Distant eQTLs and Non-coding Sequences Play Critical Roles in Regulating Gene Expression and Quantitative Trait Variation in Maize

Haijun Liu¹,², Xin Luo¹,², Luyao Niu¹, Yingjie Xiao¹, Lu Chen¹, Jie Liu¹, Xiaqing Wang¹, Minliang Jin¹, Wenqiang Li¹, Qinghua Zhang¹ and Jianbing Yan¹,*
¹National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China
²These authors contributed equally to this article.
*Correspondence: Jianbing Yan (yjianbing@mail.hzau.edu.cn)
http://dx.doi.org/10.1016/j.molp.2016.06.016

ABSTRACT
A detailed understanding of genetic architecture of mRNA expression by millions of genetic variants is important for studying quantitative trait variation. In this study, we identified 1.25M SNPs with a minor allele frequency greater than 0.05 by combining reduced genome sequencing (GBS), high-density array technologies (600K), and previous deep RNA-sequencing data from 368 diverse inbred lines of maize. The balanced allelic frequencies and distributions in a relatively large and diverse natural panel helped to identify expression quantitative trait loci (eQTLs) associated with more than 18 000 genes (63.4% of tested genes). We found that distant eQTLs were more frequent (~75% of all eQTLs) across the whole genome. Thirteen novel associated loci affecting maize kernel oil concentration were identified using the new dataset, among which one intergenic locus affected the kernel oil variation by controlling expression of three other known oil-related genes. Altogether, this study provides resources for expanding our understanding of cellular regulatory mechanisms of transcriptome variation and the landscape of functional variants within the maize genome, thereby enhancing the understanding of quantitative variations.

Key words: eQTL, RNA-seq, GBS, GWAS, non-coding regulation, Zea mays

INTRODUCTION
Identification of quantitative trait loci (QTLs) influencing the expression level of genes (eQTLs) is fundamental to exploring how genomic variants exert regulatory roles and thus contribute to the understanding of phenotypic variations, from cellular metabolites to morphological changes. Natural populations consisting of a large number of unrelated individuals are frequently used for eQTL studies, in recent years in human beings (Albert and Kruglyak, 2015) and plants (Fu et al., 2013), due to a higher mapping resolution (Albert and Kruglyak, 2015). Various advanced high-throughput technologies, for gene-expression measurement and high-density genotyping, have aided eQTL mapping studies. While it has been suggested that RNA sequencing (RNA-seq) provides higher-quality expression data than expression microarrays (Mooney et al., 2013; Wang et al., 2014), the different genotyping platforms are thought to have their own respective strengths. For example, sequencing-based technologies, including RNA-seq and genotyping by sequencing (GBS, also known as reduced genome sequencing), and array-based genotyping methods, are often used in organisms with large genomes such as maize. Targeted genotyping of known uniformly distributed variants makes data analysis easier, although data on rare alleles are difficult to obtain (Panoutsopoulou et al., 2013). RNA-seq is superior for simultaneously measuring expression quantification and genomic variation, but the identified variants are enriched within the genic region and bias conclusions against intergenic non-coding regulatory loci (Freedman et al., 2011). GBS, a cost-effective single-nucleotide polymorphism (SNP)-discovering approach, has been successfully applied, particularly in crop populations with high diversity and large genomes (Elshire
et al., 2011; He et al., 2014). However, the high ratio of missing data and uneven variant number among different individuals make further data analysis difficult and may impair QTL identification.

Moreover, the relative importance of protein-coding and non-coding regulatory loci to morphological and physiological evolution in particular has been argued for almost half a century (Britten and Davidson, 1969; Carroll, 2008; Albert and Kruglyak, 2015). Recently, non-coding regulation has attracted considerable attention, especially with the discovery of regulatory non-coding RNAs. However, most conclusions have been drawn from case studies and lack genome-wide validation, while the number of uncharacterized non-coding transcripts has increased significantly. It is important to uncover the regulation of expressed genes or proteins, cellular metabolites, and observed traits by distant non-coding sequences. Many previous studies were performed by using biparental segregating populations or association mapping populations with limited sample size or low-density markers, resulting in low resolution and the inability to analyze distant regulation factors (Albert and Kruglyak, 2015). In this study, we created an integrated map that combines variants from deep RNA-seq, GBS, and various arrays with densities of 50K (MaizeSNP50 BeadChip; Ganal et al., 2011) and 600K (Affymetrix Axiom Maize Genotyping 600K Array, hereafter 600K; Unterseer et al., 2014) in an enlarged diverse collection with 540 maize inbred lines. Through incorporating the previous measurement of the expression of 28 769 genes in the maize kernel (Fu et al., 2013) from 368 diverse unrelated individuals, we aim to: (1) provide insights into the regulatory landscape of the maize kernel; (2) dissect regulatory causality and links to phenotypic variations; and (3) elaborate regulatory “temporal-spatial” characteristics, including genomic regulation hotspots and regulation patterns in multiple tissues.

RESULTS
Creation of a Reliable Integrated Variation Map
A global collection of 540 inbred lines was genotyped in the present study, of which 513 lines were previously genotyped with MaizeSNP50 BeadChip (Ganal et al., 2011), a subset of 368 lines was genotyped by deep RNA-seq (Fu et al., 2013), 469 lines by GBS, and 153 lines by 600K array (Figure 1). High consistency between different genotyping methods was observed, at least 96.1%, with an overall average of 97.4% (Table 1). However, as three lines (2%) and 585 loci (3%) had a consistency of less than 95% for the 600K array compared with the 50K chip (Supplemental Figure 1), which might be caused by residual heterozygosity of inbred lines, these were eliminated from further analyses. After comparing the variants between RNA-seq and 600K, we found that the consistency ratio decreases as the missing rate of RNA-seq increases, and reaches a 95% consistency when the missing rate reaches 91% (Figure 1B). Consequently, genotypes with a missing rate higher than 91% in RNA-seq were excluded from further analyses. More SNPs were retained in comparison with the previous study (Fu et al., 2013), which used a 60% missing rate cutoff.

All filtered variants (see Methods) from the four different platforms were combined to improve the accuracy of imputation. Simulation results showed that with a missing rate of greater than 90%, accuracy drops significantly, from higher than 90% to lower than 75% (Figure 1C); thus only the
merged SNPs with a missing rate of less than 90% were retained. The removing-then-imputation and imputation-then-removing strategies were compared, and the former, with a higher imputation accuracy (96.93% versus 95.89% on average), was applied for the final analysis. By applying the optimized imputation parameters to all merged SNPs, the average), was applied for the final analysis. By applying the optimized imputation parameters to all merged SNPs, the median missing rate dropped significantly from 72% to 5% after imputation (Figure 1D). In total, 2.65M SNPs were obtained from 540 inbred lines, with more than half (1.4M, 52.8%) being rare (minor allelic frequency [MAF] of <5%). By examining the contribution to imputation accuracy of the four genotype platforms, we found that the additional data from the 600K array greatly improved imputation accuracy (Figure 1E) and that SNPs genotyped by 600K array both for individuals and loci always showed high accuracy of imputation.

To evaluate the reliability of our imputed integrated variation map, we collected the variants identified by resequencing of PCR products on the same panel in different laboratories and with different times of reproduction, and found that the average consistency of the total 477 different overlapped loci was 94.53% (Supplemental Table 1 and Supplemental Figure 2). As mentioned above, a few loci brought down the average consistency (Supplemental Figure 3A), and the number of resequenced lines with these inconsistencies was significantly lower (P = 0.003; Supplemental Figure 3B) than the number of lines with highly consistent loci, suggesting a high residual heterozygosity rate in some lines. Another reason for the inconsistency may have been the presence of short (especially single-nucleotide) tandem repeats with small InDels at some loci (Supplemental Figure 4), which made SNP calling more complicated and prone to error. Interestingly, the rare (MAF <5%) subset had a slightly higher accordance ratio (95.25% versus 94.13%), which suggests that allele frequency has little effect on SNP calling. The integrated map has the highest reported density (2.65M loci), the highest number of individuals (540), and a good distribution of SNPs on the chromosomes, with good coverage in intergenic regions (Figure 2). A large number of variants are predicted to produce severe phenotypes or loss of function (Figure 2B), especially those rich in splice-related variants, which always change the function of encoded proteins. Linkage disequilibrium decays rapidly in this panel and implies high resolution (especially compared with array-based methods) in association analysis (Figure 2C).

---

Table 1. Consistency between Different Genotyping Methods.
The numbers in lower triangle of the matrix represent the numbers of overlapping SNP loci considered between different genotyping platforms; the percentages in upper triangle are consistency rates for the overlapping loci identified by each pair of genotyping platforms (mean/median).

<table>
<thead>
<tr>
<th>Index</th>
<th>50K</th>
<th>600K</th>
<th>GBS</th>
<th>Rseq_raw</th>
<th>Rseq_M91%</th>
</tr>
</thead>
<tbody>
<tr>
<td>50K</td>
<td>–</td>
<td>96.97%/100%</td>
<td>96.12%/99.55%</td>
<td>96.32%/100%</td>
<td>96.53%/100%</td>
</tr>
<tr>
<td>600K</td>
<td>19 495</td>
<td>–</td>
<td>98.48%/100%</td>
<td>97.13%/100%</td>
<td>97.53%/100%</td>
</tr>
<tr>
<td>GBS</td>
<td>5885</td>
<td>39 131</td>
<td>–</td>
<td>97.57%/100%</td>
<td>97.93%/100%</td>
</tr>
<tr>
<td>Rseq_raw</td>
<td>12 691</td>
<td>141 958</td>
<td>203 580</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Rseq_M91%</td>
<td>11 363</td>
<td>123 948</td>
<td>161 352</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

---

Distant eQTL and Non-coding Sequences Are Dominant for Regulation

Fu et al. (2013) performed large-scale eQTL analysis based on the SNPs derived from RNA-seq. In this study, the integrated map was used to identify the regulatory factors affecting gene expression by using the linear mixed model (LMM; Yu et al., 2006), incorporating population structure, hidden confounding factors, and relatedness, as in the previous study (Fu et al., 2013). As expected, many more associations with gene expression under a strict cutoff (P < 7.97 × 10⁻³⁰; 1/n) were identified, of which almost one-quarter (4397 of 18 243) were new compared with the previous study (Fu et al., 2013). For traits identified with eQTLs in both studies, more than half (62%) included newly identified eQTLs (Figure 3). The average number of eQTLs identified for each trait increased significantly when using the new variation map (3.35 versus 1.14), and most of the novel eQTLs were distant to the gene involved (Figure 3B). As a result, the ratio of distant eQTLs was higher than previously reported (93% versus 45%). This was also true when a stricter threshold was used (larger eQTL distance considered as local, Figure 3C) and with lower P value cutoffs for lead SNP (until the P value reached 1 × 10⁻²⁰; Supplemental Figure 3). Most lead eQTL SNPs (71.6%) were >100 kb away relative to their regulated target genes and enriched in the 1 Mb region, while another 8% of the lead SNPs were distant eQTL SNPs were located on different chromosomes (Figure 3D). The local eQTLs tended to have larger effects, which was consistent with the previous study; however, the sum of explained phenotypic variations by either local or distant eQTLs was larger than previously determined (Supplemental Figure 4). The distribution of lead SNPs for local eQTLs showed peaks at the 5' and 3' ends of genes (Supplemental Figure 5), which was consistent with previous conclusions (Mazumder et al., 2003; Wilkie et al., 2003; Ringné and Krogh, 2005; Fu et al., 2013).

The distributions of associated lead SNPs for local and distant eQTLs were compared, and significant differences were found for all of the comparison categories classified by different predicted effect consequences (Figure 3E and Supplemental Figure 7). An enrichment was observed for synonymous SNPs both for local and distant eQTLs located to protein-coding genes, which suggested the importance of synonymous SNPs in expression regulation, while missense ones likely contributed more to structural changes. Lead SNPs were less likely (measured by percentage) to be located within non-coding regions even
Critical Distant eQTLs and Non-coding Sequences in Maize

Molecular Plant

though the number of variants was higher. However, a significant enrichment was also found for distant eQTLs within intergenic and intron regions, which suggests a potential distant regulatory role for non-coding sequences.

Distant-Acting eQTLs Clustered in Hotspots

A region affecting expression of many distant genes is called in this study a distant-acting eQTL hotspot. We applied a new statistical approach that identified hotspots taking into account dynamic genomic density (Silva et al., 2014) to identify robust hotspots, even with different window sizes (Supplemental Figure 8). In total, 518 hotspots (adjusted < 0.05; Figure 4 and Supplemental Table 2) were identified, covering 1.8 Mb of the genome (less than 0.1% of whole genome) and averaging 3.42 kb in size, which regulate the expression of 2090 genes (7.3% of testable genes and 11.5% of genes with identified eQTLs). Interestingly, some hotspots were found in gene-poor regions on several chromosomes (Figure 4A and Supplemental Figure 8). Of the total number of hotspots, 39% (or 204) were fully within genes, another 59% (or 307) overlapped with genes, and the remaining 15% (79) were completely located in non-genic regions. The importance of distant regulation for non-coding regions was recognized by a recent study (Albert and Kruglyak, 2015). In another study, 98 trans-hotspots (covering 141 Mb or 7% of the maize reference genome) were identified using a biparental population (Li et al., 2013b). Seventeen (17.3%) overlapped with our present results, including the most significant one in present study and two of the top 10 hotspots in previous work (Li et al., 2013b).

One distant-acting eQTL hotspot (Chr1: 47 905 716…47 907 183; \( P = 1.36 \times 10^{-22}; \) Figure 4A) was located closely upstream of several A-type R2R3 Myb-like transcription factors, including \( p1 \) (pericarp color1, GRMZM2G084799), \( p2 \) (GRMZM2G057027), and some others (GRMZM2G129872, GRMZM2G016020). The hotspot was also found to be associated with many flavonoid metabolites (Figure 4B; these metabolic phenotypes were measured by Wen et al., 2014), and also regulated the expression of a number of genes (Figure 4B and Supplemental Figure 9), more than half of which (7 of 11 or 63.64%) were related to a flavonoid metabolic pathway and were determined to be controlled (or affected) by \( p1 \) (Morohashi et al., 2012). This finding contributes to the understanding of the regulation of the flavonoid metabolic pathway.

Hotspot analysis presents an enhanced view of complex regulatory networks. RNA-binding proteins (RBPs) play important roles in RNA metabolism by governing all aspects of post-transcriptional gene regulation (Dreyfuss et al., 2002), including mRNA stabilization, alternative splicing, mRNA localization, and even chromatin modification. In addition to playing a role in the diverse developmental processes, they are also involved in hormone signaling to help plants to adapt to rapidly changing environments (Lorković, 2009; Ambrosone et al., 2012). We found two hotspots located upstream and downstream of a zmRBP gene (GRMZM2G171518) that has been shown to affect 27 downstream genes (Figure 4A and 4C, Supplemental Figure 10, and Supplemental Table 3), including a gene involved in nuclear mRNA splicing (GRMZM2G401561), two auxin-binding proteins (GRMZM2G078508 and GRMZM2G064371), an NAC transcription factor (GRMZM2G083347) involved in auxin signaling and the regulation of plant stress responses (Olsen et al., 2005; Nuruzzaman et al., 2013), a DNA-directed RNA polymerase (GRMZM2G129457), a ubiquitin-like modifier of autophagy-related 8d (atg8d, GRMZM2G134613), a PHD finger protein (GRMZM2G115424) involved in chromatin-mediated gene regulation, and a set of enzymes involved in diverse metabolic pathways. These targets and their downstream-regulated genes together constitute a complex regulation network, and the RBP gene is likely to be one of the key nodes.

Spatiotemporal Gene-Expression Patterns

Transcriptome-level regulatory changes in gene expression are a flexible and dynamic means of adaptation (Liu et al., 2015), and are involved in the determination of different cell types. In the present study, the expressed genes were first classified into
regulatory levels (Figure 5 and Supplemental Figure 11A): the most upstream ones only play regulatory roles (Only_Reg), and the most downstream genes are being regulated, while the intermediate level genes are both regulators and subject to regulation (named “Both”). Interestingly, genes at the upstream and intermediate levels of regulatory networks were expressed at significantly higher levels than genes only being regulated ($P = 1.31 \times 10^{-8}$ and $1.96 \times 10^{-6}$, respectively; Figure 5C) and displayed a lower expression variability (measured by coefficient of expression variation; $P = 2.25 \times 10^{-67}$ and $6.03 \times 10^{-130}$, respectively; Figure 5D). It should be noted that this trend was seen not only in the present study (whole kernel of 15 days after pollination), but also throughout the life cycle of maize and in different tissues (data from Chen et al., 2014).

Regulators acting only as distant eQTLs (Only_Dist; Supplemental Figure 11B) were found to be expressed significantly less than those uniquely playing local regulatory roles (Only_Local) and those involved both in distant and local regulation (Both_R) in different tissues (Supplemental Figure 12A). Those acting only as distant regulators (Only_Dist) also showed larger coefficient of expression variations (Supplemental Figure 12B). This expression divergence between local and distant eQTLs may reflect different effects on phenotypic variation. Gene ontology enrichment analysis provided additional support, since various binding molecular functions such as cofactor binding, nucleoside binding, and coenzyme binding were enriched (false discovery rate <0.05) for Only_Dist types, and several catalytic activity function terms, including hydrolase activity, molecular transducer activity, nucleoside-triphosphatase activity, and pyrophosphatase activity, were enriched for Only_Local types. Most of the genes examined are expressed at similar levels across tissues, while only a small number of genes are involved in tissue-specific regulation. Separately, we compared genes regulated by distant eQTLs (RB_Dist) with genes regulated only by local eQTLs (RB_Local) and genes regulated by both classes of eQTLs (RB_Both; Supplemental Figure 11C). The genes regulated only by distant eQTLs (RB_Dist) were expressed at significantly lower levels than the other two types (Supplemental Figure 13A), and displayed larger expression differences between tissues (Supplemental Figure 13B). Overall, the distant functional regulators and their regulated targets tend to be more spatiotemporally variable, and thus can contribute to
tissue-specific characteristics, consistent with the previous study (Albert and Kruglyak, 2015).

**eQTLs Link Genetic Variation with Phenotype Changes**

In previous research (Li et al., 2013a), 26 loci associated with kernel oil concentration were identified and more than one-third of the loci were shown to be significantly associated with the expression level of candidate genes based on 560K SNPs derived from expressed sequences within 368 lines. Doubtlessly a higher density of markers and a bigger sample size would improve detection power and resolution (Yang et al., 2014). In this study, the newly integrated map with 1.25M SNPs and more than 500 lines has been shown to improve detection...
Thirteen new loci affecting total oil concentration were identified compared with the previous results (Table 2). Of the previously identified 26, 19 loci were also identified using the new probability threshold ($8.0 \times 10^{-7}$ for new versus $1.8 \times 10^{-6}$ for old). More importantly, the more balanced integrated map provided the opportunity to explore the potential functions of non-genic sequences that had not been fully studied previously. Several new QTLs were mapped to intergenic regions, including one at the end of chromosome 4 (Figure 6 and Table 2). A total of eight SNPs (physical position from 141,969,034 to 142,149,527) were significantly associated with total oil concentration, and notable phenotypic differences exist between different alleles at these loci (Figure 6B and Supplemental Figure 15A). This intergenic region was also determined to be regulating expression of another three distant (>193 kb) genes ($P = 2.3 \times 10^{-4}$ for FADD, $P = 2.14 \times 10^{-9}$ for GPI, $P = 5.37 \times 10^{-15}$ for GLTP1, GRMZM2G125556; Figure 6A) whose expression level was positively correlated with the phenotypic variation (Figure 6C and Supplemental Figure 15B). Very low linkage disequilibrium (LD) observed between the associated SNPs and the variants within their targets (Figure 6D) indicated that the association could not be confused with local genome structure. Interestingly, all the eight lead SNPs displayed potentially epistatic interaction ($P = 1 \times 10^{-6}$) with the upstream candidate (GRMZM2G066618) but not the others (Supplemental Figure 16A), and many other non-significant SNPs located in the intergenic region interacted with the candidate gene GRMZM2G125556 (Supplemental Figure 16B; $P = 1 \times 10^{-5}$). The gene homologous to GRMZM2G066618 in Arabidopsis is AT4G28570, which encodes long-chain fatty alcohol dehydrogenase (FADD) and participates in fatty acid metabolism (Okuley et al., 1994; Li-Beisson et al., 2010). GRMZM2G162670 is a lipid transfer protein and functions in the first step of the glycosylphosphatidylinositol (GPI) anchor biosynthesis that is related to fatty acid remodeling (Maeda et al., 2007; Li-Beisson et al., 2010; Loizides-Mangold et al., 2012), while GRMZM2G125556 is a glycolipid transfer protein (GLTP1), which is also involved in the oil metabolism pathway (Li-Beisson et al., 2010). The function of the non-genic locus was unclear since no long non-coding RNA (Li et al., 2014), candidate microRNA (Zhang et al., 2009), or even any expressed sequence (Fu et al., 2013) were located with high power (Figure 6 and Supplemental Figure 14).
confidence in the region. Based on the eQTL and genetic analysis, we propose a model in which the unknown non-genic sequence regulates the expression of the three oil-related genes that together affect oil concentration in the kernel. However, since the whole kernels were used for RNA-seq experiments in the present study and the oil concentration is considered to be associated with the embryo size (where nearly all oil is synthesized), the newly identified intergenic QTL could regulate oil concentration by altering the embryo size. Further detailed work is needed to resolve these issues.

**DISCUSSION**

**Distant Regulation Is Critical**

Marker density and population size are two major factors affecting genome-wide association studies (GWAS) (Yan et al., 2011). In the present study marker density was increased from 560K to 1.25M with MAF >0.05, particularly the markers from non-genic regions, in a well-studied association mapping panel, and the panel size was enlarged from 368 to 540. The integrated high-density map and enlarged population size increased the QTL detection power and resolution, and presumably had a higher sensitivity for detecting weaker distant eQTLs, which provided the opportunity to reassess previous studies. Many theoretical and experimental studies have already proved that genetic long-range control for gene transcription is vital to normal development (Kleinjan and van Heyningen, 2005; Kleinjan and Lettice, 2008; Narula and Igoshin, 2010; Van Heyningen and Bickmore, 2013; Xiang et al., 2014). In the previous study with 368 inbred lines (Fu et al., 2013), it was found that the number and effect of local eQTLs was greater than that of distant eQTLs, based on expressed sequence-derived markers. The higher density and more balanced marker distribution was used to reanalyze the eQTLs with several new findings: (1) Expression QTLs for 25% more genes were identified; (2) 62% of genes were identified with more eQTLs; (3) the explained effect size was increased both for local and distant eQTLs; and (4) more importantly, the ratio of distant eQTLs increased from less than 45% to 72% with the same criterion, which implies that distant regulation might be more important than previously thought (Holloway et al., 2011; Battle et al., 2014; Bryois et al., 2014).

Some eQTL mapping studies (Holloway et al., 2011; Battle et al., 2014; Bryois et al., 2014) found more local eQTLs than distant eQTLs; however, distant regulation has been frequently proposed as a driver of phenotypic variation (especially for disease susceptibility in humans; Rotival et al., 2011; Westra et al., 2013). An earlier computational model also suggests that distant enhancer-bound proteins can significantly change the level of gene expression (Narula and Igoshin, 2010). The recently developed three-dimensional...
Table 2. List of Novel Loci and Candidate Genes for Oil Content Identified Using the New Integrated Map.

<table>
<thead>
<tr>
<th>Candidate gene</th>
<th>Chr</th>
<th>Position</th>
<th>Allele</th>
<th>MAF</th>
<th>eQTL</th>
<th>Location</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM5G814718</td>
<td>1</td>
<td>46 413 734</td>
<td>C/T</td>
<td>0.06</td>
<td>5.76 × 10^{-8}</td>
<td>NS</td>
<td>Genic Multicopper oxidase</td>
</tr>
<tr>
<td>GRMZM2G320325</td>
<td>1</td>
<td>55 071 146</td>
<td>A/T</td>
<td>0.06</td>
<td>5.78 × 10^{-7}</td>
<td>Genic</td>
<td>Uridine kinase</td>
</tr>
<tr>
<td>GRMZM2G100650</td>
<td>1</td>
<td>267 335 457</td>
<td>C/A</td>
<td>0.10</td>
<td>6.12 × 10^{-7}</td>
<td>Genic</td>
<td>Glycolipid transfer protein, GLTP</td>
</tr>
<tr>
<td>GRMZM2G425999</td>
<td>4</td>
<td>55 078 588</td>
<td>T/C</td>
<td>0.06</td>
<td>4.32 × 10^{-7}</td>
<td>Genic</td>
<td>Transmembrane transporter activity</td>
</tr>
<tr>
<td>GRMZM2G125556</td>
<td>4</td>
<td>141 969 034</td>
<td>G/A</td>
<td>0.11</td>
<td>4.66 × 10^{-7}</td>
<td>Genic</td>
<td>Glycolipid transfer protein, GLTP</td>
</tr>
<tr>
<td>GRMZM2G162670</td>
<td>4</td>
<td>142 046 103</td>
<td>T/A</td>
<td>0.11</td>
<td>2.15 × 10^{-7}</td>
<td>Genic</td>
<td>Glycolipid transfer protein, GLTP</td>
</tr>
<tr>
<td>GRMZM2G019358</td>
<td>4</td>
<td>228 628 353</td>
<td>C/G</td>
<td>0.07</td>
<td>3.13 × 10^{-8}</td>
<td>Genic</td>
<td>ATP binding</td>
</tr>
<tr>
<td>GRMZM2G139765</td>
<td>5</td>
<td>188 667 327</td>
<td>G/A</td>
<td>0.09</td>
<td>1.32 × 10^{-8}</td>
<td>Genic</td>
<td>Uridine kinase</td>
</tr>
<tr>
<td>GRMZM2G032095</td>
<td>7</td>
<td>96 823 747</td>
<td>C/A</td>
<td>0.06</td>
<td>6.53 × 10^{-7}</td>
<td>Genic</td>
<td>Catalytic activity</td>
</tr>
<tr>
<td>GRMZM2G127687</td>
<td>8</td>
<td>26 859 184</td>
<td>G/T</td>
<td>0.06</td>
<td>3.29 × 10^{-7}</td>
<td>Genic</td>
<td>ATP binding</td>
</tr>
<tr>
<td>GRMZM2G028570</td>
<td>8</td>
<td>71 717 794</td>
<td>A/C</td>
<td>0.05</td>
<td>3.46 × 10^{-7}</td>
<td>Genic</td>
<td>ATP binding</td>
</tr>
<tr>
<td>GRMZM2G029856</td>
<td>9</td>
<td>107 405 589</td>
<td>G/T</td>
<td>0.07</td>
<td>6.96 × 10^{-8}</td>
<td>Non-genic</td>
<td>UDP-glucose 4-epimerase</td>
</tr>
<tr>
<td>GRMZM2G098179</td>
<td>9</td>
<td>116 898 313</td>
<td>G/T</td>
<td>0.06</td>
<td>6.11 × 10^{-9}</td>
<td>Non-genic</td>
<td>Myb MYB30</td>
</tr>
</tbody>
</table>

**Table 2. List of Novel Loci and Candidate Genes for Oil Content Identified Using the New Integrated Map.**

- *A* candidate gene in the locus or the nearest annotated gene to the lead SNP.
- *b* Position according to version v2 of maize reference sequence.
- *c* P value for the SNP located within 100 kb of candidate gene. NS, not significant (P > 9.83 × 10^{-7}).
- *d* Each candidate gene is annotated according to MaizeGDB (Andorf et al., 2016).

eQTL Analysis Helps Reveal the Gene Regulatory Network and Genotype–Phenotype Relationship

With the recent accumulation of high-density genotypic data for large numbers of individuals across many species and related high-throughput phenotypic data, much effort is being devoted to exploring the genomic regions that underlie phenotypic changes for various traits. GWAS is a powerful approach for identifying candidate associations by examining the frequencies of different genotypes with respect to phenotypes, but it is insufficient for offering insight into biological mechanisms and defining the functions of the genes involved. Study of eQTLs could provide insights into the gene-expression effects of associated variants (Westra and Franke, 2014) and help to unravel the genotype–phenotype relationships. As shown in the cases of pF and the novel intergenic QTL on chromosome 4 with respect to flavonoid content, cob color, and oil concentration (Figures 4 and 6), understanding aspects of transcriptomic regulation can help define complex regulatory networks. This is especially true for those eQTL hotspots in which a number of genes are controlled by a common eQTL. Co-localizing eQTLs and QTLs could be important for exploring the genetic architecture of complex traits. For example, the study of plant development and phenotypic variation in *Populus* (Drost et al., 2010), combining association mapping and the co-expression network, resulted in the identification of candidate genes underlying glucosinolate traits (Chan et al., 2011). Studies such as these promise to increase knowledge of regulatory sequences and thereby allow for accurate mechanistic interpretations.

**METHODS**

**Plant Germplasm, RNA Sequencing, and Phenotyping**

The 540 maize inbred lines included in this study were from a global collection (Yang et al., 2011) including representative temperate and tropical/subtropical inbred lines. Detailed information on this panel can be found in Supplemental Table 4. A subset of 513 lines were genotyped (Yang et al., 2011) to collect data for approximately 100,000 single-nucleotide polymorphisms (SNPs) and 10,000 expression data points using the Illumina iSelect platform.
A total of 192 inbred lines were genotyped and 185 samples passed all QC obtained for a given line. Finally, 469 lines remained (Figure 1A).

The cutoff used to filter associated SNPs was $P = 7.97 \times 10^{-7}$ ($1/n$, where $n$ represents the number of SNPs). These steps were followed to identify eQTL regions. First, all significantly associated SNPs were grouped into clusters when the distance between two consecutive SNPs was $<10$ kb, and the clusters with at least five significant SNPs were regarded as candidate eQTLs, represented by their most significant SNP (named as lead SNP). Next, those candidate eQTLs in LD ($\rho \geq 0.1$) with other more significant candidates for the same gene were considered as false-positive associations introduced by intrinsic LD structure and were thus removed. The joint effect, estimated by multiple linear regression, of associated SNPs within each eQTL was then compared. When the significance of the candidate eQTLs in LD ($\rho \geq 0.1$) were equal, the eQTLs with larger joint effects were retained. The procedure for eQTL identification is similar to, but more rigorous than, the procedure used previously (Fu et al., 2013). To make the conclusions more reliable, we considered three methods to distinguish between local and distant eQTLs, from lenient to strict. First, as in the previous study, the eQTLs identified in this study (named “new” in main text and Figure 3C) were considered local if the lead SNP was
Critical Distant eQTLs and Non-coding Sequences in Maize

Molecular Plant

located within 20 kb of its targets and otherwise were considered distant. Second, eQTLs for the same traits located within 100 kb of each other were merged, and the remaining significant eQTLs were then considered local if the lead SNP was located within 100 kb of its targets, while all other QTLs were defined as distant (named as “new_100K_100K”). The most rigorous method was to merge eQTLs within 500 kb for the same traits and those with the most significant or largest effect were retained, and regarded as local if the lead SNP was located within 100 kb of its targets. The remaining ones were defined as distant (referred to as “new_500K_100K”). Results from “new_100K_100K” were used if there is no special explanation in the main text, and the full list of eQTL results are given in Supplemental Table 6.

To identify the distant hotspots, we applied a new local-scan statistical method (Silva et al., 2014). Different initial window sizes (5, 10, and 20 kb) were applied, the significance level of adjusted P value was set to 0.05, and 5 kb was finally used to achieve single gene level resolution. This method depends on the scanning window size and requires calibration. It shrinks the initial window as appropriate to detect and best define the hotspot size (Silva et al., 2014). Thus, the final hotspot regions are usually smaller than the initial window, sometimes down to a single SNP, which could be associated with several targets.

Epistatic Interaction Analysis

We have investigated whether the significant SNPs identified in the intergenic (e)QTL of chromosome 4 for oil concentration have epistatic interaction with the three candidates. Given each inspected variant pair (A versus B) for oil concentration \( Y \), linear regression was used to fit the model:

\[
Y = \beta_0 + \beta_1g_A + \beta_2g_B + \beta_3g_Ag_B
\]

where \( g_A \) and \( g_B \) are allele counts. Then the \( \beta_3 \) coefficients are tested for significance of epistatic interaction for A versus B. A linear regression-based test within plink (Purcell et al., 2007) was used in the implementation. The significance was measured as adjusted P value \( \leq 1 \times 10^{-4} \), and the distance between two examined variants less than 50 kb was excluded, of which the significant interactions could be likely caused by LD.

ACCESSION NUMBERS

The raw RNA-seq reads have been deposited in NCBI Sequence Read Archive (SRA) under accession SRR026161, and the GBS data have been deposited in the SRA with accession code SRR070875. The finally merged genotyping set (with hapmap format) and separately raw ones genotyped from different strategies are available at www.maizego.org/Resources.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at Molecular Plant Online.

FUNDING

This research was supported by the National Natural Science Foundation of China (31525017) and the National Key Research and Development Program of China (2016YFD0101001), the National Youth Top-notch Talent Support Program, and the Fundamental Research Funds for the Central Universities.

AUTHOR CONTRIBUTIONS

J.Y. designed and supervised this study. H.L., X.L., Y.X., and M.J. performed the data analysis. L.N., J.L., X.W., and W.L. contributed to materials collection. Q.Z. helped in GBS sequencing. L.C. helped to upload the sequenced data to NCBI. J.Y., H.L., and X.L. prepared the manuscript, and all authors critically read and approved the manuscript.

ACKNOWLEDGMENTS

We are grateful to the editor and two anonymous reviewers for their helpful comments and suggestions. No conflict of interest declared.

Received: March 29, 2016
Revised: June 23, 2016
Accepted: June 27, 2016
Published: July 2, 2016

REFERENCES


Feng, S., Cokus, S.J., Schubert, V., Zhai, J., Pellegrini, M., and Jacobsen, S.E. (2014). Genome-wide Hi-C analyses in wild-type and


Molecular Plant


Critical Distant eQTLs and Non-coding Sequences in Maize