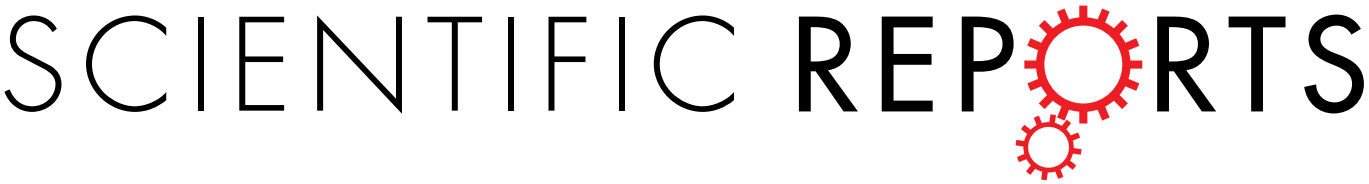

OPEN

Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation

Received: 07 October 2015
Accepted: 01 December 2015
Published: 05 January 2016

Minliang Jin[1,*], Haijun Liu[1,*], Cheng He[2,3,*], Junjie Fu[3], Yingjie Xiao[1], Yuebin Wang[1], Weibo Xie[1], Guoying Wang[3] & Jianbing Yan[1]

Gene expression variation largely contributes to phenotypic diversity and constructing pan-transcriptome is considered necessary for species with complex genomes. However, the regulation mechanisms and functional consequences of pan-transcriptome is unexplored systematically. By analyzing RNA-seq data from 368 maize diverse inbred lines, we identified almost one-third nuclear genes under expression presence and absence variation, which tend to play regulatory roles and are likely regulated by distant eQTLs. The ePAV was directly used as "genotype" to perform GWAS for 15 agronomic phenotypes and 526 metabolic traits to efficiently explore the associations between transcriptomic and phenomic variations. Through a modified assembly strategy, 2,355 high-confidence novel sequences with total 1.9 Mb lengths were found absent within reference genome. Ten randomly selected novel sequences were fully validated with genomic PCR, including another two NBS_LRR candidates potentially affect flavonoids and disease-resistance. A simulation analysis suggested that the pan-transcriptome of the maize whole kernel is approaching a maximum value of 63,000 genes, and through developing two test-cross populations and surveying several most important yield traits, the dispensable genes were shown to contribute to heterosis. Novel perspectives and resources to discover maize quantitative trait variations were provided to better understand the kernel regulation networks and to enhance maize breeding.

Maize shows an amazing degree of phenotypic variation due to the outcrossing nature, and to natural and artificial selection during the rapid worldwide population expansion[1]. Phenotypic variation has been explored by QTL mapping and genome-wide association studies (GWAS)[2]. As it becomes clear that the differences in transcript abundance are a major contributor to phenotypic evolution[3,4], allelic variation effects on the transcriptome, which reflect both genetic and epigenetic regulation, should be explored at a genome-wide level[5].

Presence/absence genomic sequence variation (PAV) is important in reshaping individual performance[6]. PAV at the genomic level would be reflected in the transcriptome ePAV (expression Presence and Absence Variation). The ePAV not only reflect genomic structural variation, but also the variations in genetic and epigenetic regulatory elements. Thus it is essential to characterize the ePAV genes and their possible functions.

Most genome-wide genetic studies focus the genetic elements present in the reference genome. It is now recognized that a portion of the genomic content is only present in a subset of individuals within a species, (termed the dispensable genome) especially in diverse species, such as maize. The genome-wide comparison between B73 and Mo17[6] and within an expanded panel including teosinte (ancestral maize) lines[7] demonstrated that a considerable portion of the genome (~50%) was not shared. The widespread dispensable genes, i.e those showing present/absent variation, have been proposed to be important for phenotypic diversity in inbred collections and for heterotic performance in hybrids[8,9].

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China. [2]College of Agriculture and Biotechnology, China Agricultural University, Beijing 100193, China. [3]Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to F.J. (email: fujunjie@caas.cn) or Y.J. (email: yjianbing@mail.hzau.edu.cn)

The rapid development of next generation sequencing technology and the decrease in cost provide us an opportunity to sequence many individuals within a species to build up the pan genome, or the sequences which, taken as a whole from all individuals, define a species. RNA sequencing (RNA-seq) has been successfully used to define the transcriptome and to find novel transcripts absent from the reference genome[10]. Compared to genome sequencing, RNA-seq is more economical, especially in the exploration of the complex maize genome containing more than 85% repetitive sequences[11]. The construction of the maize pan-transcriptome is especially useful for the discovery of functional dispensable genes. Recently, the maize pan-transcriptome and its diversity have been studied in diverse lines[9,12], however, we still lack knowledge about many dispensable gene function at the genome-wide level.

Here, with the help of deep RNA-seq of kernels at 15 DAP in a diverse panel with 368 inbred lines[5], we characterized the extreme variation at the transcript level (ePAV), relative to the reference genome, and performed association studies between ePAVs and more than 600 quantitative traits. By *de novo* assembly, we also constructed the maize pan-transcriptome and explored its contribution to phenotypic and transcriptomic diversity.

Results

Expression presence/absence is prevalent and trans-regulated. Gene expression levels of the annotated genes in the B73 reference genome were quantified using RNA-seq of maize kernels in 368 inbred lines[5]. We define the expression level differences between subsets of individuals at the given tissue or developmental stage as a polymorphism at the transcription level: expression present/absent variation (ePAV). By filtering the genes showing expression in less than 19 inbred lines or more than 348 inbred lines (MAF $\leq$ 5%) and applying an adapted distribution-based measure with no subjective set cutoff (see Methods), 13,382 nuclear genes among 38,032 total with ePAVs were obtained (5% $\leq$ MAF $\leq$ 95%). Among them, 6,656 (49.9%) were not explored in a previous study[5] since they were expressed in less than 50% but great than 5% of the inbred lines (see Supplementary Fig. S1 online).

Almost half (46%, 6,726) of the ePAV genes expressed in more than 50% of the inbred lines have been clearly identified as regulated by expression quantitative trait loci (eQTLs) in the previous study[5]. The ePAV genes were more likely to be regulated by distant eQTLs when compared with non-ePAV genes (also called core expression genes, expressed in more than 95% of the lines; $P < 2.2E\text{-}6$, $\chi 2$ test; Fig. 1a). The effects of local eQTL were found to be greater than distant eQTL both for ePAV (P = 7.05E-22) and non-ePAV genes (P = 1.92E-135; Fig. 1b). The eQTL effects for ePAV genes were greater than those for non-ePAV genes in both local (P = 1.34E-18) and distant (P = 7.18E-56) types. The ePAV genes were enriched in regulation-related processes, while the non-ePAV genes tended to play roles as structural genes (Fig. 1c; Supplementary Table S1 online). The dominant regulation by distant eQTLs and defined as regulators indicate the ePAVs may act as intermediate regulators to downstream genes, which mostly consist of non-ePAV (or core) genes. This is supported by the observation that most (92%; P < 2.2E-16) of the potential regulation targets of ePAV genes were non-ePAV genes. Additionally, the non-ePAV genes tend to be regulated by distant eQTLs as well, with up to 81.2% of regulated non-ePAV genes located on different chromosomes than their ePAV regulators, and for those located on the same chromosome, 86% were separated by over 20 Mb (Fig. 1d). All the above suggest that the dispensable expression genes are functionally essential, and play key roles in the intermediate regulation layer.

PAV contribute rarely to the causation of ePAV. Before using ePAV for further analysis, we confirmed that the undetectable gene expression in a given tissue was not due to sequencing bias or low sequencing coverage. PAV gene expression should always show ePAV patterns that provide excellent samples to test the reliability of ePAV detection. A ~2.4 Mb fragment on chromosome 6 is present in B73 but absent in the Mo17 genome where 62 genes were annotated[6]. This region was also confirmed by PCR in our inbred lines, among which 209 lines had the same haplotype of B73 and another 15 lines were consistent with Mo17 (see Supplementary Table S2 online). Among the 62 genes within this region, 61 were detected by RNA-seq and 52 of them were considered as ePAV genes based on the standard (expressed more than 5% and less than 95% lines). The consistency between ePAV status of the 52 ePAV candidates and PCR validation at the genomic level was 74%. This suggests that the ePAV label is acceptable in that most (96.4%) of the inconsistencies were likely caused by non-expression in kernel tissue with presence in DNA sequence, and that the frequency of apparent expression without sequence evidence was rare, at 3.6%.

To determine how many of the ePAVs are caused by genomic PAVs, the reference genome B73 and deep sequenced genome Mo17 were compared. Only 54 (~1%) of the identified 5,838 ePAV genes were supported as sequence PAVs by the re-sequencing results of Mo17 (Supplementary Fig. S2 online)[8]. However, these two inbreds represent only a fraction of the total maize sequence diversity. Therefore, we used genotyping data generated from Illumina MaizeSNP50 array (50 K) for the whole panel[13] and from the Affymetrix® Axiom® Maize Genotyping Array (600 K)[14] for 38 lines; again, we found ~ 1% (102 and 122, or 0.76% and 0.91% for 50 K and 600 K datasets, respectively) of ePAVs were predicted as PAVs (see Methods). These results together imply that only a small proportion (~1%) of ePAVs were due to PAV in the genomic sequence and therefore, most were likely to be the result of suppression at the expression level.

We further chose 10 putative PAV genes in ePAVs for experimental validation in a subset of 96 inbred lines by genomic PCR. All ten ePAV genes represent genomic PAV genes. The consistency of the ePAV and PAV labels detected by PCR in the 96 lines ranged from 70% to 89% (see Supplementary Fig. S3 and Supplementary Table S3 online), which provided an estimate of the reliability of the predicted ePAV correspondence to sequenced PAVs.

Novel expressed sequence discovery from *de novo* assembly. RNA-seq reads from each inbred were *de novo* assembled to detect the novel expressed sequences and construct the maize-transcriptome (Supplementary Table S4 online). We applied Trinity[15] to detect novel sequences by comparing the two strategies: "align-then-assemble" and "assemble-then-align"[10] (detail in Methods).

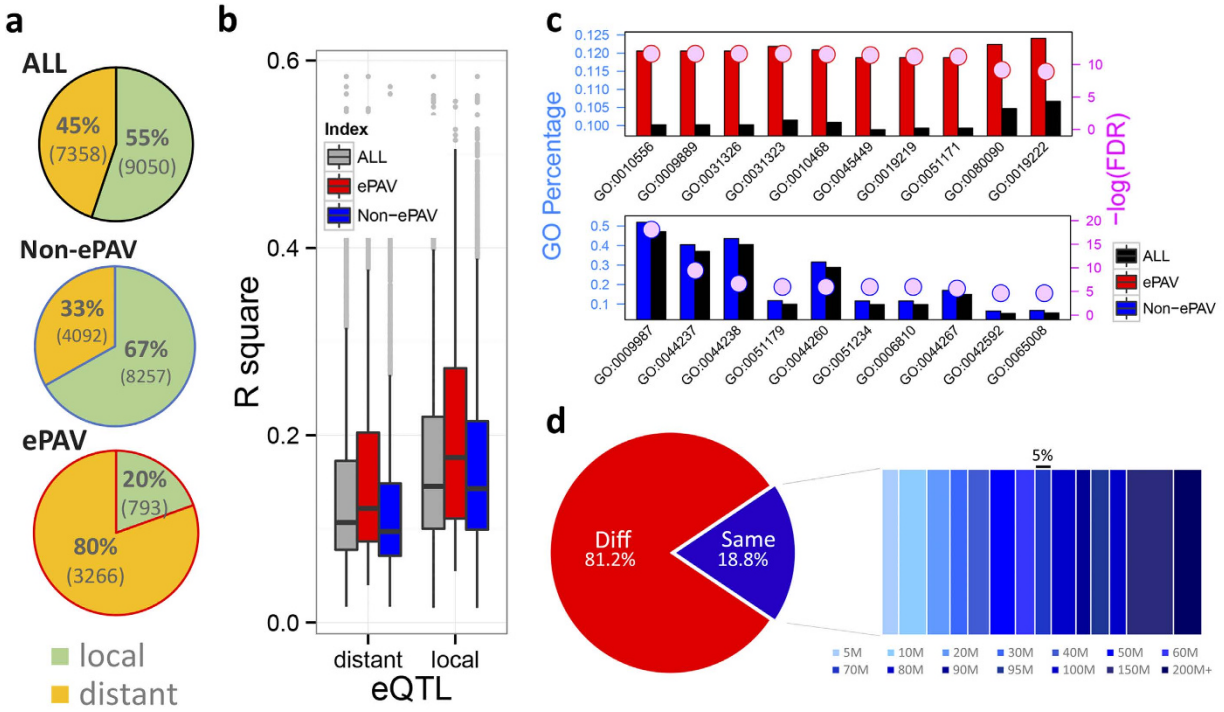

Figure 1. ePAV candidates played key roles in distant-regulation. (**a**) The ratio of local- (green) and distant- (orange) eQTLs among ePAV, non-ePAV and ePAV + non-ePAV together, expressed as percentages. (**b**) The effects of local eQTL were larger than those of distant eQTL both for ePAV (P = 7.05E-22; student test) and non-ePAV genes (P = 1.92E-135). The eQTL effects for ePAV genes were greater than those for non-ePAV genes in both local (P = 1.34E-18) and distant (P = 7.18E-56) types. (**c**) Top 10 GO enrichment terms in biological processes of ePAV (red) and non-ePAV (blue) are displayed. The left y-axis represents the percentage of genes belonging to each GO term. The colored circles and right y-axis represent the significance level (FDR). Red, blue and black colors means ePAV, non-ePAV and reference levels, respectively. The corresponding GO term description for each GO number could be available in Supplementary Table S1 online. (**d**) ePAV candidates as distant-eQTL affecting expression of Non-ePAV genes. "Diff" means the eQTL is located on a different chromosome with its regulating gene and "Same" represents both are located on the same chromosome (expressed as %). And even for the "Same" cases, the eQTLs tend to be located far away with their regulated targets (the colored rectangles represent the different distance windows, and the width represents the corresponding ratio.

Based on the 'align-then-assemble' strategy, 7,775 contigs with a total length of 3.46 Mb were obtained, of which N50 size was 445 bp, much shorter than the average length of reference transcripts (1826 bp). Most of these contigs had no hits to protein databases, so it seems that they do not correctly represent the transcripts. We suspect that there was a large proportion of incomplete fragments due to filtering conserved reads and breaking long contigs into short ones. Based on the 'assemble-then-align' strategy, 2,355 novel sequences with a total length of 1.9 Mb (N50 = 922 bp) were obtained (see Supplementary Fig. S4 online, Additional Information and methods), resulting in longer and more complete contigs compared to the results of 'align-then-assemble' (Supplementary Fig. S5). Further, comparison of the results of the two strategies indicated that some of the sequence reads of conserved functional domains might be filtered out when applying 'align-then-assemble' strategy. For example, Unigene_441 from the 'assemble-then-align' strategy was identical to Unigene_ref71 from the 'align-then-assemble' strategy but longer and containing the unknown protein domain of DUF789. The distribution of unique reads and further PCR re-sequencing both confirmed that the result from 'assemble-then-align' was correct (see Supplementary Fig. S6, Supplementary Table S3 and Supplementary Fig. S9 online). Thus, only the results from 'assemble-then-align' strategy were used for further analyses.

To evaluate the reliability of the assembled novel sequences, we first compared 2,355 novel sequences to the 4,712 novel genomic contigs obtained in a study of deep sequencing six elite maize inbred lines[8], showing that 447 (19%) of our novel sequences align to those novel contigs (see Additional Information). Second, the novel sequences were compared with 8,681 novel representative transcripts from whole seedling RNA-seq on a panel of 503 diverse maize inbred lines[12]. Nearly 60% (1,380 among 2,355) of the novel sequences identified in the present study had above 85% identity in the alignment with novel transcripts detected from seedling tissue (see Additional Information). In total, about 62% of our novel sequences were found to have hits in at least one of the previous studies.

We validated the present/absent variation of 10 randomly selected novel sequences in a set of 96 inbred lines including B73 and Mo17 using genomic PCR (see Supplementary Fig. S7 online and Additional Information). Two of these novel sequences were present in all 96 inbred lines (Unigene_31, Unigene_361), possibly due to the presence in the genome but absence at the expression level. The other eight were determined to be PAVs, and the consistency of present/absent status between the transcriptome assembly and PAV detected by genomic PCR ranged from 31% to 99%, with an average of 72% (Supplementary Fig. S7). Most (89.2%) of the inconsistency was also due to the presence at the genomic level without expression in kernel (Supplementary Fig. S7). We further re-sequenced the amplified products from genomic DNA of the 10 randomly selected novel genes in 5 diverse genotypes and all were consistent with assembly sequences (Supplementary Fig. S8 and Supplementary Fig. S9 online). Cross-comparison with other studies and experimental results not only validates the assembled novel sequences, but also indicates that the predicted present/absent variants are reliable.

Annotation and mapping of novel expressed sequences. To annotate novel sequences identified in this study, we first compared the sequences with the non-redundant (nr) protein database[16] using NCBI BLAST, which showed that 1,359 of them had significant matches (E-value < 1e-6) and most (93.57%) of the best matches were within Poaceae. The majority of the significant hits (1,318 of 1,359, 97%) were functionally classified into six types of known enzymes (Supplementary Fig. S10) and conserved domains or annotated motifs (Supplementary Fig. S11; Additional Information). In the GO enrichment analysis, the overrepresented processes included several metabolic processes and biotic stimuli (Supplementary Fig. S12 and Supplementary Table S5). In addition, 145 of the 1,037 unannotated novel sequences were considered to have coding potential, having at least 120 amino acid-long predicted open reading frames (ORFs) and a homolog in the non-redundant protein database at a lax standard (E-value < 1e-3). Furthermore, 248 of remaining novel sequences were annotated as smRNA precursors against small RNA database[17] (E-value < 1e-10), and the remaining 644 were predicted to be high confidence novel lncRNAs in maize (see Supplementary Fig. S11 and Additional Information).

To locate possible physical chromosomal positions of the novel sequences, the linkage disequilibrium (LD) mapping strategy was used between novel SNPs within new sequences and high density SNPs in the whole inbred line collection (Supplementary Fig. S13). After multiple sequences alignment, 27,466 SNPs from 664 novel sequences were provisionally identified (See Methods). Based on the LD between SNPs located in novel sequences and high density SNPs with known positions in the whole panel, 625 novel sequences (94.4%) were mapped onto the reference genome (see Supplementary Fig. S14 and Supplementary Table S6 online). The locations of the common expressed genes and the SNPs show the similar trends with enrichment at the ends of the chromosomes, while the distribution pattern of the novel sequences demonstrates fluctuation, and in some cases concentrates near the centromeres on some chromosomes, thus physically complementing the reference genome-containing variations (Supplementary Fig. S14).

Maize pan-transcriptome plays an important role in regulating phenotypic variation. To systematically explore the genetic consequences of the above described expression variation, and considering that the metabolic phenotype provides a link between gene sequence and visible phenotype, genome wide association study (GWAS) was performed to study the potential effects of ePAV genes on 616 metabolites detected in mature kernels[18] and 17 agronomic traits[19] measured in the same panel. Among the ePAV genes, 56 (0.42%) were significantly (P < 7.49E-5, 1/n) associated with 15 agronomic traits and 1,967 (14.74%) associated with 526 metabolic traits including content of 18 amino acids[4] (see Additional Information).

A major secondary metabolite group in plants, the flavonoids, is widely distributed and has variety of functions[20]. The pericarp color1 (p1) gene encoding an R2R3 Myb-like transcription factor[21] regulates flavonoid biosynthesis by promoting a suite of structural genes, and conditions pigment in several floral organs including the seed coat, cob glumes, tassel glumes, and silk under both genetic and epigenetic regulation mechanisms[20–22]. In this study, the quantification of 39 flavonoid metabolites together with cob color were used for GWAS, and results indicated the ePAV pattern of the p1 gene was highly associated with cob color (P = 1.33E-19), and was correlated with six different flavonoid metabolites (P < 2.34E-05; Fig. 2a). We also identified a structural gene (GRMZM2G162755; anthocyanidin 3-O-glucosyltransferase) significantly associated with cob color (P = 7.05E-20) and the same six flavonoid metabolites (P < 6.23E-05; Fig. 2a). This gene was shown to be regulated by p1 in a previous eQTL mapping study[5] and ChIP-Seq analysis[23]. Another copy of the R2R3 Myb-like transcription factor (p2, GRMZM2G057027) could regulate the other two flavonoid metabolites (P < 3.41E-06; Fig. 2a). Notably, we found these three ePAV candidates, as regulators, could also control expression of other genes that are related to flavonoids such as c2, chi1, a1, pr1 and whp1[21,23] (Fig. 2b; P ≤ 9.65E-10), providing further support for the hypothesis that these ePAV genes are involved in the flavonoid pathway and functioning through the PAV differences in expression level.

The novel expressed sequences play critical roles in regulation of the transcriptome and metabolome. For the novel sequences, a re-mapping strategy was applied to correct the PAV distribution for each novel gene, to be used in GWAS (see Methods). We found that 26 (1.1%) of the novel genes were associated (P < 4.25E-4) with 13 agronomic traits (see Additional Information). Eleven were associated with flowering time (i.e. Days to Tasseling, Days to Pollen Shed and Days to Silking). We also identified a novel gene (Unigene_55) that encodes a late embryogenesis abundant (LEA) protein that is associated with kernel width (P = 2.14E-5). LEA proteins have been described as accumulating late in embryogenesis and could protect other proteins from aggregation under various environmental stresses[24]. Here we provided a clue that LEA may also affect kernel size.

Moreover, 788 novel genes (33.46%) were associated (P < 4.25E-4, 1/n) with 487 metabolic traits measured in maize kernels[18] (see Additional Information), which implied that those novel genes could play more complex roles within cellular metabolism processes; thus, this study provides fresh resources for the genetic study of maize kernel quality and production. Metabolic processes are commonly controlled by transcription regulation. Therefore, it is valuable to examine whether the identified novel genes were widely involved at multiple regulatory levels. We

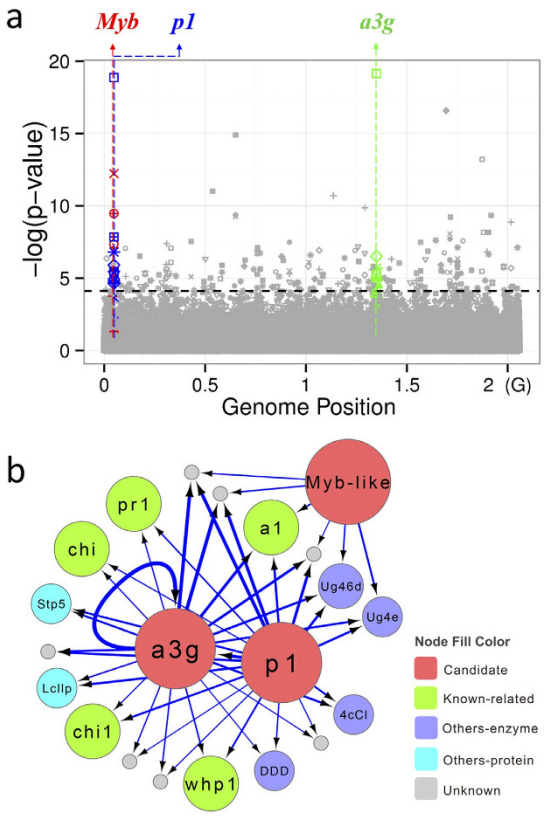

Figure 2. ePAV candidates contributed to both maize cob color and various kinds of flavonoids. (**a**) Manhattan plot of the association of three ePAV candidates, maize cob color and several flavonoids. Different shapes represent different traits, and points with different color represent different kinds of ePAV candidates: Blue: *pericarp color1* (*p1*, GRMZM2G084799); Red: *p2*, another copy of R2R3 Myb-like transcription factor (GRMZM2G057027); Green: anthocyanidin 3-O-glucosyltransferase (GRMZM2G162755); Grey: other ePAVs. Black dashed horizontal line was the cut-off (P = 7.47E-5) of significant level. (**b**) The three ePAV candidates were also significantly associated with expression of related genes within maize flavonoid pathway. Nodes in red are the three ePAV candidates above, green nodes represent several identified genes located in the maize flavonoid pathway, purple nodes are other genes encode enzymes, light blue were other genes encoding non-enzyme proteins (such as transporters), and grey nodes had no annotation. The blue arrow edges link the ePAV candidates and its associated targets and the a3g links to itself meaning self-regulation in expression level, while thicker lines represented more significant associations.

found the novel sequences were significantly responsible for expression levels of annotated genes or their expression presence/absence states at a strict cutoff ($P \leq 1E\text{-}4$), including the novels annotated as non-coding RNAs (see Additional Information). By combing the metabolome and transcriptome findings, 23 novel genes were found playing roles in both metabolic processes and expression regulation.

Plant NBS-LRR proteins can directly or indirectly recognize pathogen-deployed proteins and triggers plant defense responses[25,26], and exhibit high levels of PAV polymorphism in various plants[27–29]. Here, we identified two novel NBS-LRR genes (Fig. 3a,b) that have high homology to rice NBS-LRR genes Os11gRGA4 and Os11gRGA5, which were shown to interact functionally and physically to mediate resistance to the fungal pathogen Magnaporthe oryzae[30,31]. Recent studies have verified that inducing plant immunity impacts flavonoid biosynthesis[32,33], and that flavonoid compounds significantly contribute to plant resistance[34]. Os11gRGA4 was found to be associated with the flavonoid Naringenin O-malonylhexoside[35]. Interestingly, we found that these two novel NBS-LRR like sequences were both associated with Apigenin C-pentosyl-O-coumaroylhexoside and C-pentosyl-apigenin O-caffeoylhexoside contents, two flavonoid metabolites (Fig. 3c). In addition, these two novel sequences were also associated with several gene expression presence/absence states (Fig. 3d), including transcription factors with DNA binding activity (TGA6 or GRMZM2G000842; GRMZM2G405170), spliceosomal complex (GRMZM2G011034), nucleic acid binding genes (GRMZM2G088348), translation release factor (GRMZM5G864412), actin cytoskeleton (GRMZM2G552644) and other enzymes functioning in metabolic processes. These observed associated targets were consistent with previous observations that alternative splicing is important in the regulation of NBS-LRR proteins and plant immunity[36], and that TGA6 and other bZIP transcription factors are significant in plant defense against pathogens[37,38]. Actin cytoskeleton dynamics also play an important role in mediating resistance[39], and translation release factors are critically involved in the elimination of aberrant mRNA. The regulatory targets in pathogen defense response are indeed R-genes, especially for the abundant and alternatively spliced NBS-LRR R-genes[40].

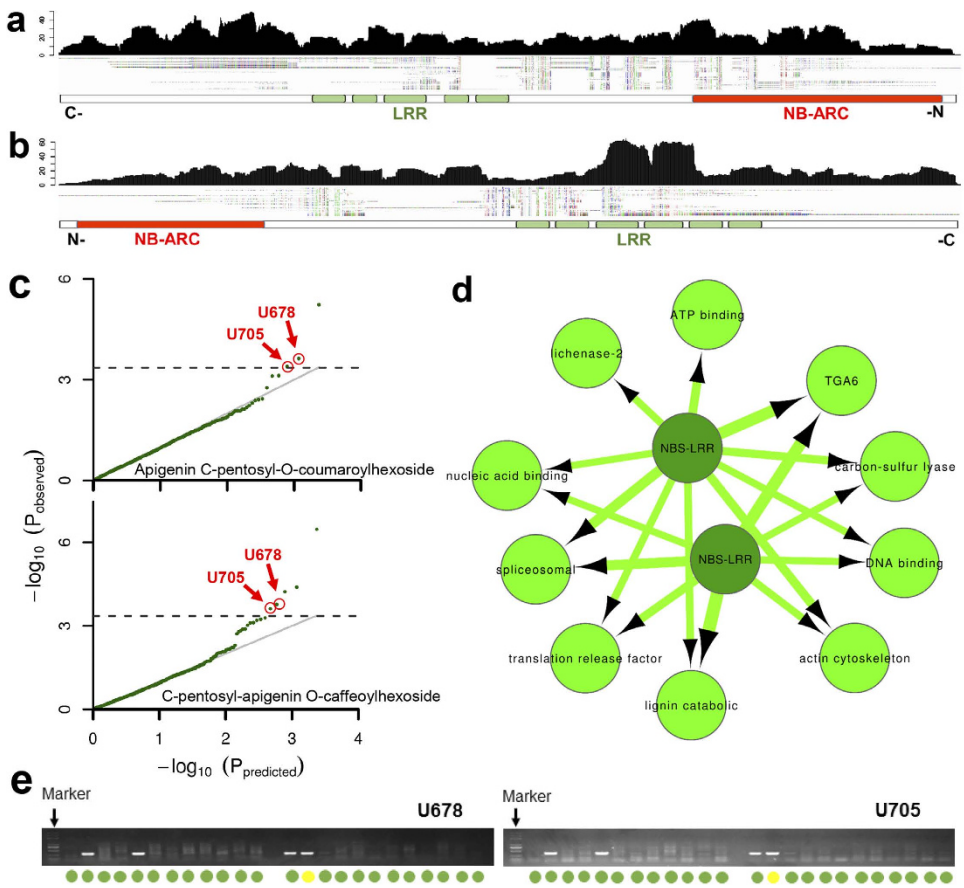

Figure 3. Two novel NBS-LRR genes showed significant association with flavonoid metabolites and with expressed genes involved in flavonoid pathway. Read distribution and predicted conserved domains of novel reference gene Unigene_678 (**a**) and Unigene_705 (**b**) and sequence alignments for all presence genotypes. (**c**) Q-Q plot of association mapping for different flavonoids. (**d**) The two novel NBS-LRR genes were also significantly associated with other genes with expression presence-absence variation. (**e**) Validation of the PAV of the two novel genes. Green represents consistency between experiment and prediction. Yellow means the gene was absent in our prediction but exists in the genome.

The two novel sequences were confirmed by PCR sequencing (Supplementary Fig. S9 and Supplementary Table S3). Moreover, the consistency of presence/absence variation in the association panel used in our study between observed and predicted variants was greater than 98% (Unigene_678) and 96% (Unigene_705), respectively. The PAV states of these two genes on the genome level also showed a significant relationship with the metabolic traits mentioned above. These results show that the dispensable novel expressed sequences were important both in morphological adaptation processes as previously reported[12], and in cellular metabolome and transcriptome regulation.

Present and absent genes may contribute to trait heterosis. Complementation of gene content variation is assumed to be important in heterosis[6,8,41]. Since our identified expressed novel genes have been shown to be functionally important, their combination of inbred-specific sequences in hybrids could provide novel trans-interactions potentially resulting in non-additive expression. This provides an opportunity to test the link between gene content PAV and heterosis. We crossed the association panel with the Mo17 inbred line to develop a suitable population to test this. Six yield-related traits were measured for each hybrid in different environments over two years (see Methods). The degree of heterosis increased with more complementary (present in one and absent in the other inbred parent) novel genes in the hybrids among five of the six measured traits (Fig. 4A). This trend is more significant for those traits with relatively stronger heterotic effect, and novel sequences identified in this study have a greater effect than ePAVs (Fig. 4B). However, only a small portion ($< 10\%$) of observed heterosis was explained by novel sequences and/or ePAVs, which implies that heterosis is complexly affected by many different factors[6,8,9,42].

Discussion

Expression PAV is a kind of variation at transcript level, mostly due to genetic or epigenetic regulation. With the conservative distribution-based approach (see Methods), we identified more than 13,000 genes as ePAVs, about one third of the maize annotated genes. These are genes that are only expressed in a subset of the association panel. This finding was based on one tissue (kernel of 15DAP) and limited inbred lines ($n = 368$), therefore, more ePAVs

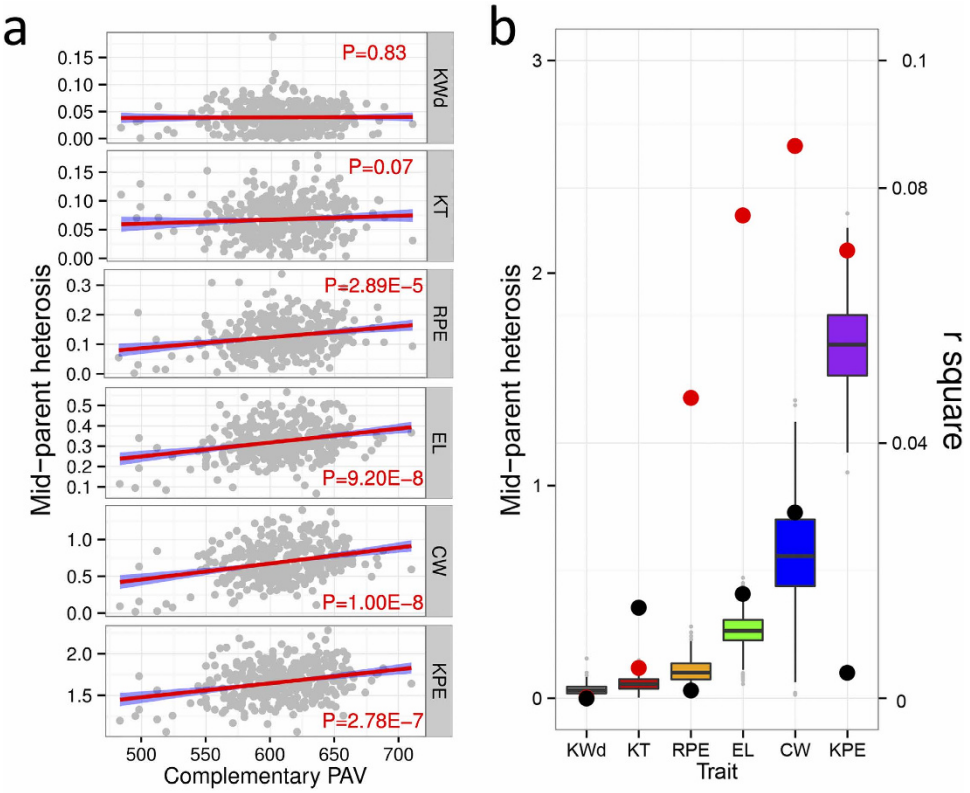

Figure 4. PAV status of novel genes and ePAV both correlated with heterosis of most yield-related traits. (**a**) Correlation between mid-parent heterosis and the number of complementary novel genes exhibiting PAV between the parents of the F1 population. Six panels represent different yield-related traits. KWd: Kernel Width; KT: Kernel Thick; RPE: Rows Per Ear; EL: Ear length; CW: Cob Weight; KPE Kernels Per each row of Ear. (**b**) Boxplot in different colors represents different traits ordered by mid-parent heterosis (the left y axis). The points in red and black represent Pearson's r^2 of correlation between mid-parent heterosis and the number of complementary novel genes and ePAVs showing presence-absence variation between parents of the F_1 population.

will be identified when more tissues and materials are studied. The number of ePAV should vary under different sequencing coverage and even different cutoffs; however, this study demonstrates that ePAV is a common phenomenon and that the underlying mechanisms and ramifications need to be explored. Among our findings, the ePAV genes were enriched in regulation-related processes and were usually regulated by distant eQTLs while core expression genes were commonly regulated by local eQTLs (Fig. 1a). The different regulatory patterns imply that the two kinds of genes may affect phenotypic variation by different mechanisms.

Transcript variation as an independent variable can be regarded as a molecular marker to perform GWAS (i.e. ePAV-GWAS), corrected for population structure and relatedness, and this should provide additional insights into the architecture and regulation of quantitative traits and help understand certain important biological questions such as adaptation[12,43]. Interestingly, about 15% of the identified ePAVs were found to associate with agronomic and metabolic traits, which confirms the expectation that gene expression presence and absence can affect the phenotypic variation directly. Combining the ePAV-GWAS results with SNP-GWAS and eQTL mapping information aid not only in the identification of gene candidates, but also to better understand molecular mechanisms. Here, we use cob color and several related flavonoid metabolites as an example for further exploration. After strict filtering of our genotypic data, there were no SNPs left within the p1 locus (see Supplementary Fig. S15b), known to control cob color, but an associated region was found upstream of the the p1 locus (~200 K, Supplementary Fig. S15a,b). This makes it difficult to unambiguously identify a single causal gene for cob color. Using ePAVs as markers (ePAV-GWAS), the p1 locus was exactly identified and the p2 was promoted as another candidate (Fig. 2a, Supplementary Fig. S15b, panel4). Another significant locus (GRMZM2G162775, a3g) on chromosome 6 was detected by ePAV-GWAS (Supplementary Fig. S15a, panel4), and may have been detected as regulated by p1 by the previous eQTL mapping[5] (Supplementary Fig. S15a, panel 3) and ChIP-Seq studies[23]. This trans-regulation pattern could not be discovered by applying SNP-GWAS, even when the SNP density was doubled to 1.25 million (data not shown). Thus, the ePAV will provide a unique complementarity to SNP marker for the genetic exploitation.

Although more than 13,000 ePAVs were identified, only small proportion (~1%) included genomic PAVs which are sometimes called dispensable genes. Previous studies revealed that the B73 reference genome included only 70% of the total low-copy sequences available in the maize species[44], which implied that many dispensable genes are present beyond the reference genome. It is necessary to explore these novel genes that may be phenotypically important in certain genotypes. We applied *de novo* assembly to detect such novel sequences. Of the two combined

strategies were available, 'assemble-then-align' resulted in longer and more complete contigs. Although in theory, 'align-then-assemble' should be more sensitive and *de novo* assembly was likely to work only for the most abundant transcripts[45], in practice, the align-first strategy probably enrich some low abundance (and possibly extraneous) reads and assemble them into contigs. Only a small portion (4%, identity $\geq$ 85%, coverage $\geq$ 85%) of the contigs from the align-first strategy could identify high confidence matches in assemble-first contigs. After stringent filtering, 2,355 high confidence novel sequences with a total length of 1.9 Mb were obtained.

The enrichment analysis of these novel sequences suggested their roles in metabolic processes responsive to stimuli. Almost 34% of them were found to be associated with metabolic traits. The differences in metabolism-related genes may be associated with differential environmental effects[4]. The novel sequences involved in development, such as beta-tubulin, also likely contribute to adaptation. In hybrids, the number of novel genes in heterozygous (present in one parent, absent in the other) state was correlated with heterosis of yield-related traits, which supports the complementation hypothesis of this phenomenon. We showed that the "dispensable" genes, whether they were present on reference genome or not, indeed play an indispensable roles at the population level.

The construction of the maize pan-transcriptome is more effective than a maize pan-genome because the high proportion of repetitive sequences present in the genome complicates assembly. However, limitations in tissue and availability of diverse genotypes could result in underestimating the size of maize pan-transcriptome (see Supplementary Fig. S16a). We compared tissue-specific and genotype-specific efficiency in discovering novel transcripts. We selected five diverse tissues (16DAP Whole Seed, V3_Stem and SAM, V9_Immature Leaves, R2_Thirteenth Leaf, 6DAS_GH_Primary Root) from a previous study[46] and repeated the de novo assembly process as described. We found more novel transcripts when adding new tissues than by adding new individuals (Supplementary Fig. S16b). This indicated that the expression divergence is significantly larger between tissues than individuals ($P = 1.07E\text{-}58$).

We estimated the size of maize pan-transcriptome based on our RNA-seq data from one tissue but multiple genotypes (see Supplementary Fig. S16c and Methods). As expected, when adding more genotypes to the analysis the number of additional novel sequences detected eventually leveled off and the total is expected to reach an asymptotic maximum of ca. 28,000. Using the reference genome and a similar procedure, we found that number of core expression genes decreased and became nearly invariable when more than 200 lines were included. The minimum number of core expressed genes and maximum for dispensable ones were 22,043 and 13,382, respectively (Supplementary Fig. S16d). Combining the reference based genes and newly identified genes from the current study, we estimated the size of the pan-transcriptome of the maize whole kernel is about 63,000. Under the simple assumption that maize kernels only express 70% ~ 80% of the total genes[5], the whole pan-genome of maize is close to 78,000 ~ 84,000. Thus, the present reference genome may only capture half of the predicated maize pan-gene, which is similar to previous prediction[12]. To identify the pan-gene and study the functions will help to understand the genome better thus enhancing crop improvement.

Materials and Methods

Detection of ePAV. In the previous study[5], authors quantify the expression of 38,032 reference genes, read counts for each expressed gene and individual transcripts of that gene were calculated and scaled according to the definition of RPKM (reads per kilobase of exon model per million mapped reads). The genes showing expression (RPKM $>$ 0) in less than 19 (5%) inbred lines were excluded in the following analysis. In this study, we further filtered the genes which expressed (RPKM $>$ 0) in more than 348 (95%) inbred lines. The remaining genes were considered to have presence/absence variation in expression, and several further distribution-based steps were used to acquire an ePAV pattern: (1) extract non-zero expression data of a ePAV gene in 368 inbred lines; (2) sort it from smallest to largest and make the frequency distribution (10 groups); (3) turn the abnormal low (data in 1st group) and high (data in groups that frequency $<$ 3) expression values according to the frequency distribution as "NA" (4) convert the rest of no-zero expression data to '1' and no expression data to '0' to get the ePAV pattern of each gene.

Prediction of PAV through genotyping from 50 K and 600 K SNP arrays. The ePAV genes (including 1Kb upstream and downstream regions) containing at least two SNPs in the array genotyping dataset were used to analyze their PAVs. A gene was regarded as potential PAV if all its SNPs were genotyped as missing in a particular line. Further, for each ePAV, if the potential PAV occurred at more than a certain ratio (5%, or 19 for 50 K and 2 for 600 K datasets, respectively), it was considered as non-random, thus to be candidate PAV.

***De novo* transcript assembly.** The poly(A) + transcriptomes of immature kernels (15 DAP) were sequenced using 90-bp paired-end Illumina sequencing with libraries of 200-bp insert sizes. The sequencing data for this project can be downloaded in the NCBI Sequence Read Archive under accession code SRP026161. Average 73.9 million reads were obtained in each sample and 367 inbred lines were used in assembly process[5] (Supplementary Table S4). The adaptors and low quality reads were filtered using Trimmomatic software[47], resulting in a total 24.7 billion high-quality reads, used for the assembly (Supplementary Table S4 and Supplementary Fig. S17).

While applying the "align-then-assemble" strategy, the mapping process was first performed by Bowtie2[48] (version 2.0.2) and TopHat2[49] (version 2.0.6) with the parameters −i 5, −I 60000, −r 20, —mate-std-dev 75 and gene annotation was provided. The unmapped reads from each individual were assembled by Trinity[15], which is based on the de Bruijn graphs algorithm. Min count for K-mers to be assembled by Inchworm most influenced the result. We found that there was a large increase in the numbers of transcripts that align to the B73 reference transcripts when applying min K-mers between 2 and 3 and we chose the parameters: -seqType fq,-min_kmer_cov 2, -min_contig_length 200. In the "assemble-then-align" strategy, the whole cleaned RNA-seq reads from 367 inbred lines were de novo assembled with the same parameters.

Identification of novel sequences. To detect truly novel sequences and remove the ones that were alleles or paralogs of sequences present in the B73 reference genome, the assembled transcripts from each line were aligned to B73 5b pseudomolecules using GMAP[50], a genome alignment program for mRNA sequences. We randomly chose 200 assembled transcripts from each inbred line to determine GMAP parameters and the identity cutoffs were then set to 0.85. The representative transcripts that did not align to the reference sequence were clustered by the TGI Clustering tool (TGICL)[51]. Transcripts present in at least 19 inbred lines (5% of 367) with non-homology (identity $<$ 95%; coverage $<$ 90%) to B73 cDNA 5b pseudomolecules (FGS) were retained as candidate novel sequences. DeconSeq[52] was further used to remove sequence contamination to improve the reliability of novel sequence identification. Reference genomes of human, human microbiome, and virus were used as the "contamination-datasets", and plant datasets including Zea mays, Oryza sativa, Sorghum bicolor, Setaria italica, and Brachypodium distachyon were used as "retain-datasets" under the parameters of identity $\geq$ 98% and coverage $\geq$ 90%. Finally, RNA-seq reads were aligned back to novel sequences for quality assessment by running alignReads.pl in Trinity software[15] with the —bowtie and —phred64-quals options. The 12 sequences which had breakpoints in distribution of reads were excluded. This indicates there may be minor errors in the transcripts assembly process. All the procedures and related results were shown in Supplementary Fig. S17 online. On average, 57,628 assembled transcripts with N50 size 1,078 bp were obtained for each inbred line (Supplementary Table S4). After excluding the transcripts present in the B73 reference and other contaminations, an average of 1,388 unmapped transcripts were retained in each inbred line. We clustered these remaining transcripts from all inbred lines, and the longest one was selected as a representative sequence in each unigene cluster. Each unigene cluster would then be retained if it was present in at least 19 inbred lines (5% of the panel). Finally, 2,355 novel representative sequences with a total length of 1.9 Mb and N50 size 922 bp were obtained (Supplementary Fig. S4 and Supplementary Table S4).

An improved re-mapping strategy to correct the distribution of novel sequences. After the clustering step, we obtained the PAV patterns for each novel gene among all genotypes. When considering those present in genomic sequence but non-expressed as "inconsistent", the consistent ratio reached to average 67%, using a simple clustering step. However, we found that some novel genes appear to be also expressed in predicted "Absence" lines

This may be caused by incorrect assignment to the genomic location of short or well-conserved expressed sequences. In these cases we applied a second re-mapping step to recover correct genomic matches, by using BLASTx with "identity $\geq$ 0.96, query-coverage $\geq$ 0.5, subject-coverage $\geq$ 0.96", which improved the consistency ratio to an average of 72%.

Annotation of novel sequences. Blast2go[53] is an all in one tool for functional annotation of novel sequences and the analysis of annotation data. For BLASTx to nr database[16], a minimum E-value of 1e-6 was used and only best hits were considered. BLAST XML result file was imported in Blast2go. GO mapping and InterProScan[54] were also performed to complete the annotation. A total of 1,359 novel sequences had matches in the nr protein database using BLASTx (E-value $\leq$ 1e-6). Nearly all of them (1,318 of 1,359, 97%) can be functionally classified into families and contained conserved domains and functional sites. The remaining 1,037 unannotated sequences were left. Among annotated ones, 166 could encode enzymes (see Supplementary Fig. S10 and Additional Information). 640 can be grouped into at least one GO term and used in the next GO enrichment analysis.

The remaining unannotated novel sequences were used to predict the protein coding potential. Three important criteria were used: transcript length, open reading frame (ORF) size, and presence of homology with known proteins. Transcript length was set to 200 bp. Only three ORFs longer than 120 amino acid were identified in 14 known long non-coding RNAs (lncRNAs)[55], ORFs longer than 120 amino acid considered potential coding candidates. Coding regions and the corresponding amino acid sequences were extracted from novel sequences using TransDecoder in the Trinity software[15]. In addition, transcripts were aligned to UniProtKB/Swiss-Prot database[56] identify transcripts with potential protein-coding ability (E-value $\leq$ 1e-3). The unaligned transcripts were considered non-coding RNAs (ncRNAs). Using BLASTN (E-value $\leq$ 1e-10) and Infernal software[57] ("INFERence of RNA ALignment"; score $\geq$ 40), 248 sequences were matched to NONCODE database[58], smRNA transcriptome databases including predicted microRNAs (miRNAs), other predicted short hairpin forming RNAs (shRNAs) and predicted small interfering RNAs (siRNAs) or Rfam database[17,58]. These sequences were all considered the precursors of small RNAs. The remained 644 novel sequences were predicted to be high confidence maize lncRNAs (see Supplementary Fig. S11 and Additional Information).

SNP analysis and LD mapping of novel sequences. MAFFT[59] was used to align novel sequences from all inbreds to their corresponding representative ones (the longest one in each unigene cluster; see above). SNP_SITES software (https://github.com/sanger-pathogens/snp_sites) was used to identify SNPs in the multiple alignment. Biallelic SNPs with minor allele frequencies (MAFs) larger than 0.05 were retained for analysis. A total of 27,466 SNPs were identified in 664 novel sequences. Pairwise LD between SNPs within novel sequences and between SNPs in B73 reference genome was computed by a script on the assumption of equal probability for either phase relationship of the alleles. B73 reference gene with the highest LD (and at least $r2 > 0.1$) to SNPs within the novel gene was considered the likely location of the novel gene. Using this approach, of the 664 novel sequences with SNPs, 627 were mapped to the B73 reference.

Validation of PAVs within ePAV and novel ones. Genomic PAV for 10 ePAV genes, and the 10 novel expressed genes across a set of 96 diverse inbred lines were evaluated using touchdown PCR. Inbred lines information, primer sequences and experiment results are available in Supplementary Fig. S3, Supplementary Fig. S7

and Supplementary Table S3 online. The thermo cycler program for touchdown PCR were included: 1 = 94 °C 5 min; 2 = 94 °C 30 s; 3 = 64 °C 30 s −0.5 °C/cycle; 4 = 72 °C 50 s; 5 = GOTO 2 12repeats; 6 = 94 °C 30 s;7 = 58 °C 30 s; 8 = 72 °C 50 s; 9 = GOTO 6 23repeats; 10 = 72 °C 5 min; 11 = 25 °C 2 min; 12 = END. The PCR products of 10 novel genes in 5 lines were then re-sequenced and subjected to multiple alignment to evaluate the correctness of de novo assembly.

Novel sequences and ePAVs both contributed to heterosis. To test whether the PAV pattern of the novel genes contributes maize heterosis, all population inbred lines were planted with randomized complete experimental design by single replication in 2011 (Chongqing city; Hebi city, Henan province; Honghe autonomous prefecture, Yunnan province and Sanya city, Hainan province) and 2012 (Chongqing city; Hebi city, Henan province; Honghe autonomous prefecture, Yunnan province and Wuhan city, Hubei province) and 6 yield-related traits including Kernel Width (KWd), Kernel Thickness (KT), Rows Per Ear (RPE), Ear Length (EL), Cob Weight (CW), Kernels Per Ear (KPE) were measured. Generally, the average values from five individuals were calculated to represent each line in each experiment and the BLUP values from different environments and years were used for next analysis. The number of PAVs in heterozygous state (present in one parent, absent in the other) were then used to evaluate their correlation (R-square was measured; Fig. 5) with observed mid-parent heterosis for each trait.

The estimation of the maize pan-transcriptome size. Five diverse tissues (16DAP Whole Seed, V3_Stem and SAM, V9_Immature Leaves, R2_Thirteenth Leaf, 6DAS_GH_Primary Root) from a previous study[46] were chosen to repeat de novo assembly process. We then compared each pair of tissues and individuals, by measuring the ratio of shared genes to total genes. To eliminate the effect of biased sample size (5 tissues vs 367 individuals), we randomly selected five pairwise comparisons and repeated this process 1000 times, then compared resulting distribution.

To determine whether the maize pan-transcriptome is open (the size of pan genome grows continuously with the number of sequenced individuals increases) or closed (the size of pan genome reached a constant value with the number of sequenced individuals increases) and to estimate the size of it, a simulation process on real data was used. There were three parts to form the maize pan-transcriptome: reference based core genes, reference based dispensable genes (ePAV) and novel sequences. We randomly chose 20 samples in 367 maize inbred lines in a clustering run to estimate the number of novel sequences among them and then add another 20 lines to do the same cluster recursively until a total of 360 inbred lines were in the set. Ten independent simulations were run and the mean of each run from n = 20 to 360 inbred lines was used to estimate the maximum number of novel sequences. The same simulation process was also performed to estimate the maximum value of core genes and dispensable genes on references genome.

References

1. Yan, J., Warburton, M. & Crouch, J. Association mapping for enhancing maize (L.) genetic improvement. *Crop Sci.* **51,** 433–449 (2011).
2. Huang, X. & Han, B. Natural Variations and Genome-Wide Association Studies in Crop Plants. *Annu Rev Plant Bio.* **65,** 531–551 (2014).
3. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet.* **16,** 197–212 (2015).
4. Liu, H. *et al.* Genomic, transcriptomic and phenomic variation reveals the complex adaptation of modern maize. *Mol Plant*. **8,** 871–84 (2015).
5. Fu, J. *et al.* RNA sequencing reveals the complex regulatory network in the maize kernel. *Nat Commun.* **4,** 2832 (2013).
6. Springer, N. M. *et al.* Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5,** e1000734 (2009).
7. Swanson-Wagner, R. A. *et al.* Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* **20,** 1689–1699 (2010).
8. Lai, J. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet.* **42,** 1027–1030 (2010).
9. Hansey, C. N. *et al.* Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. *PLoS One.* **7,** e33071 (2012).
10. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat Rev Genet.* **12,** 671–682 (2011).
11. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science.* **326,** 1112–1115 (2009).
12. Hirsch, C. N. *et al.* Insights into the maize pan-genome and pan-transcriptome. *Plant Cell.* **26,** 121–135 (2014).
13. Li, Q. *et al.* Genome-Wide Association Studies Identified Three Independent Polymorphisms Associated with α-Tocopherol Content in Maize Kernels. *PLoS One.* **7,** e36807 (2012).
14. Unterseer, S. *et al.* A powerful tool for genome analysis in maize: development and evaluation of the high density 600k SNP genotyping array. *BMC Genomics.* **15,** 823 (2014).
15. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* **29,** 644–652 (2011).
16. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40,** D130–D135 (2012).
17. Wang, X. *et al.* Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *Plant Cell.* **21,** 1053–1069 (2009).
18. Wen, W. *et al.* Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat Commun.* **5,** 3438 (2014).
19. Yang, N. *et al.* 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,** 821–833 (2014).
20. Koes, R., Verweij, W. & Quattrocchio, F. Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. *Trends Plant Sci.* **10,** 236–242 (2005).
21. Grotewold, E., Drummond, B. J., Bowen, B. & Peterson, T. The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. *Cell.* **76,** 543–553 (1994).
22. Sekhon, R. S., Peterson, T. & Chopra, S. Epigenetic modifications of distinct sequences of the p1 regulatory gene specify tissue-specific expression patterns in maize. *Genetics.* **175,** 1059–1070 (2007).

23. Morohashi, K. *et al.* A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. *Plant Cell.* **24,** 2745–2764 (2012).
24. Goyal, K., Walton, L. & Tunnacliffe, A. LEA proteins prevent protein aggregation due to water stress. *Biochem J.* **388,** 151–157 (2005).
25. DeYoung, B. J. & Innes, R. W. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat Immunol.* **7,** 1243–1249 (2006).
26. McHale, L., Tan, X., Koehl, P. & Michelmore, R. W. Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* **7,** 212 (2006).
27. Shen, J., Araki, H., Chen, L., Chen, J. Q. & Tian, D. Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in Arabidopsis thaliana. *Genetics.* **172,** 1243–1250 (2006).
28. Yang, S. *et al.* Genetic variation of NBS-LRR class resistance genes in rice lines. *Theor Appl Genet.* **116,** 165–177 (2008).
29. Wu, P. *et al.* Loss/retention and evolution of NBS-encoding genes upon whole genome triplication of Brassica rapa. *Gene.* **540,** 54–61 (2014).
30. Okuyama, Y. *et al.* A multifaceted genomics approach allows the isolation of the rice Pia-blast resistance gene consisting of two adjacent NBS-LRR protein genes. *Plant J.* **66,** 467–479 (2011).
31. Césari, S. *et al.* The NB-LRR proteins RGA4 and RGA5 interact functionally and physically to confer disease resistance. *EMBO J.* **33,** 1941–1959 (2014).
32. Ali, M. B. *et al.* Berry skin development in Norton grape: distinct patterns of transcriptional regulation and flavonoid biosynthesis. *BMC Plant Biol.* **11,** 7 (2011).
33. Serrano, M. *et al.* Repression of sucrose/ultraviolet B light-induced flavonoid accumulation in microbe-associated molecular pattern-triggered immunity in Arabidopsis. *Plant Physiol.* **158,** 408–422 (2012).
34. Treutter, D. Significance of flavonoids in plant resistance: a review. *Environ Chem Lett.* **4,** 147–157 (2006).
35. Chen, W. *et al.* Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat Genet.* **46,** 714–721 (2014).
36. Zhang, Z. *et al.* Splicing of Receptor-like kinase-encoding SNC4 and CERK1 is regulated by two conserved splicing factors that are required for plant immunity. *Mol Plant*. **7,** 1766–1775 (2014).
37. Xiang, C., Miao, Z. & Lam, E. DNA-binding properties, genomic organization and expression pattern of TGA6, a new member of the TGA family of bZIP transcription factors in Arabidopsis thaliana. *Plant Mol Biol.* **34,** 403–415 (1997).
38. Alves, M. S. *et al.* Plant bZIP transcription factors responsive to pathogens: a review. *Int J Mol Sci.* **14,** 7815–7828 (2013).
39. Wang, X. *et al.* The rpg4-mediated resistance to wheat stem rust (Puccinia graminis) in barley (Hordeum vulgare) requires Rpg5, a second NBS-LRR gene, and an actin depolymerization factor. *Mol Plant Microbe In*. **26,** 407–418 (2013).
40. Riehs-Kearnan, N., Gloggnitzer, J., Dekrout, B., Jonak, C. & Riha, K. Aberrant growth and lethality of Arabidopsis deficient in nonsense-mediated RNA decay factors is caused by autoimmune-like response. *Nucleic Acids Res.* **40,** 5615–5624 (2012).
41. Schnable, P. S. & Springer, N. M. Progress toward understanding heterosis in crop plants. *Annu Rev Plant Biol.* **64,** 71–88 (2013).
42. Guo, M. & Rafalski, J. A. Gene Expression and Heterosis in Maize Hybrids, in Polyploid and Hybrid Genomics (eds Z. J. Chen & J. A. Birchler) 59–84 (John Wiley & Sons, Inc., Oxford, UK, 2013).
43. Harper, A. L. *et al.* Associative transcriptomics of traits in the polyploid crop species Brassica napus. *Nat Biotechnol.* **30,** 798–802 (2012).
44. Gore, M. A. *et al.* A first-generation haplotype map of maize. *Science.* **326,** 1115–1117 (2009).
45. Haas, B. J. & Zody, M. C. Advancing RNA-seq analysis. *Nat Biotechnol.* **28,** 421–423 (2010).
46. Sekhon, R. S. *et al.* Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. *PLoS One.* **8,** e61005 (2013).
47. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* **30,** 2114–2120 (2014).
48. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* **9,** 357–359 (2012).
49. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
50. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics.* **21,** 1859–1875 (2005).
51. Pertea, G. *et al.* TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics.* **19,** 651–652 (2003).
52. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One.* **6,** e17288 (2011).
53. Conesa, A. & Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics.* 2008, 619832 (2008).
54. McDowall, J. & Hunter, S. InterPro protein classification. *Methods Mol Biol.* **694,** 37–47 (2011).
55. Boerner, S. & McGinnis, K. M. Computational identification and functional predictions of long noncoding RNA in Zea mays. *PLoS One.* **7,** e43047 (2012).
56. Jungo, F., Bougueleret, L., Xenarios, I. & Poux, S. The UniProtKB/Swiss-Prot Tox-Prot program: a central hub of integrated venom protein data. *Toxicon.* **60,** 551–557 (2012).
57. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* **29,** 2933–2935 (2013).
58. Burge, S. W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* **41,** D226–D232 (2012).
59. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol.* **537,** 39–64 (2009).

Acknowledgements

This research was supported by the National Natural Science Foundation of China (31123009, 31222041 and 31525017) and the National Hi-Tech Research and Development Program of China (2012AA10A307), the National Youth Top-notch Talent Support Program, the Fundamental Research Funds of ICS-CAAS (Grant to J.F.) and The Agricultural Science and Technology Innovation Program of CAAS.

Author Contributions

J.Y. and J.F. designed and supervised this study. M.J., H.L. and C.H. performed the data analysis. Y.X. provided the simulation code for pan-genome size estimation. G.W. and W.X. contributed to materials collection and suggested analysis procedure. M.J. and Y.W. performed the experiments. M.J., H.L., J.Y. and J.F. prepared the manuscript, and all the authors critically read and approved the manuscript.

Additional Information

Data availability: The raw RNA sequencing data have been deposited in NCBI Sequence Read Archive (SRA) under accession SRP026161. The raw sequences (with fasta format) and full annotation information of novel assembled genes, and together with the variation called from both novel genes and ePAV candidates, and related association mapping results could all be available at www.maizego.org/Resources.

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Jin, M. *et al.* Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Sci. Rep.* **6**, 18936; doi: 10.1038/srep18936 (2016).