CellPress
PARTNER JOURNAL

Molecular Plant
Research Article

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

Haijun Liu[1,2], Xin Luo[1,2], Luyao Niu[1], Yingjie Xiao[1], Lu Chen[1], Jie Liu[1], Xiaqing Wang[1], Minliang Jin[1], Wenqiang Li[1], Qinghua Zhang[1] and Jianbing Yan[1,*]

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

[2]These 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*

Liu H., Luo X., Niu L., Xiao Y., Chen L., Liu J., Wang X., Jin M., Li W., Zhang Q., and Yan J. (2017). Distant eQTLs and Non-coding Sequences Play Critical Roles in Regulating Gene Expression and Quantitative Trait Variation in Maize. Mol. Plant. **10**, 414–426.

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

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

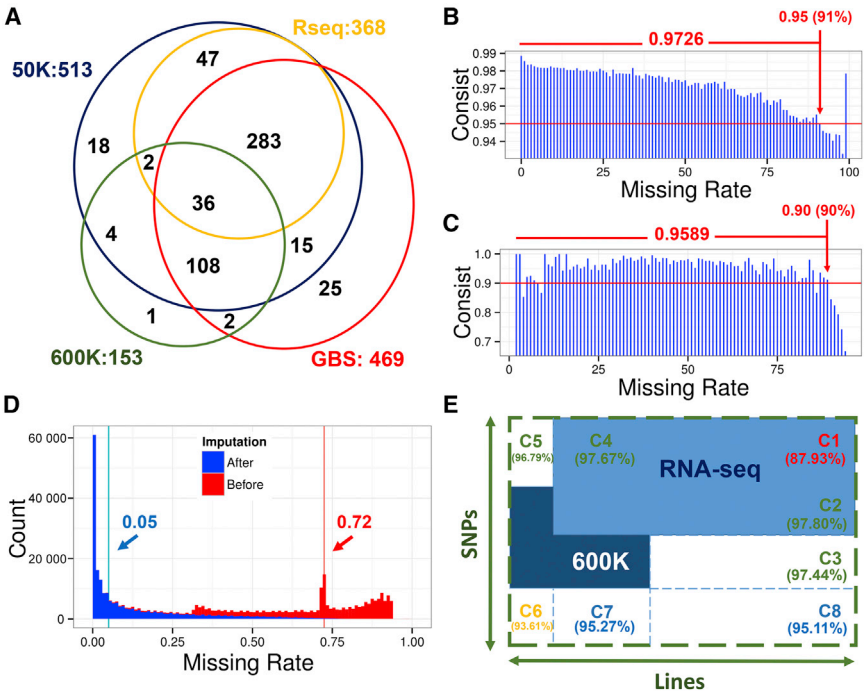

Figure 1. Creation of the Integrated Map.

(A) Numbers of individuals genotyped with different platforms. 50K, Illumina MaizeSNP50 array; 600K, Affymetrix Axiom Maize 600K array; Rseq, RNA sequencing for maize kernel (15 days after pollination [15DAP]); GBS, genotyping by sequencing.

(B) The consistency rates between original overlapped genotypes derives from 600K and RNA-seq with different missing rates of RNA-seq. The missing frequency less than 91% is selected, under which the coincident ratio is 97.26%.

(C) Accuracy rates for imputation with different missing rates of merged genotypes. The missing frequency less than 90% is retained, and the reliability for whole genotypes without missing larger than 90% is 95.89%.

(D) Distribution of missing rates before and after imputation.

(E) The contribution to imputation accuracy of different classifications based on different genotyping methods and panels of individuals. The proportion within each type represents the accuracy rates of imputation. As 50K and GBS covers most of the lines and the data are well distributed among loci, we assumed that they contributed uniformly both in loci and individually, and we mostly focused on the subset derived from 600K and RNA-seq.

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

Index	50K	600K	GBS	Rseq_raw[a]	Rseq_M91%[b]
50K	–	96.97%/100%	96.12%/99.55%	96.32%/100%	96.53%/100%
600K	19 495	–	98.48%/100%	97.13%/100%	97.53%/100%
GBS	5885	39 131	–	97.57%/100%	97.93%/100%
Rseq_raw	12 691	141 958	203 580	–	–
Rseq_M91%	11 363	123 948	161 352	–	–

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).
[a]Raw genotypes from RNA-seq.
[b]Raw genotypes for RNA-seq filtered by missing rates larger than 91%, which was included in further analysis.

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 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 genotyping 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).

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 ($P < 7.97 \times 10^{-7}$; $1/n$) 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 5). 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 for 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 6A). The distribution of lead SNPs for local eQTLs showed peaks at the 5′ and 3′ ends of genes (Supplemental Figure 6B), which was consistent with previous conclusions (Mazumder et al., 2003; Wilkie et al., 2003; Ringnér 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

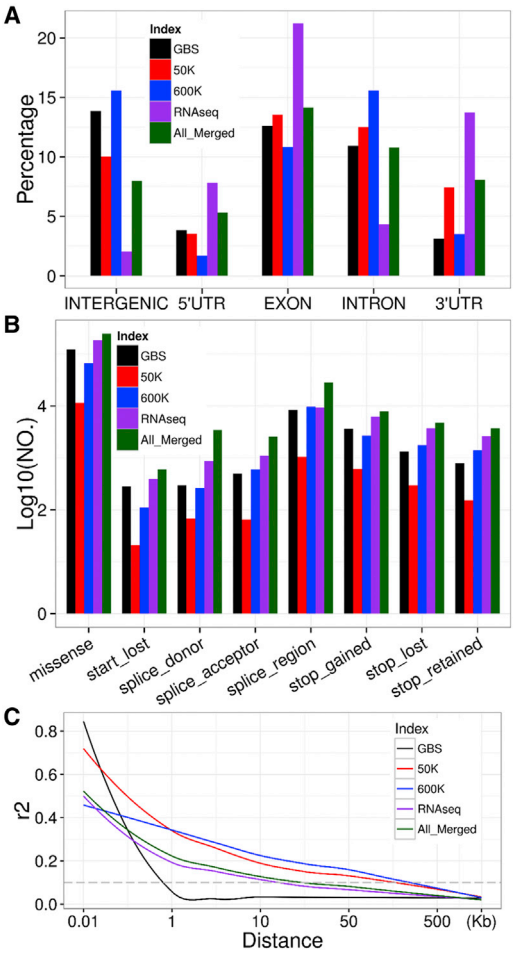

Figure 2. Features of the New Integrated Map.

(A) The integrated map has a more uniform physical distribution; it has a higher proportion of SNPs in genic regions than array-based methods and a higher proportion of SNPs in intergenic regions than RNA-seq.

(B) The integrated map identifies a larger number of variants that may cause severe phenotypic effects including loss of function.

(C) LD decay is intermediate, with a good balance between resolution and power. This ensures almost single gene resolution unlike array-based methods (50K and 600K) and confers high power to identify associated variants relative to sequence-based technologies (RNA-seq and GBS).

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 ($P_{adjusted}$ < 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

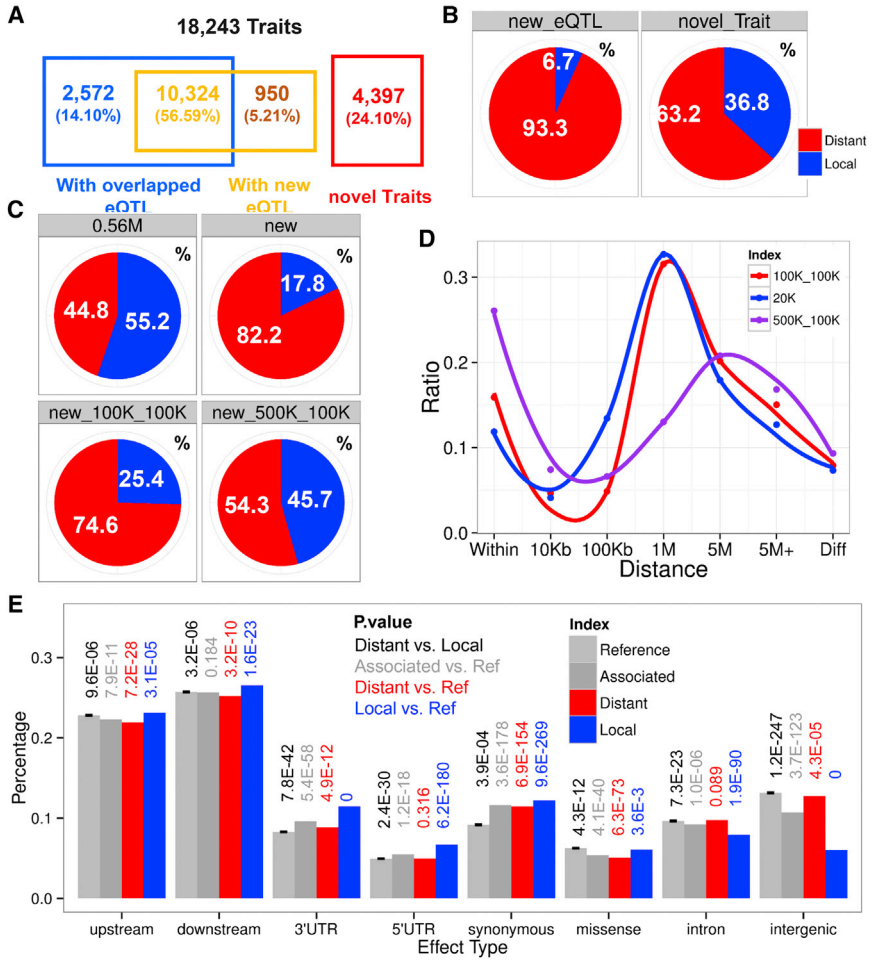

Figure 3. Distant Regulation and the Significance of Non-coding Regions.

(A) The identification of eQTLs by a new integrated map and its comparison with previous study (Fu et al., 2013). The blue rectangle represents the number of traits for which eQTLs were both identified, while the number in the yellow rectangle represents traits for which novel eQTLs were identified in the present study. Therefore, the number in blue font represents the traits for which eQTLs were consistently identified in both studies, the number in light yellow font represents the number of traits for which additional eQTLs were found, and the dark orange number represents traits with non-overlapping new eQTLs. Traits in the red rectangle (and red font) were those for which novel eQTLs were found in this study but not in the previous study.

(B) The ratio of distant and local eQTLs for additional eQTLs for traits that identified with eQTLs (called new eQTL) and for traits that were not found in eQTLs previously (marked as novel Trait; Fu et al., 2013) with the same threshold in the present study.

(C) The ratio of distant and local eQTLs at different thresholds. 0.56M shows the results from Fu et al. (2013); the other three methods used to distinguish local and distant eQTLs included the most lenient ("new," same as in Fu et al., 2013) and the most restrictive ("new_500K_100K," see Methods).

(D) Distribution of distance between lead SNPs and their regulated targets located on the same chromosome.

(E) Comparison of variant percentages in each effect type among distant eQTLs, local eQTLs, combined associated sets (local + distant), and the simulated distribution of reference random ones. A Chi-square goodness-of-fit test was used in comparison between ratios from local and distant eQTLs, and pnorm function in R was used to calculate the probabilities for each effect type for the other three classes to satisfy the random reference distribution.

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 1.96×10^{-6}, respectively; Figure 5C) and displayed a lower expression variability than genes being regulated (measured by coefficient of expression variation; $P = 2.25 \times 10^{-67}$ and 6.03×10^{-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

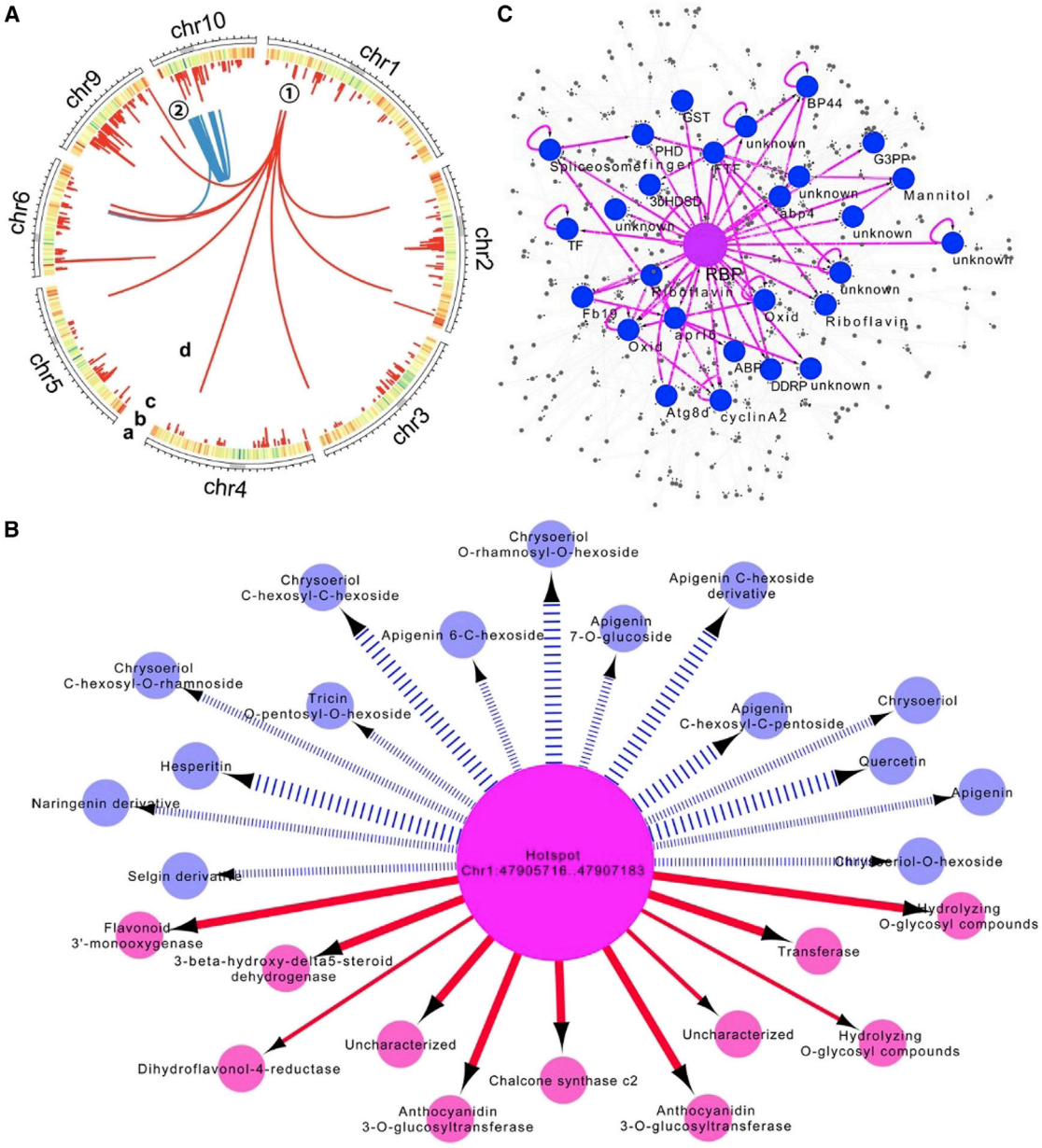

Figure 4. Distant-Acting eQTL Hotspots and the Identification of Complex Regulatory Networks.

(A) The identification of distant-acting hotspots. Two hotspots are displayed along with their regulated targets on the maize chromosome map, and the ① and ② represent the *p1* and *zmRBP* cases, respectively; centromeres are colored gray. b, heatmap showing the counts of targets within specific distant eQTL intervals; c, the histogram of significance (measured by −log(*P* value)) of each hotspot; d, links representing the association between the distant hotspot and their regulated genes.

(B) One of the distant-acting hotspots located in chromosome 1 (large pink circle, region from 47 905 716 to 47 907 183). The pink nodes represent distant regulated expressed genes and the light blue nodes represent associated cellular metabolites, all of which are flavonoids. The width of each edge was correlated with the significance of association (measured by square root of −log(*P* value)).

(C) One of the maize RNA-binding proteins (RBP, GRMZM2G171518) is characterized as a significant distant-acting hotspot, while it is directly associated with many other genes (blue nodes), which further regulate more genes with gray nodes.

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

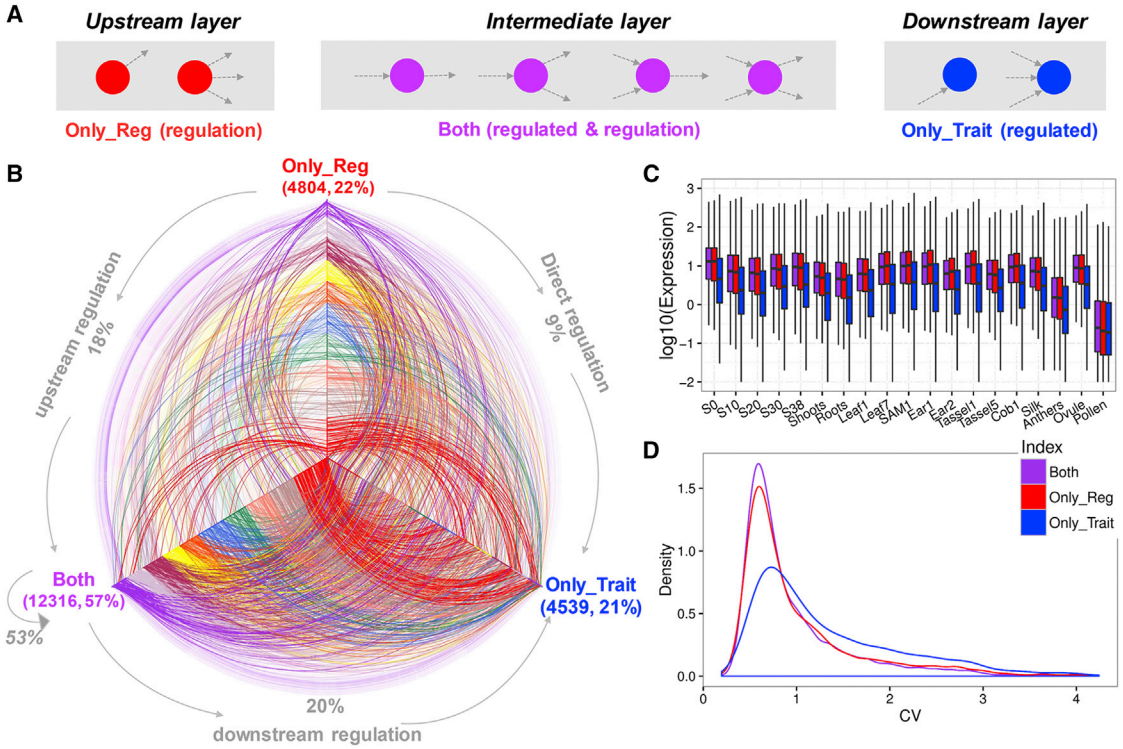

Figure 5. The Classification and Temporal-Spatial Expression of Genes in Different Regulation Layers.

(A) The expressed genes in this study were classified into three different regulation layers: only playing regulatory roles (Only_Reg, upstream layer), only to be regulated (Only_Trait, downstream layer), and as both regulator and to be regulated (Both, intermediate layer).

(B) The number and ratio of each layer and the percentage of different regulation layers. The three coordinates are subset genes classified into different regulation layers, while the different colors represent the different chromosomes (chromosomes 1 to 10 from inside to outside). Each link represents the regulation relationship and is colored as the corresponding chromosome from which eQTL is derived. To highlight the distant ones, we set the transparency for local regulated ones higher (with much more transparency) than distant regulations. Upstream and downstream regulation denote the regulation from upstream to intermediate layer and from intermediate to downstream layer, respectively. Direct regulation represents the direction of regulation from upstream layer to downstream layer.

(C) Comparison of gene-expression levels in different layers and tissues (tissue represented on the *x* axis).

(D) The distribution of the range of expression levels among different tissues (measured by coefficient of variation, CV), separately for genes from different regulation layers (expressed data for multiple tissues from Chen et al., 2014).

power (Figure 6 and Supplemental Figure 14). 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 (8.0×10^{-7} for new versus 1.8×10^{-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 ($P = 2.31 \times 10^{-11}$ for *FADD*, GRMZM2G066618; $P = 2.14 \times 10^{-19}$ for *GPI*, GRMZM2G162670; and $P = 5.37 \times 10^{-15}$ 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 ($P \leq 1 \times 10^{-4}$) 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; $P \leq 1 \times 10^{-4}$). 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

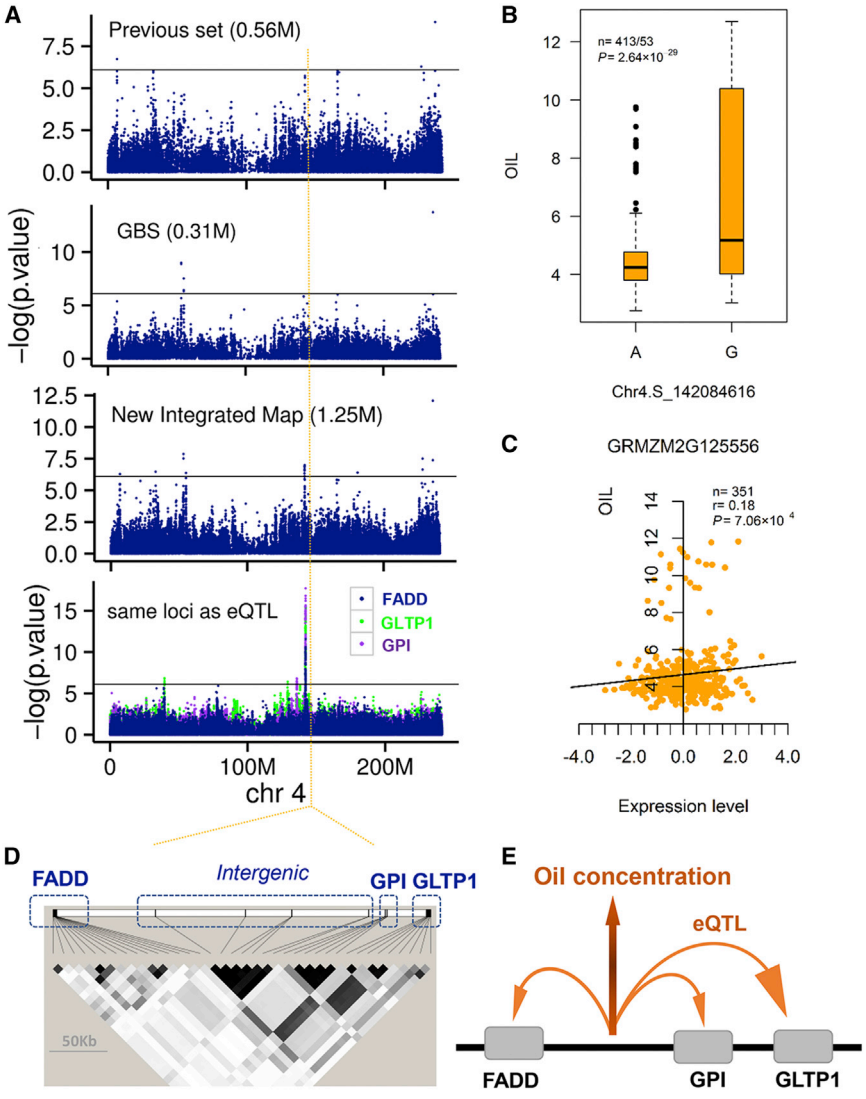

Figure 6. The Improved Statistical Power in GWAS and Co-localization of a Novel QTL for Kernel Oil Concentration with eQTLs for Three Oil-Related Genes.

(A) Comparison of GWAS results among genotypes from different platforms for kernel oil content. In the Manhattan plots, the dashed horizontal line represents the significance cutoff ($P = 9.83 \times 10^{-7}$). The novel QTL on chromosome 4 is significantly associated with kernel oil concentration and expression of three genes (*FADD*, GRMZM2G066618; *GPI*, GRMZM2G162670; and *GLTP1*, GRMZM2G125556).

(B) The divergence of oil concentrations between different alleles of the lead SNP.

(C) Positive correlation between gene-expression level (*GLTP1*, GRMZM2G125556 as a case) and oil concentration.

(D) Pairwise LD between associated SNPs and SNPs within three regulated genes. Triangles represent significant SNPs and squares represent candidate genes. The shading from black to white represents the decreasing intensity of LD.

(E) Illustration of the co-localized intergenic QTL in association with oil concentration and gene expression.

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

Candidate gene[a]	Chr	Position[b]	Allele	MAF	*P* value	eQTL[c]	Location	Annotation[d]
GRMZM5G814718	1	46 413 734	C/T	0.06	5.76×10^{-8}	NS	Genic	Multicopper oxidase
GRMZM2G320325	1	55 071 146	A/T	0.06	5.78×10^{-7}	8.46×10^{-22}	Genic	Uridine kinase
GRMZM2G100650	1	267 335 457	C/A	0.10	6.12×10^{-7}	7.92×10^{-9}	Genic	Glycolipid transfer protein, GLTP
GRMZM2G425999	4	55 075 588	T/C	0.06	4.32×10^{-7}	2.90×10^{-18}	Genic	Transmembrane transporter activity
GRMZM2G125556	4	141 969 034	G/A	0.11	4.66×10^{-7}	5.37×10^{-15}	Non-genic	Glycolipid transfer protein, GLTP
GRMZM2G162670	4	142 046 103	T/A	0.11	2.15×10^{-7}	2.14×10^{-19}	Non-genic	GPI anchor
GRMZM2G019358	4	228 628 353	C/G	0.07	3.13×10^{-8}	1.08×10^{-13}	Genic	Unknown
GRMZM2G139765	5	188 667 327	G/A	0.09	1.32×10^{-8}	9.20×10^{-13}	Genic	Transcription factor
GRMZM2G032095	7	96 823 747	C/A	0.06	6.53×10^{-7}	NS	Non-genic	Catalytic activity
GRMZM2G127687	8	26 859 184	G/T	0.06	3.29×10^{-7}	5.98×10^{-18}	Geneic	ATP binding
GRMZM2G028570	8	71 717 794	A/C	0.05	3.46×10^{-7}	NS	Genic	Transporter activity
GRMZM2G029856	9	107 405 589	G/T	0.07	6.96×10^{-8}	3.57×10^{-9}	Non-genic	UDP-glucose 4-epimerase
GRMZM2G098179	9	116 898 313	G/T	0.06	6.11×10^{-9}	1.26×10^{-9}	Non-genic	Myb MYB30

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

[a]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 ($P > 9.83 \times 10^{-7}$).

[d]Each candidate gene is annotated according to MaizeGDB (Andorf et al., 2016).

genomic architecture technology (Feng et al., 2014; Rao et al., 2014) sheds new light on the interactions between enhancers and their target promoters and aids in identifying the genetic and epigenetic regulatory elements, and, thus, the underlying long-range regulatory mechanisms (Noonan and McCallion, 2010).

As regular practice for RNA-seq, the reads from samples were mapped to the reference genome to obtain expression quantification; thus the substantial genomic variation among diverse individuals would potentially cause mapping differences, which would consequently affect local eQTL mapping. It was found in the present study that the individuals with reference allele indeed "expressed" significantly higher than non-reference ones in general. Aligning reads to variant-corrected reference genomes could bring more accuracy to both expression quantification and eQTL mapping, and should be considered in future studies.

Non-coding Regulatory Sequences in Maize

Several studies have disclosed the mechanisms by which non-coding regions functionally contribute to disease (Kleinjan and van Heyningen, 2005; Visel et al., 2009). It was also found that TASs (trait-associated SNPs) were enriched in non-genic regions (Freedman et al., 2011; Li et al., 2012) and that nearly half of the large number of *trans*-acting eQTLs associated with splicing are located in non-genic regions in maize (Thatcher et al., 2014). This study provides an opportunity to understand the distribution of these putative regulators on the genome-wide level. It was found that, for distant eQTLs, 42% of lead SNPs were located in the non-genic regions (78% in non-coding sequences), which might have been missed in the previous study (Fu et al., 2013). The maize genome contains a large proportion of repetitive elements, including retrotransposons and transposons, which are often found to be expressed and involved in expression regulation (Gebert and Rosenkranz, 2015).

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 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 *p1* 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) with the Illumina MaizeSNP50 array (Ganal et al., 2011). Poly-A^+ RNA, collected from the whole kernels at 15 days after pollination, was sequenced from 368 selected diverse lines, and 560K SNPs with MAF >0.05 were previously identified (Fu et al., 2013). A total of 28 769 genes expressed in more than 50% of the inbred lines were used for further analyses, and the overall distribution of expression levels for each gene was normalized using a normal quantile transformation (qqnorm function in R, http://www.r-project.org). More details on library construction, sequencing, SNP detection, positive control of SNP accuracy, and quantile normalization of expression are provided as additional notes in Supplemental Information. In the present study, 469 and 153 lines were further genotyped by GBS (Elshire et al., 2011) and Affymetrix Axiom Maize 600K array (Unterseer et al., 2014), respectively (Figure 1A). The oil concentration of the mature kernel was measured in multiple environments, as described in the previous study (Li et al., 2013a).

Construction of Reduced Representation Libraries and GBS

Nucleic DNA extraction from young leaves (0.4 g) followed the method of Murray and Thompson (1980). DNA concentrations were normalized to 10 ng/μl. *Ape*KI (NEB, R0643L) restriction enzyme was used to digest DNA at 75°C for 2 h. DNA fragments were ligated to adapters with different indexes for each line using T4 DNA Ligase (NEB, MK0202L). After cleanup, a DNA panel including 96 lines was pooled into one GBS library and 16 cycles of PCR amplification were performed. Size selection was done using the MinElute Gel Extraction Kit (Qiagen, Germany). DNA concentrations of the libraries were measured using quantitative PCR (Bio-Rad, California). Cluster generation was performed on a cBot (program: SR_Amp_Lin_Block_Hyb_v8.0, Illumina) using a flow cell v3 and reagents from TruSeq SR Cluster Kit v3 (Illumina) according to the manufacturer's instructions. DNA sequencing was performed on a HiSeq2000, equipped with on-instrument HCS version 1.4.8 and Real-Time Analysis version 1.12.4.2 (Illumina). Sequencing was performed in paired-end mode with a 100-bp read length.

SNP Allele Calling

GBS

The Java program TASSEL was used to call the variants (v3.0; Glaubitz et al., 2014). More than 1.25 billion reads were obtained from 522 lines (Supplemental Table 5) with an average 2.74M reads per line covering 9% of the genome coverage. Sequenced reads from each line were then pooled together to define a tag (at least 5 reads), and 15 499 006 tags were obtained. BWA (Li and Durbin, 2009) was used to map the tags to the B73 reference genome (AGPv2, FGS 5b; Schnable et al., 2009). 57.3% (or 8 881 958) of tags were mapped to the reference genome and 69.3% (or 6 152 365) of these were uniquely mapped and were used for SNP calling with default parameters (Glaubitz et al., 2014). Index sequences were used to distinguish different individuals. In total, 670 412 SNPs were acquired with an average missing rate of 69.2%. The outlier lines were excluded from further analysis based on the following standards: (1) compared with the maize SNP 50K data, the consistency was less than 90%; or (2) fewer than 10K SNPs were obtained for a given line. Finally, 469 lines remained (Figure 1A).

600K

A total of 192 inbred lines were genotyped and 185 samples passed all QC metrics, including low call rates, and low-quality and possibly contaminated samples. Genotypes were produced by the AxiomGTv1 algorithm (Nicolazzi et al., 2014) with inbred penalty and generic prior-model clustering, and the average reproducibility among eight sets of hidden replicates was 99.8%. SNPolisher (Nicolazzi et al., 2014) was then used to classify the QC SNP genotype into six different types, and the type of polymorphic high resolution (PHR) and PHR reidentified from an optional off-target variant calling algorithm were both retained and merged for the next analysis, to obtain a total of 503 030 SNPs. A filtration similar to that used with GBS was applied, and 153 lines were used for further analysis (Figure 1A).

Creation of the Integrated Map

After strict quality controls for each dataset, the genotypes from four different genotyping platforms were merged, and in cases where the different platforms disagreed about specific loci, priority was given in this order: 600K > 50K > RNA-seq > GBS. Beagle (v4.0; Browning and Browning, 2007) was then used to perform genotype imputation. Markers from chromosome 10 were used to select the best parameters for Beagle and to evaluate the accuracy of imputation, and 15 000 (~3% of total within chromosome 10) randomly selected known genotypes (loci × individual) were masked as missing. The reliability of imputation was evaluated by comparing the known and imputed genotypes. Various parameters were tested, and the best parameter for this study was determined to be: window = 50 000, overlap = 5000, ibd = true. With these parameters, the average accuracy of imputation was 96.93% when SNPs with a high missing rate (>90%) were excluded before imputation (removing-then-imputation strategy). An alternative strategy, imputation-then-removing, in which those SNPs with a high missing rate (>90%) are excluded after imputation, resulted in a lower accuracy rate of 95.89%. Thus the former strategy was used for final imputation analysis. Finally, the integrated map, with more than 2.65M loci, was obtained for 540 individuals, 1.25M of which had a MAF ≥5% and were used for further studies. The finally merged genotyping set (with hapmap format) is available at www.maizego.org/Resources.

SNP Validation and Annotation

To evaluate the reliability of the newly integrated map, we collected the variants identified from resequenced PCR products by different laboratories working on this same population. In all, 477 independently genotyped loci within the subset of this panel were evaluated, and the consistency reached 94.53% (Supplemental Table 1). Most of the inconsistencies were located within complex regions such as tandem repeats (Supplemental Figure 4). The MAF of 307 among 477 loci was greater than 5%. The effect of each variant was annotated by the SnpEff program (Cingolani et al., 2012) to classify genotypes based on different genomic regions and different effect consequences. The Chi-square goodness-of-fit test of the ratio of different types of SNP effects (missense, intergenic, etc.) was performed.

Association Analysis

The merged SNPs dataset (MAF >0.05) was used to perform eQTL analysis for each gene, based on the LMM (Yu et al., 2006) implemented in the R package EMMA (Kang et al., 2008), where the population structure, kinship matrix, and other hidden confounding factors were fitted to control false-positive associations. All are similar to those of the previous study as described by Fu et al. (2013), so the results are comparable.

Identification of eQTLs and Hotspots

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 ($r^2 \geq 0.1$) 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 ($r^2 \geq 0.1$) 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

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:

$$Y = \beta_0 + \beta_1 g_A + \beta_2 g_B + \beta_3 g_A g_B$$

where g_A and g_B are allele counts. Then the β_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 those pairs with *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 SRP026161, and the GBS data have been deposited in the SRA with accession code SRP070875. 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

Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. **16**:197–212.

Ambrosone, A., Costa, A., Leone, A., and Grillo, S. (2012). Beyond transcription: RNA-binding proteins as emerging regulators of plant response to environmental constraints. Plant Sci. **182**:12–18.

Andorf, C.M., Cannon, E.K., Portwood, J.L., 2nd, Gardiner, J.M., Harper, L.C., Schaeffer, M.L., Braun, B.L., Campbell, D.A., Vinnakota, A.G., Sribalusu, V.V., et al. (2016). MaizeGDB update: new tools, data and interface for the maize model organism database. Nucleic Acids Res. **44**:D1195–D1201.

Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. **24**:14–24.

Britten, R.J., and Davidson, E.H. (1969). Gene regulation for higher cells: a theory. Science **165**:349–357.

Browning, R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. **81**:1084–1097.

Bryois, J., Buil, A., Evans, D.M., Kemp, J.P., Montgomery, S.B., Conrad, D.F., Ho, K.M., Ring, S., Hurles, M., Deloukas, P., et al. (2014). Cis and trans effects of human genomic variants on gene expression. PLoS Genet. **10**:e1004461.

Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. **134**:25–36.

Chan, E.K., Rowe, H.C., Corwin, J.A., Joseph, B., and Kliebenstein, D.J. (2011). Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. PLoS Biol. **9**:1713.

Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A., and Lai, J. (2014). Dynamic transcriptome landscape of maize embryo and endosperm development. Plant Physiol. **166**:252–264.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) **6**:80–92.

Dreyfuss, G., Kim, V.N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat. Rev. Mol. Cell Biol. **3**:195–205.

Drost, D.R., Benedict, C.I., Berg, A., Novaes, E., Novaes, C.R., Yu, Q., Dervinis, C., Maia, J.M., Yap, J., Miles, B., et al. (2010). Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of *Populus*. Proc. Natl. Acad. Sci. USA. **107**:8492–8497.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. **6**:e19379.

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

mutants reveal high-resolution chromatin interactions in *Arabidopsis*. Mol. Cell. **55**:694–707.

Freedman, M.L., Monteiro, A.N., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., et al. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. Nat. Genet. **43**:513–518.

Fu, J., Cheng, Y., Linghu, J., Yang, X., Kang, L., Zhang, Z., Zhang, J., He, C., Du, X., Peng, Z., et al. (2013). RNA sequencing reveals the complex regulatory network in the maize kernel. Nat. Commun. **4**:2832.

Ganal, M.W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E.S., Charcosset, A., Clarke, J.D., Graner, E.M., Hansen, M., Joets, J., et al. (2011). 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.

Gebert, D., and Rosenkranz, D. (2015). RNA-based regulation of transposon expression. Wiley Interdiscip. Rev. RNA. **6**:687–708.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., and Buckler, E.S. (2014). TASSEL-gbs: a high capacity genotyping by sequencing analysis pipeline. PLoS One. **9**:e90346.

He, J., Zhao, X., Laroche, A., Lu, Z.X., Liu, H., and Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci. **5**:484.

Holloway, B., Luck, S., Beatty, M., Rafalski, J.A., and Li, B. (2011). Genome-wide expression quantitative trait loci (eQTL) analysis in maize. BMC Genomics. **12**:336.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics **178**:1709–1723.

Kleinjan, D.A., and Lettice, L.A. (2008). Long-range gene control and genetic disease. Adv. Genet. **61**:339–388.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am. J. Hum. Genet. **76**:8–32.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. **25**:1754–1760.

Li, X., Zhu, C., Yeh, C.T., Wu, W., Takacs, E.M., Petsch, K.A., Tian, F., Bai, G., Buckler, E.S., Muehlbauer, G.J., et al. (2012). Genic and nongenic contributions to natural variation of quantitative traits in maize. Genome Res. **22**:2436–2444.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., et al. (2013a). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. **45**:43–50.

Li, L., Petsch, K., Shimizu, R., Liu, S., Xu, W., Ying, K., Yu, J., Scanlon, M.J., Schnable, P.S., Timmermans, M.C., et al. (2013b). Mendelian and non-Mendelian regulation of gene expression in maize. PLoS Genet. **9**:e1003202.

Li, L., Eichten, S.R., Shimizu, R., Petsch, K., Yeh, C.T., Wu, W., Chettoor, A.M., Givan, S.A., Cole, R.A., Fowler, J.E., et al. (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol. **15**:R40.

Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M.X., Arondel, V., Bates, P.D., Baud, S., Bird, D., Debono, A., Durrett, T.P., et al. (2010). Acyl-lipid metabolism. Arabidopsis Book **8**:e0133.

Liu, H., Wang, X., Warburton, M.L., Wen, W., Jin, M., Deng, M., Liu, J., Tong, H., Pan, Q., Yang, X., et al. (2015). Genomic, transcriptomic, and phenomic variation reveals the complex adaptation of modern maize breeding. Mol. Plant **8**:871–884.

Loizides-Mangold, U., David, F.P., Nesatyy, V.J., Kinoshita, T., and Riezman, H. (2012). Glycosylphosphatidylinositol anchors regulate glycosphingolipid levels. J. Lipid Res. **53**:1522–1534.

Lorković, Z.J. (2009). Role of plant RNA-binding proteins in development, stress response and genome organization. Trends Plant Sci. **14**:229–236.

Maeda, Y., Tashima, Y., Houjou, T., Fujita, M., Yoko-o, T., Jigami, Y., Taguchi, R., and Kinoshita, T. (2007). Fatty acid remodeling of GPI-anchored proteins is required for their raft association. Mol. Biol. Cell. **18**:1497–1506.

Mazumder, B., Seshadri, V., and Fox, P.L. (2003). Translational control by the 3′-UTR: the ends specify the means. Trends Biochem. Sci. **28**:91–98.

Mooney, M., Bond, J., Monks, N., Eugster, E., Cherba, D., Berlinski, P., Kamerling, S., Marotti, K., Simpson, H., Rusk, T., et al. (2013). Comparative RNA-seq and microarray analysis of gene expression changes in B-cell lymphomas of *Canis familiaris*. PLoS One. **8**:e61088.

Morohashi, K., Casas, M.I., Falcone Ferreyra, M.L., Mejía-Guerra, M.K., Pourcel, L., Yilmaz, A., Feller, A., Carvalho, B., Emiliani, J., Rodriguez, E., et al. (2012). A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. Plant Cell. **24**:2745–2764.

Murray, M.G., and Thompson, W.F. (1980). Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. **8**:4321–4325.

Narula, J., and Igoshin, O.A. (2010). Thermodynamic models of combinatorial gene regulation by distant enhancers. IET Syst. Biol. **4**:393–408.

Nicolazzi, E.L., Iamartino, D., and Williams, J.L. (2014). AffyPipe: an open-source pipeline for Affymetrix Axiom genotyping workflow. Bioinformatics. **30**:3118–3119.

Noonan, J.P., and McCallion, A.S. (2010). Genomics of long-range regulatory elements. Annu. Rev. Genomics Hum. Genet. **11**:1–23.

Nuruzzaman, M., Sharoni, A.M., and Kikuchi, S. (2013). Roles of NAC transcription factors in the regulation of biotic and abiotic stress responses in plants. Front Microbiol. **4**:248.

Okuley, J., Lightner, J., Feldmann, K., Yadav, N., Lark, E., and Browse, J. (1994). *Arabidopsis* FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. Plant Cell. **6**:147–158.

Olsen, A.N., Ernst, H.A., Leggio, L.L., and Skriver, K. (2005). NAC transcription factors: structurally distinct, functionally diverse. Trends Plant Sci. **10**:79–87.

Panoutsopoulou, K., Tachmazidou, I., and Zeggini, E. (2013). In search of low-frequency and rare variants affecting complex traits. Hum. Mol. Genet. **22**:R16–R21.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. **81**:559–575.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. **159**:1665–1680.

Ringnér, M., and Krogh, M. (2005). Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. PLoS Comput. Bio. **1**:e72.

Rotival, M., Zeller, T., Wild, P.S., Maouche, S., Szymczak, S., Schillert, A., Castagné, R., Deiseroth, A., Proust, C., Brocheton, J., et al. (2011). Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet. **7**:e1002367.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326**:1112–1115.

Silva, I.T., Rosales, R.A., Holanda, A.J., Nussenzweig, M.C., and Jankovic, M. (2014). Identification of chromosomal translocation hotspots via scan statistics. Bioinformatics. **30**:2551–2558.

Thatcher, S.R., Zhou, W., Leonard, A., Wang, B.B., Beatty, M., Zastrow-Hayes, G., Zhao, X., Baumgarten, A., and Li, B. (2014). Genome-wide analysis of alternative splicing in *Zea mays*: landscape and genetic regulation. Plant Cell. **26**:3472–3487.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., Meitinger, T., Strom, T.M., Fries, R., Pausch, H., et al. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600k SNP genotyping array. BMC Genomics. **15**:823.

Visel, A., Rubin, E.M., and Pennacchio, L.A. (2009). Genomic views of distant-acting enhancers. Nature **461**:199–205.

Van Heyningen, V., and Bickmore, W. (2013). Regulation from a distance: long-range control of gene expression in development and disease. Philos. Trans. R. Soc. Lond. B Biol. Sci. **368**:20120372.

Wang, C., Gong, B., Bushel, P.R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., et al. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat. Biotechnol. **32**:926–932.

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., et al. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat. Commun. **5**:3438.

Westra, H.J., and Franke, L. (2014). From genome to function by studying eQTLs. Biochim. Biophys. Acta **1842**:1896–1902.

Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. **45**:1238–1243.

Wilkie, G.S., Dickson, K.S., and Gray, N.K. (2003). Regulation of mRNA translation by 5′- and 3′-UTR-binding factors. Trends Biochem. Sci. **28**:182–188.

Xiang, J.F., Yin, Q.F., Chen, T., Zhang, Y., Zhang, X.O., Wu, Z., Zhang, S., Wang, H.B., Ge, J., Lu, X., et al. (2014). Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Res. **24**:513–531.

Yan, J., Warburton, M., and Crouch, J. (2011). Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. Crop Sci. **51**:433–449.

Yang, X.H., Gao, S.B., Xu, S.T., Zhang, Z.X., Prasanna, B.M., Li, L., Li, J.S., and Yan, J.B. (2011). Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. Mol. Breed. **28**:511–526.

Yang, N., Lu, Y., Yang, X., Huang, J., Zhou, Y., Ali, F., Wen, W., Liu, J., Li, J., and Yan, J. (2014). Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. Plos Genet. **10**:e1004573.

Yu, J., Pressoir, G., Briggs, W.H., Vroh, Bi I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38**:203–208.

Zhang, L., Chia, J.M., Kumari, S., Stein, J.C., Liu, Z., Narechania, A., Maher, C.A., Guill, K., McMullen, M.D., and Ware, D. (2009). A genome-wide characterization of microRNA genes in maize. PLoS Genet. **5**:e1000716.