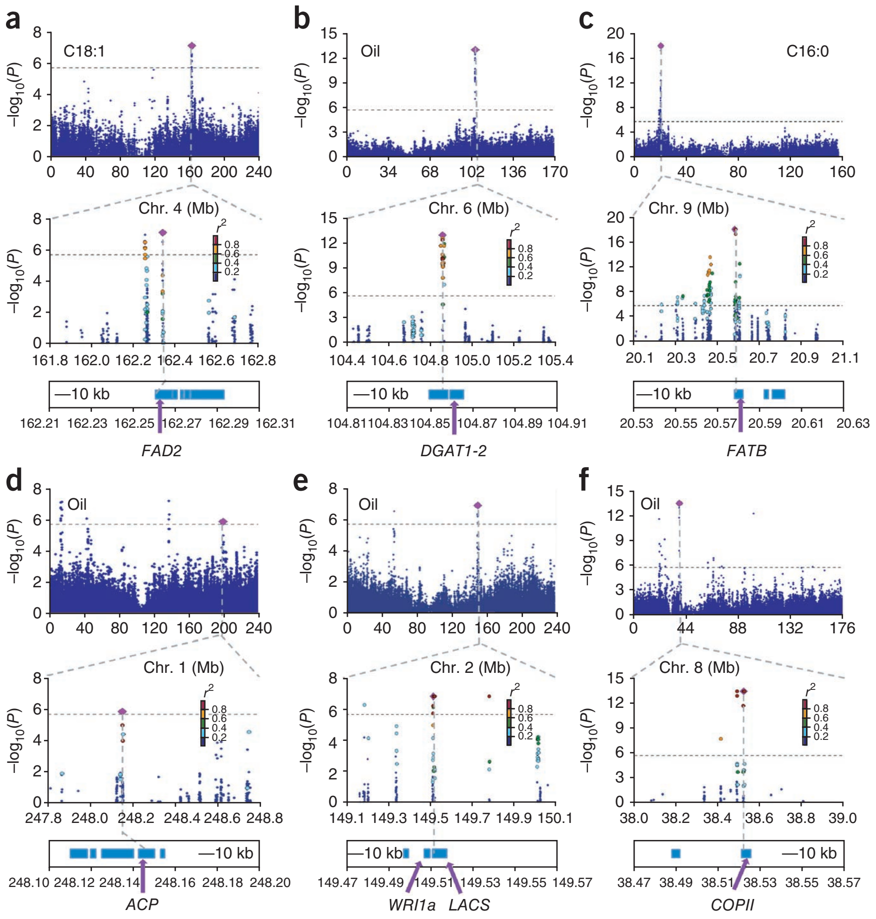

Figure 3 Associations and genomic locations of known and new loci associated with oil concentration and composition. (**a**–**f**) Four previously identified genes, *FAD2* (**a**), *DGAT1-2* (**b**), *FATB* (**c**) and *WRI1a* (**e**) were significantly associated with C18:1, oil concentration, C16:0 and oil concentration, respectively. Three newly identified genes, *ACP* (**d**), *LACS* (**e**) and *COPII* (**f**) were significantly associated with oil concentration. Top, association results of oil-related traits for one chromosome. Bottom, a 0.5-Mb region on each side of the lead SNP (the SNP with the lowest *P* value), whose position is indicated by a vertical gray dashed line. In each plot, the most significantly associated SNP is shown with a purple diamond. Color coding of the remaining markers reflects r^2 values with the most significantly associated SNP. Dashed horizontal lines depict the Bonferroni-adjusted significance threshold (1.8×10^{-6}). The *x* axis shows the genomic position, and the *y* axis shows the significance expressed as $-\log_{10} P$ value.

Table 2 Polymorphisms identified by resequencing of 5 candidate genes

Candidate gene	Trait	Marker[a]	Site[b]	Allele[c]	Frequency	Location	Amino acid change	*P* value
FAD2	C18:1	SNPG/T	Chr. 4_162263608	G,T	154/323	Exon	p.Ser230Ala	2.9×10^{-4}
ACP	Oil	Indel_8	Chr. 1_248150429	0,8	377/32	3′ UTR	No	4.5×10^{-6}
LACS	Oil	Indel_146/472	Chr. 2_149517352	146,472	31/409	3′ UTR	No	2.4×10^{-9}
WRI1a	Oil	Indel_2000	Chr. 2_149342079	0,2,000	87/417	3′ UTR	No	7.0×10^{-4}
COPII	Oil	Indel_20	Chr. 8_38520907	0,20	472/29	5′ UTR	No	2.2×10^{-11}

[a]Candidate functional polymorphisms. [b]Position for the SNP and indel markers according to version 5b.60 of the B73 reference sequence (MaizeSequence, see URLs). [c]The favorable allele for the corresponding trait is underlined. The number represents the insertion size of an allele.

investigate the associations between the allelic variation of these candidate genes and phenotypic variation in the association panel, we chose five candidate genes involved in oil metabolism (*FAD2*, *ACP*, *LACS*, *WRI1a* and *COPII*) to investigate the potential functional polymorphisms capable of causing changes in the phenotype. We did this by resequencing PCR products that encompassed the genetically associated polymorphisms in a subset of 155 inbred lines[26]. The additional polymorphisms identified by resequencing were then genotyped in the complete panel.

FAD2 (*GRMZM2G064701*), which functions in the endoplasmic reticulum (ER), was significantly associated with oleic acid composition (**Fig. 3a** and **Supplementary Fig. 8**). Our GWAS results identified a strongly associated SNP in the first intron of this gene ($P = 9.8 \times 10^{-10}$, $n = 471$; **Supplementary Table 3**). Resequencing the coding and untranslated regions of *FAD2* indicated that the polymorphism (SNPG/T; Chr. 4_162263608) affecting residue 230 (encoding a p.Ser230Ala alteration), resulting in a polarity change in the amino acid, was significantly associated with oleic acid composition ($P = 2.9 \times 10^{-4}$, $n = 477$; **Table 2** and **Supplementary Fig. 8**). The SNPG/T variant localized with a previously mapped QTL for oleic acid composition in the By804/B73 recombinant inbred line population (**Supplementary Fig. 8**). Subsequent investigation revealed that expression of this gene was negatively correlated with oleic acid composition ($r = -0.15$, $P = 4.9 \times 10^{-3}$) and positively correlated with linoleic acid composition ($r = 0.17$, $P = 1.4 \times 10^{-3}$), and thus negatively correlated with the ratio between oleic acid and linoleic acid ($r = -0.18$, $P = 7.0 \times 10^{-4}$; **Supplementary Fig. 8**). We did not, however, detect a significant difference between the expression of the two alleles of SNPG/T or the SNP originally detected in the promoter region ($P = 0.58$), suggesting that these two sites do not cause the observed differences between the expression levels and the target traits in the association panel. Additional sequencing efforts of untranslated and regulatory regions are needed.

ACP (*GRMZM2G110298*) encodes an acyl carrier protein. The homologous gene in *Arabidopsis* functions as the mobile carrier of the growing fatty acid chain in each cycle reaction of fatty acid synthesis[27,28]. Overexpression of an ACP isoform in *Arabidopsis* remarkably increases the fatty acid composition[29]. Our resequencing results identified an 8-bp indel in the 3′ UTR of *ACP* (indel_8) that was strongly associated with oil concentration ($P = 4.5 \times 10^{-6}$, $n = 409$) (**Fig. 3d** and **Table 2**). We found a significant difference between the expression levels of the two alleles at indel_8 ($P = 9.0 \times 10^{-3}$, $n = 367$; **Supplementary Fig. 9**), suggesting that the regulation of this gene at the level of expression can explain at least part of the phenotypic variation and that indel_8 may be the cause of this expression difference.

LACS (*GRMZM2G079236*) contains a Ser/Thr/Gly-rich domain with long-chain Acyl-CoA ligase activity. The *Arabidopsis* homolog activates fatty acyl chains to fatty acid CoAs and participates in the last step of fatty acid synthesis and in cutin, polyester and wax biosynthesis in *Arabidopsis*[30]. Resequencing results identified two completely linked indels (indel_146 and indel_472, a 146-bp insertion with 472-bp deletion and a 146-bp deletion with a 472-bp insertion, respectively) in the 3′ UTR of *LACS* that were significantly associated with oil concentration ($P = 2.4 \times 10^{-9}$, $n = 440$; **Table 2**). The lines containing the 146-bp insertion (with the 472-bp deletion) had lower *LACS* expression but higher oil concentration compared with the lines containing the 472-bp insertion (with the 146-bp deletion; **Fig. 3e**, **Table 2** and **Supplementary Fig. 10**).

WRI1a (*GRMZM2G124524*), located 200 kb upstream of *LACS*, encodes a transcription factor that affects kernel oil accumulation in *Arabidopsis* and maize[31,32]. There are two *WRI1* genes in maize, which are located on chromosome 2 (*WRI1a*) and chromosome 4 (*WRI1b*). In our GWAS results, the lead SNP at *WRI1a* (M2c149341792) in the 3′ UTR was associated with oil concentration ($P = 1.2 \times 10^{-5}$, $n = 368$). On the basis of resequencing, a 2,000-bp indel in the 3′ UTR was also significantly associated with oil concentration ($P = 6.9 \times 10^{-4}$, $n = 504$; **Fig. 3e** and **Table 2**). However, it was not the primary factor responsible for the expression difference, and we uncovered no other significantly associated polymorphisms via resequencing. Although expression of this gene was significantly correlated with oil concentration ($r = 0.30$, $P = 9.8 \times 10^{-9}$; **Supplementary Fig. 11**), the causal polymorphism is still unknown. It may lie in another gene entirely, acting through altered regulation.

Coexpression analysis between *WRI1a* and all other 28,768 genes showed that the expression levels of 2,482 genes were significantly correlated at $P < 1.0 \times 10^{-12}$ with the expression level of *WRI1a*. Among the top 200 genes (**Supplementary Table 10**), 11 were annotated as transcription factors, with a high similarity to either the B3 region of the VP1/ABI3-like protein, or the AP2/ERF family of transcription factors; both of these are crucial for seed development and interacted with *WRI1a* directly or indirectly in this study. There were 33 genes involved in late glycolysis and fatty acid biosynthesis in the plastid, 10 of which were reported previously to be involved in kernel oil biosynthesis[33] (**Supplementary Fig. 12** and **Supplementary Table 10**). Although it is difficult to establish a direct link to the process of lipid metabolism, the other 127 functionally annotated genes were mainly related to carbohydrate metabolic, amino-acid metabolism and transmembrane transport processes, which might provide resources and energy for lipid metabolism (**Supplementary Fig. 12** and **Supplementary Table 10**).

COPII (*GRMZM2G003022*) encodes a sec23 or sec24 protein, which transfers membrane proteins and certain lipids between cellular organelles in the secretory pathways in *Arabidopsis*[30]. A 20-bp indel in the 5′ UTR (indel_20) was significantly associated with oil concentration ($P = 2.2 \times 10^{-11}$, $n = 501$; **Fig. 3f** and **Supplementary Fig. 13**). Indel_20 segregated in the parents of the K22 and Dan340 segregating population, in which a QTL for oil concentration was identified near *COPII*. Indel_20 did not, however, associate with expression of *COPII* (**Supplementary Fig. 13**), suggesting that phenotypic variation may not be regulated via expression differences or that indel_20 may be linked to but not cause these differences.

DISCUSSION

Rapid LD decay and abundant diversity make maize an ideal species for GWAS[11,34]. The resolution of maize GWAS in most cases can reach the gene level, which is much more precise than in self-pollinated plant species, such as rice[35] and *Arabidopsis*[36]. With the rapid development of next-generation sequencing technologies and the continuing decrease in the associated costs, GWAS are rapidly becoming a standard tool for detecting natural variation that accounts for complex quantitative phenotypes in plants[35,36]. The mixed model is a popular method to detect genotype-phenotype associations in plant GWAS, but resource consumption becomes impractical because of large sample size and high-throughput marker density[37]. The improved method used in this study[17] can save a substantial amount of computer time while decreasing false positive rates. However, it may decrease detection power, as it may be too strict in using the Bonferroni threshold as the cutoff after controlling for population stratification and kinship. In addition, LD between strongly selected factors of large effect, not captured by the kinship in the mixed model, will cause overestimation of individual effect size[38]. Users must make decisions based on knowledge of the trait under study.

Using RNA deep sequencing, we obtained more than 1 million high-quality SNPs in 368 diverse maize lines. Using GWAS, we identified 74 loci associated with oil concentration or composition, including 3 previously cloned genes involved in oil biosynthesis (*DGAT1-2*, *FATB* and *FAD2*). We identified complex coexpression networks between the identified genes, and one-third of these genes affected phenotype via transcriptional regulation (**Supplementary Fig. 7** and **Supplementary Table 10**). By resequencing the candidate genes in a large and diverse germplasm collection, we identified polymorphisms that were either causal or in high LD with causal polymorphisms for trait associations with five genes. These include four members of the oil metabolic pathway (*FAD2*, *LCACS*, *ACP* and *COPII*) and one transcription factor (*WRI1a*), which regulates many other genes involved in lipid metabolism and is itself regulated by several other transcription factors. *WRI1a* could prove to be a key regulator of pathways involved in oil biosynthesis that are as yet uncharacterized (**Supplementary Fig. 12**). We found insertions and deletions (some very long) in the UTRs or promoter regions in four of the five validated genes, potentially accounting for gene expression differences seen in the RNA-seq results. Transposable elements are a key source of new genetic variation in maize[39], and transposable element insertions have been documented to be the causal variants in biosynthesis pathways[40,41] and maize domestication genes[42,43]. An example in this study includes the 2,000-bp insertion in the *WRI1a* 3′ UTR, which has high sequence homology to a nonautonomous *Helitron*-type transposable element including nonspecified gene fragments (CENSOR). These *Helitron* elements cause mutations in maize by altering RNA splicing[44,45]. In the present study, we found the *Helitron* sequences in the *WRI1a* 3′ UTR in most inbred lines with regular oil concentration, where they may influence RNA stability and protein translation, resulting in lower oil concentration. Additional studies, including linkage validation in near-isogenic lines or transgenic analysis, will be necessary to conclusively identify this insertion as the causal variant.

Oil concentration and composition are inherited in a mainly additive manner, which has also been observed in previous studies[8–10] and will make breeding for these genes more straightforward. Mutation and favorable allele accumulation are probably two major routes for increasing oil concentration during the selection of high-oil lines[46]. Our results provide evidence for the latter hypothesis. The 23 high-oil lines in the association panel had favorable alleles for 24.6 ± 4.2 (±s.d.) of the 58 loci associated with oil concentration and composition (excluding those loci associated only with derived ratios of composition). The 328 lines with regular oil levels in this study had favorable alleles for only 9.8 ± 2.8 of those loci. Although oil concentration is a polygenic trait controlled by many genes, each with only a small effect, a few genes (seven or less in the present study) with much larger effects were associated with oil composition. Oil concentration is the result of complex biosynthesis pathways, including many known (**Supplementary Fig. 5**) and unknown components. Oil composition, in contrast, represents the interim products of a pathway and is regulated by a few key genes.

In the IHO population, ~50 genes have a role in oil biosynthesis, each of which increases oil concentration by ~0.2% (ref. 8). Genetic improvement for oil concentration via marker-assisted selection or genetic engineering will be inefficient under those circumstances. There were, however, a few key genes with a relatively higher phenotypic effect on the traits in our study, including *DGAT1-2*, the locus with the second strongest association signal for oil concentration (**Fig. 1** and **Table 1**). In a recent study, a line with the favorable allele of this gene, By804, was backcrossed into two lines with regular oil concentration, Zheng58 and Chang7-2, the parents of the most widely planted hybrid in China for the past 10 years. This allele increased oil concentration by >1% in the near-isogenic background without significantly ($P > 0.05$) affecting yield potential or other major agronomic traits[47]. This relatively larger effect may help to explain the difference in selection response in different high-oil populations. The IHO population, which was originally chosen from the open-pollinated variety Burr's White, has been selected for >100 cycles, and the kernel oil concentration has increased by ~15% (ref. 5). However, in two other high-oil populations with more diverse backgrounds and higher selection intensity, it took only 7 and 18 cycles of selection to increase kernel oil concentration by ~8% and~10%, respectively[7]. More favorable genes and alleles with larger effects must have been present for selection to act on the more diverse populations. The identification of major loci in this study will provide the genetic resources and markers needed for additional rapid oil improvement in maize as well as other crops.

URLs. MaizeSequence, http://www.maizesequence.org/; InterProScan, http://www.ebi.ac.uk/interpro/; Primer Premier 5, http://www.premierbiosoft.com/primerdesign/index.html; Primer3, http://frodo.wi.mit.edu/primer3/; CENSOR, http://www.girinst.org/censor/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Gene resequencing data are available under GenBank accession codes JX404032–JX405439, and the SNP data set generated by RNA-seq is available from http://www.maizego.org/Resources.html.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We appreciated the helpful comments on the manuscript from A. Rafalski, J. Dudley, T. Setter and B. Shen. This research was supported by the National Hi-Tech Research and Development Program of China (2012AA10A307), the National Natural Science Foundation of China (31101156, 31123009), the National Basic Research Program of China (2009CB118404), the Ministry of Agriculture of China (2011ZX08009-001) and the National Gene Bank Project of China.

AUTHOR CONTRIBUTIONS

J.Y., X.Y., J. Li and G.W. designed and supervised this study. H.L., W.W., Y.H., Y. Chai, P.Z. and X.H. performed the experiments. H.L., X.Y., W.W., Z.P., J.F., T.G., N.Y., Y.L., J. Liu, Y. Cheng and J.Y. analyzed data. J.W. and J.Z. contributed new regents. J.Y., H.L. and M.L.W. prepared the manuscript, and all the authors critically read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2484.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Thelen, J.J. & Ohlrogge, J.B. Metabolic engineering of fatty acid biosynthesis in plants. *Metab. Eng.* **4**, 12–21 (2002).
2. Graham, I.A. & Eastmond, P.J. Pathways of straight and branched chain fatty acid catabolism in higher plants. *Prog. Lipid Res.* **41**, 156–181 (2002).
3. Beisson, F. *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.* **132**, 681–697 (2003).
4. Baud, S. & Lepiniec, L. Physiological and developmental regulation of seed oil production. *Prog. Lipid Res.* **49**, 235–249 (2010).
5. Dudley, J.W. & Lambert, R.J. 100 generation of selection for oil and protein in corn. *Plant Breed. Rev.* **24**, 79–110 (2004).
6. Lambert, R.J., Alexander, D.E. & Mejaya, I.J. Single kernel selection for increased grain oil in maize synthetics and high-oil hybrid development. *Plant Breed. Rev.* **24**, 153–175 (2004).
7. Song, T.M. & Chen, S.J. Long term selection for oil concentration in five maize populations. *Maydica* **49**, 9–14 (2004).
8. Laurie, C.C. *et al.* The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. *Genetics* **168**, 2141–2155 (2004).
9. Yang, X.H. *et al.* Major and minor QTL and epistasis contribute to fatty acid compositions and oil concentration in high-oil maize. *Theor. Appl. Genet.* **120**, 665–678 (2010).
10. Cook, J.P. *et al.* Genetic Architecture of Maize Kernel Composition in the Nested Association Mapping and Inbred Association Panels. *Plant Physiol.* **158**, 824–834 (2012).
11. Yan, J.B., Warburton, M. & Crouch, J. Association mapping for enhancing maize genetic improvement. *Crop Sci.* **51**, 433–449 (2011).
12. Myles, S. *et al.* Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* **21**, 2194–2202 (2009).
13. Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* **27**, 72–79 (2011).
14. Ganal, M.W. *et al.* A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE* **6**, e28334 (2011).
15. Yang, X.H. *et al.* Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. *Mol. Breed.* **28**, 511–526 (2011).
16. Yu, J.M. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
17. Zhang, Z.W. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
18. Buckler, E.S. *et al.* The genetic architecture of maize flowering time. *Science* **325**, 714–718 (2009).
19. Tian, F. *et al.* Genome-wide association study of maize identifies genes affecting leaf architecture. *Nat. Genet.* **43**, 159–162 (2011).
20. Kump, K.L. *et al.* Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* **43**, 163–168 (2011).
21. Poland, J.A., Bradbury, P.J., Buckler, E.S. & Nelson, R.J. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc. Natl. Acad. Sci. USA* **108**, 6893–6898 (2011).
22. Beló, A. *et al.* Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. *Mol. Genet. Genomics* **279**, 1–10 (2008).
23. Zheng, P.Z. *et al.* A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nat. Genet.* **40**, 367–372 (2008).
24. Li, L. *et al.* An 11-bp insertion in *Zea mays* fatb reduces the palmitic acid content of fatty acids in maize grain. *PLoS ONE* **6**, e24699 (2011).
25. Hurley, J.H. & Misra, S. Signaling and subcellular targeting by membrane-binding domains. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 49–79 (2000).
26. Yang, X.H. *et al.* Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. *Theor. Appl. Genet.* **121**, 417–431 (2010).
27. Shintani, D.K. & Ohlrogge, J.B. The characterization of a mitochondrial acyl carrier protein isoform isolated from *Arabidopsis thaliana*. *Plant Physiol.* **104**, 1221–1229 (1994).
28. Sanchez, J., Agrawal, V.P. & Stumpf, P.K. *Structure, Function and Metabolism of Plant Lipids* (eds. Siegenthaler, P.A. & Eichenberger, W.) (Elsevier, 1984).
29. Branen, J.K., Chiou, T.J. & Engeseth, N.J. Overexpression of acyl carrier protein-1 alters fatty acid composition of leaf tissue in *Arabidopsis*. *Plant Physiol.* **127**, 222–229 (2001).
30. Li-Beisson, Y.H. *et al.* Acyl-lipid metabolism (eds. Somerville, C.R. & Meyerowitz, E.M.) *The Arabidopsis Book*, vol. 1 (American Society of Plant Biologists, 2010).
31. Cernac, A. & Benning, C. *WRINKLED1* encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in *Arabidopsis*. *Plant J.* **40**, 575–585 (2004).
32. Shen, B. *et al.* Expression of *ZmLEC1* and *ZmWRI1* increases seed oil production in maize. *Plant Physiol.* **153**, 980–987 (2010).
33. Pouvreau, B. *et al.* Duplicate maize *Wrinkled1* transcription factors activate target genes involved in seed oil biosynthesis. *Plant Physiol.* **156**, 674–686 (2011).
34. Gore, M.A. *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
35. Huang, X.H. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
36. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
37. Zhang, Z.W., Buckler, E.S., Casstevens, T.M. & Bradbury, P.J. Software engineering the mixed model for genome-wide associated studies on large samples. *Brief. Bioinform.* **10**, 664–675 (2009).
38. Platt, A., Vilhjálmsson, B.J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045–1052 (2010).
39. McClintock, B. The significance of responses of the genome to challenge. *Science* **226**, 792–801 (1984).
40. Harjes, C.E. *et al.* Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. *Science* **319**, 330–333 (2008).
41. Yan, J.B. *et al.* Rare genetic variation at *Zea mays crtRB1* increases β-carotene in maize grain. *Nat. Genet.* **42**, 322–327 (2010).
42. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat. Genet.* **43**, 1160–1163 (2011).
43. Zhou, L.L., Zhang, J.Y., Yan, J.B. & Song, R.T. Two transposable element insertions are causative mutations for the major domestication gene teosinte branched 1 in modern maize. *Cell Res.* **21**, 1267–1270 (2011).
44. Lal, S.K., Giroux, M.J., Brendel, V., Vallejos, C.E. & Hannah, L.C. The maize genome contains a helitron insertion. *Plant Cell* **15**, 381–391 (2003).
45. Gupta, S. *et al.* A novel class of Helitron- related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol. Biol.* **57**, 115–127 (2005).
46. Moose, S.P., Dudley, J.W. & Rocheford, T.R. Maize selection passes the century mark: a unique resource for 21st century genomics. *Trends Plant Sci.* **9**, 358–364 (2004).
47. Chai, Y.C. *et al.* Validation of DGAT1–2 polymorphisms associated with oil content and development of functional markers for molecular breeding of high-oil maize. *Mol. Breed.* **29**, 939–949 (2011).
48. Song, X.F., Song, T.M., Dai, J.R. & Rocheford, T.R. QTL mapping of kernel oil concentration with high–oil maize by SSR markers. *Maydica* **49**, 41–48 (2004).
49. Mangolin, C.A. *et al.* Mapping QTLs for kernel oil content in a tropical maize population. *Euphytica* **137**, 251–259 (2004).
50. Zhang, J. *et al.* Mapping quantitative trait loci for oil, starch, and protein concentrations in grain with high–oil maize by SSR markers. *Euphytica* **162**, 335–344 (2008).
51. Newman, J.W., Morisseau, C. & Hammock, B.D. Epoxide hydrolases: their roles and interactions with lipid metabolism. *Prog. Lipid Res.* **44**, 1–51 (2004).
52. Nawrath, C. The biopolymers cutin and suberin, vol. 2 (eds. Somerville, C.R. & Meyerowitz, E.M.) *The Arabidopsis Book* (American Society of Plant Biologists, 2010).
53. Safford, R. *et al.* Plastid–localised seed acyl–carrier protein of *Brassica napus* is encoded by a distinct, nuclear multigene family. *FEBS J.* **174**, 287–295 (1988).
54. Evans, D.E., Taylor, P.E., Singh, M.B. & Knox, R.B. The interrelationship between the accumulation of lipids, protein and the level of acyl carrier protein during the development of *Brassica napus* L. pollen. *Planta.* **186**, 343–354 (1992).
55. Kurz, E.U. & Lees-Miller, S.P. DNA damage–induced activation of ATM and ATM–dependent signaling pathways. *DNA Repair* **3**, 889–900 (2004).
56. Shockey, J.M., Fulda, M.S. & Browse, J.A. Arabidopsis contains nine long–chain acyl–coenzyme a synthetase genes that participate in fatty acid and glycerolipid metabolism. *Plant Physiol.* **129**, 1710–1722 (2002).
57. Fulda, M., Shockey, J., Werber, M., Wolter, F.P. & Heinz, E. Two long-chain acyl-CoA synthetases from *Arabidopsis thaliana* involved in peroxisomal fatty acid beta-oxidation. *Plant J.* **32**, 93–103 (2002).
58. Hellyer, S.A., Chandler, I.C. & Bosley, J.A. Can the fatty acid selectivity of plant lipases be predicted from the composition of the seed triglyceride? *Biochim. Biophys. Acta* **1440**, 215–224 (1999).
59. Rosnitschek, I. & Theimer, R.R. Properties of a membrane–bound triglyceride lipase of rapeseed (*Brassica napus* L.) cotyledons. *Planta* **148**, 193–198 (1980).
60. Paloccia, C. *et al.* Lipolytic isoenzymes from *Euphorbia* latex. *Plant Sci.* **165**, 577–582 (2003).
61. Benveniste, I. *et al.* CYP86A1 from *Arabidopsis thaliana* encodes a cytochrome P450–dependent fatty acid omega–hydroxylase. *Biochem. Bioph. Res. Co.* **243**, 688–693 (1998).
62. Song, W.C., Funk, C.D. & Brash, A.R. Molecular cloning of an allene oxide synthase: a cytochrome P450 specialized for the metabolism of fatty acid hydroperoxides. *Proc. Natl. Acad. Sci. USA* **90**, 8519–8523 (1993).
63. Murata, N. & Tasaka, Y. Glycerol-3-phosphate acyltransferase in plants. *Biochim. Biophys. Acta* **1348**, 10–16 (1997).
64. Yang, W.L. *et al.* A distinct type of glycerol–3–phosphate acyltransferase with sn–2 preference and phosphatase activity producing 2–monoacylglycerol. *Proc. Natl. Acad. Sci. USA* **107**, 12040–12045 (2010).

65. Tamada, T. *et al.* Substrate recognition and selectivity of plant glycerol-3-phosphate acyltransferases (GPATs) from *Cucurbita moscata* and *Spinacea oleracea*. *Acta Crystallogr. D* **60**, 13–21 (2004).
66. Oakes, J. *et al.* Expression of fungal *diacylglycerol acyltransferase2* Genes to increase kernel oil in maize. *Plant Physiol.* **155**, 1146–1157 (2011).
67. Bouvier-Navé, P., Benveniste, P., Oelkers, P., Sturley, S.L. & Schaller, H. Expression in yeast and tobacco of plant cDNAs encoding acyl CoA: diacylglycerol acyltransferase. *FEBS J.* **267**, 85–96 (2011).
68. Lung, S.C. & Weselake, R.J. Diacylglycerol acyltransferase: a key mediator of plant triacylglycerol synthesis. *Lipids* **41**, 1073–1088 (2005).
69. Xu, J.Y. *et al.* Cloning and characterization of an acyl–CoA–dependent *diacylglycerol acyltransferase* 1 (*DGAT1*) gene from *Tropaeolum majus*, and a study of the functional motifs of the DGAT protein using site–directed mutagenesis to modify enzyme activity and oil content. *Plant Biotechnol. J.* **6**, 799–818 (2008).
70. Nickel, W., Brugger, B. & Wieland, F.T. Protein and lipid sorting between the endoplasmic reticulum and the Golgi complex. *Semin. Cell Dev. Biol.* **9**, 493–501 (1998).
71. Mikkilineni, V. & Rocheford, T.R. Sequence variation and genomic organization of fatty acid desaturase–2 (*fad2*) and fatty acid desaturase–6 (*fad6*) cDNAs in maize. *Theor. Appl. Genet.* **106**, 1326–1332 (2003).
72. Wassom, J.J., Mikkelineni, V., Bohn, M.O. & Rocheford, T.R. QTL for fatty acid composition of maize kernel oil in Illinois High Oil x B73 backcross–derived lines. *Crop Sci.* **48**, 69–78 (2007).
73. Okuley, J. *et al.* Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *Plant Cell* **6**, 147–158 (1994).
74. Dyer, J.M. & Mullen, R.T. Immunocytological localization of two plant fatty acid desaturases in the endoplasmic reticulum. *FEBS Lett.* **494**, 44–47 (2001).

ONLINE METHODS

Association panels: genetic relationship and phenotyping. Association tests were done in an association mapping panel composed of 508 diverse inbred lines (AM508, 473 regular and 35 high-oil lines). This panel was characterized with 36,618 high-quality SNPs from IlluminaMaizeSNP50 BeadChip[14] to estimate population structure and kinship coefficients[75]; three subpopulations were identified[15,75]. One subset of the AM508 panel (CAM155), including 155 temperate Chinese inbred lines, was used for candidate gene resequencing[26]. Another subset of 368 lines (345 regular and 23 high-oil lines), randomly selected from the AM508 panel, was used for RNA sequencing.

The AM508 panel was planted in 2009 in Sichuan, Yunnan and Hainan and in 2010 in Guangxi, all in China. One replicate was planted in each location; field experiments have been described previously[15]. CAM155 inbred lines were planted in Beijing, China, in 2006 and 2007 and Hainan, China, in 2007. Two replications were planted for all lines in each enviroment[26]. More than six ears in each row were self-pollinated for all AM508 and CAM155 lines. The protocols for maize kernel lipid extraction and measurement have been reported previously[9].

RNA preparation and sequencing. All lines in the AM508 panel were divided into two groups (temperate and tropical/subtropical) on the basis of their pedigree information and planted in one-row plots in an incompletely randomized block design within the group with two replicates in Jingzhou, China, in the summer of 2010. Six to eight ears in each block were self-pollinated, and five immature seeds from three to four ears in each block were collected at 15 d after pollination. Equal amounts of immature seeds from two replicates were mixed together for total RNA extraction. Additionally, 3 inbred lines (replicated twice) were added as positive controls to the analysis with the 368 inbred lines. RNA extraction, library construction with 200-bp insert size, 90-bp paired-end Illumina sequencing, read mapping and SNP calling followed published protocols. On average, 73.8 ± 0.7 million reads were generated for each sample, leading to 2,445.9 Gb of high-quality raw sequencing data. In total, 1.03 million high-quality SNPs and 28,769 genes, covering about 70% of the maize predicted genes, were identified. The SNP density in the transcript region was ~1 SNP per 54 bp, and there were 40.3 SNPs per gene. Overall LD decay was rapid, reaching 500 bp ($r^2 = 0.1$) in the 368 lines. There were 10,117 SNPs in common with the SNPs identified by the commercially available MaizeSNP50 BeadChip, and the overall mean concordance rate was 96.7%. For the three technical replicates, the concordance rates between each pair of replicates were greater than 99.6% (J. Yan, J. Wang and G. Wang, unpublished data).

Genome-wide association analysis. A GWAS on kernel oil traits was performed using a mixed linear model[16,17] that took into account population structure and relative kinship to test for statistical association between phenotypes and genotypes in 2 data sets, including 1.03 million high-quality SNPs genotyped by RNA-seq and 56,110 SNPs genotyped by the MaizeSNP50 BeadChip. To combine association results across the two studies, we set a uniform threshold (P $1/n = 1.8 \times 10^{-6}$, n = total markers used). To uncover the unique candidate gene underlying association signals, we performed LD analysis of the significant SNPs on the same chromosome and used a cutoff of <0.2 for the LD statistic r^2. Among the unique association signals identified, several candidate genes in or near (within 50 kb up- and downstream of the lead SNP) known genes were validated. Associated SNPs that were not in or near annotated oil metabolism–related genes were considered more likely to be linked to a more distant gene, the closest of which was considered to be the most likely candidate gene. The physical location of the SNPs was identified based on the maize genomic sequence version 5b.60 (MaizeSequence, see URLs).

Conditional analysis of significant signals. To identify additional independent oil-associated SNPs, we repeated the GWAS for each of the 21 oil-related traits, using the lead SNPs identified in the first GWAS iteration as additional covariates. For some lead SNPs detected by the MaizeSNP50 BeadChip, we performed a conditional analysis using merged genotypes from the MaizeSNP50 BeadChip and RNA-seq data.

Candidate gene resequencing and analysis. Maize gene sequences were obtained from the B73 reference sequence at the MaizeSequence database (see URLs). Primers were designed using Primer Premier 5 (see URLs) or Primer3 (see URLs) to cover the full length and/or the 5′ and 3′ sequences of the maize genes (**Supplementary Table 11**). Sequencing was performed by the Tianyi Huiyuan Bioscience & Technology and the SinoGenoMax Companies, using 3,730 sequencers (ABI). The sequences were aligned using MUSCLE[76] and refined manually in BioEdit[77]. Nucleotide polymorphisms, including SNPs and indels at a frequency of ≥0.05, were extracted in TASSEL[78]. TASSEL was also used to calculate r^2 among polymorphisms using 1,000 permutations. Several markers developed from the candidate functional sites of validated genes were used to genotype the AM508 panel using the primers and PCR conditions listed in **Supplementary Table 11**.

QTL mapping. Linkage analysis was done in three linkage populations: the By804/B73 recombinant inbred line (RIL), and the K22/Dan340 $F_{2:3}/F_{2:4}$ and CI7/K22 $F_{2:3}/F_{2:4}$ populations. The By804/B73 RIL population was derived from a cross between regular line B73 and high-oil line By804, and QTL analysis results have been described in detail previously[9]. As only QTLs for oil concentration and four major oil components (C16:0, C18:0, C18:1 and C18:2) had been identified previously[9], additional QTL mapping for other oil components and the derived ratios mentioned in this GWAS was performed. Over 500 individuals each in the K22/Dan340 and CI7/K22 F_2 populations were planted to develop F_3 families by self-pollinating in Beijing in 2009; a total of 465 $F_{2:3}$ families (K22/Dan340, 237; CI7/K22, 218) were collected for phenotyping and offspring validation. These $F_{2:3}$ families were further planted in two environments (Sanya, 2009; Hubei, 2010) to obtain $F_{2:4}$ families. All 465 F_2 individuals, together with 3 parents, were genotyped using GoldenGate assays (Illumina) containing 1,536 SNPs[79]. A linkage map was constructed using Mapmaker version 3.0 (ref. 80), and QTL mapping using the composite interval mapping method[81] was performed with QTL cartographer version 2.5 (ref. 82).

eQTL mapping. To determine whether SNPs near significantly associated genes act as regulators, the associations between the expression level of each associated gene and SNPs within 100 kb upstream and downstream of the lead SNP were performed using the method described for the GWAS. Only those genes expressed in more than 50% of the 368 sequenced lines that had a mean quantification of more than 10 reads were used in this analysis.

Epistatic interactions. Epistatic interactions were tested in 368 inbred lines using a subset of the SNPs because of the prohibitive number of pairwise comparisons in the original data set. One SNP from each gene was chosen at random throughout the genome. A previously reported method was used for detecting epistatic interactions[18]. We first developed an additive model by stepwise regression (using $P < 1.0 \times 10^{-4}$ as a significance threshold); this was followed by a two-way ANOVA for preliminary screening of significant pairwise marker interactions with the smallest P values using all chosen markers in pairwise combinations. Finally, we tested all significant pairwise interactions to identify epistasis combined with the additive model, and permutation tests were done to find the significance of the effects. If the minimum P value from the additive data model alone was smaller than the fifth percentile of minimum P values from 100 permutations, we concluded that at least 1 epistatic interaction was present in our panel.

Coexpression analysis. To develop a correlated expression network of the chosen genes, we calculated pairwise relative expression coefficients in R[83] and used these coefficients to filter the genes ($P < 1.0 \times 10^{-12}$, $n = 368$). The program Cytoscape[84] was used to draw the network with only the 54 most highly connected genes.

Phenotypic variation explained by multiple SNPs. Stepwise regression was performed to examine the effect of multiple alleles with different functional polymorphisms on oil traits and estimate total variance explained (R^2), using the lm function in R[83]. Before fitting the model, each marker was recoded, substituting the value 1 for inbred lines with a given allele and value 0 for all other inbred lines. To avoid linear dependency, the recoded variables were transformed into a set of independent linear combinations. The model was

doi:10.1038/ng.2484

then fitted using least square estimation. The forward-backward (stepwise) selection of markers on the basis of Akaike information criterion (AIC) was started from fitting the null model (no marker). At each forward step, the global significance of the model was evaluated, as well as the significance of the newly added marker. In additional to AIC, the criteria for accepting a new marker was $P < 0.05$, based on a partial F test for each marker. At each backward step, the least significant marker was dropped from the model. R^2 was calculated as the proportion of total phenotypic variation explained by the optimal regression model.

75. Li, Q. *et al.* Genome-wide association study identifies three independent polymorphisms for α-tocopherol content in maize kernels. *PLoS ONE* **7**, e36807 (2012).
76. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
77. Hall, T.A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98 (1999).
78. Bradbury, P.J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
79. Yan, J.B. *et al.* High-throughput SNP genotyping with the GoldenGate assay in maize. *Mol. Breed.* **25**, 441–451 (2010).
80. Lincoln, S.E., Daly, M.J. & Lander, E.S. Construction genetic maps with MAPMAKER/EXP 3.0. *Whitehead Institute Technical Report*, 3rd edn (Whitehead Institute, 1992).
81. Zeng, Z.B. Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468 (1994).
82. Wang, S., Basten, C.J. & Zeng, Z. Windows QTL Cartographer 2.5. (North Carolina State University, Raleigh, North Carolina, USA, 2005).
83. Ihaka, R. & Gentleman, R.R. A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5**, 299–314 (1996).
84. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011).

doi:10.1038/ng.2484