The Plant Cell, Vol. 26: 121–135, January 2014, www.plantcell.org

LARGE-SCALE BIOLOGY ARTICLE

Insights into the Maize Pan-Genome and Pan-Transcriptome[W][OPEN]

Candice N. Hirsch,[a,b,1] Jillian M. Foerster,[c,2] James M. Johnson,[c,3] Rajandeep S. Sekhon,[c,d] German Muttoni,[c,4] Brieanne Vaillancourt,[a,b] Francisco Peñagaricano,[e] Erika Lindquist,[f] Mary Ann Pedraza,[f] Kerrie Barry,[f] Natalia de Leon,[c,d] Shawn M. Kaeppler,[c,d] and C. Robin Buell[a,b,5]

[a] Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824
[b] Department of Energy Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Michigan 48824
[c] Department of Agronomy, University of Wisconsin, Madison, Wisconsin 53706
[d] Department of Energy Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, Wisconsin 53706
[e] Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin 53706
[f] Department of Energy, Joint Genome Institute, Walnut Creek, California 94598

Genomes at the species level are dynamic, with genes present in every individual (core) and genes in a subset of individuals (dispensable) that collectively constitute the pan-genome. Using transcriptome sequencing of seedling RNA from 503 maize (*Zea mays*) inbred lines to characterize the maize pan-genome, we identified 8681 representative transcript assemblies (RTAs) with 16.4% expressed in all lines and 82.7% expressed in subsets of the lines. Interestingly, with linkage disequilibrium mapping, 76.7% of the RTAs with at least one single nucleotide polymorphism (SNP) could be mapped to a single genetic position, distributed primarily throughout the nonpericentromeric portion of the genome. Stepwise iterative clustering of RTAs suggests, within the context of the genotypes used in this study, that the maize genome is restricted and further sampling of seedling RNA within this germplasm base will result in minimal discovery. Genome-wide association studies based on SNPs and transcript abundance in the pan-genome revealed loci associated with the timing of the juvenile-to-adult vegetative and vegetative-to-reproductive developmental transitions, two traits important for fitness and adaptation. This study revealed the dynamic nature of the maize pan-genome and demonstrated that a substantial portion of variation may lie outside the single reference genome for a species.

INTRODUCTION

There is a large amount of natural phenotypic variation in plants, with some species demonstrating extreme phenotypic plasticity as evidenced from long-term selection experiments (Odhiambo and Compton, 1987; Dudley and Lambert, 1992; de Leon and Coors, 2002; Russell, 2006) and adaptation to changing environmental conditions. Understanding genomic and transcriptomic variation within a species can provide insights into phenotypic variation, plasticity, and environmental adaption. Initial studies exploring whole-genome variation in plant species focused primarily on single nucleotide polymorphism (SNPs), simple sequence repeats, and small insertions and deletions (Clark et al., 2007; Gore et al., 2009; McNally et al., 2009). Additionally, these studies focused largely on genomic variation for only those sequences present in a single reference genome sequence. With the advent of next-generation sequencing and the rapid reduction in sequencing costs, it is becoming increasingly feasible to determine the complete genomic content of many individuals within a species.

Initial work on resequencing individuals within a species was conducted in bacteria due to the relative simplicity of bacterial genomes compared with animal and plant genomes. These studies demonstrated that extensive variation in genome content could be seen between individuals within a species (Medini et al., 2005; Tettelin et al., 2005, 2008; Hogg et al., 2007). It is now understood that within a species there is a portion of the genome that is present in all individuals (termed the core genome) and a portion of the collective genomic content that is present in only a subset of individuals (termed the dispensable genome), the sum of which is the pan-genome for the species. The number of individuals required to capture the full pan-genome varies between species. After sequencing eight group B *Streptococcus* genomes, it was determined that the dispensable genome accounted for 33.4% of the pan-genome, and the size of the dispensable genome continually expanded with each

[1] Current address: Department of Agronomy and Plant Genetics, University of Minnesota, Saint Paul, MN 55108.
[2] Current address: DuPont Pioneer, Johnston, IA 50131.
[3] Current address: AgReliant Genetics, Lebanon, IN 46052.
[4] Current address: Monsanto Company, Lebanon, IN 46052.
[5] Address correspondence to buell@msu.edu.
The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) are: Natalia de Leon (ndeleongatti@wisc.edu), Shawn M. Kaeppler (smkaeppl@wisc.edu), and C. Robin Buell (buell@msu.edu)
[W] Online version contains Web-only data.
[OPEN] Articles can be viewed online without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.113.119982

additional genome (Tettelin et al., 2005). Based on modeling, the group B *Streptococcus* dispensable genome is expected to continue to expand even after sequencing hundreds of strains (Tettelin et al., 2005) with similar results observed with group A *Streptococcus*. By contrast, for *Bacillus anthracis*, only four strains were required to characterize the entire pan-genome (Tettelin et al., 2005).

The concept of the pan-genome has been observed on various levels in many plant species, such as *Arabidopsis thaliana*, maize (*Zea mays*), and rice (*Oryza sativa*) (Brunner et al., 2005; Morgante et al., 2007; Ossowski et al., 2008; Gore et al., 2009; Springer et al., 2009; Weigel and Mott, 2009; Lai et al., 2010; Swanson-Wagner et al., 2010; Cao et al., 2011; Gan et al., 2011; Chia et al., 2012). In maize, a comparison across four randomly chosen genomic regions revealed strong evidence for presence/absence variation (PAV) between the inbred lines B73 and Mo17 (Brunner et al., 2005; Morgante et al., 2007). For these four regions, only ~50% of the sequences were present in both genotypes. The 50% of the genome that was not in common was composed predominantly of transposable elements (TEs), although non-TE genes were also represented. Variation in gene copy number and PAV on a whole-genome scale between B73 and Mo17 for single-copy expressed genes has also been demonstrated (Springer et al., 2009). Similar results were also observed in an expanded panel of 34 maize and teosinte lines (Swanson-Wagner et al., 2010). Additionally, a study that generated over 32 Gb of sequence in the low-copy region of the genome across 27 diverse maize lines estimated that the B73 genome represents only ~70% of the available low-copy sequence in the maize pan-genome (Gore et al., 2009). More recently, a study of the maize pan-genome was conducted by resequencing six elite maize inbred lines important for commercial hybrid production in China and several hundred complete dispensable genes were identified (Lai et al., 2010). Similarly, the second-generation maize HapMap identified pervasive structural variation and showed that structural variants are enriched at loci associated with important traits (Chia et al., 2012). Even with the relatively small number of individuals sampled to date, there is strong evidence for the existence in maize of a core and a dispensable genome, which may be a major contributor to phenotypic variation seen in inbred lines and heterosis observed in hybrids.

Dispensable genes in bacteria are thought to contribute to diversity and adaptation (Medini et al., 2005; Tettelin et al., 2005, 2008; Kahlke et al., 2012). In plants, the timing of developmental transitions is important for evolution and adaptation to new environments; in crop species, such transitions have had an important role in crop domestication and cultivar improvement. In maize, two key events in development include the juvenile-to-adult vegetative and the vegetative-to-floral transitions, and components of pathways responsible for these developmental transitions have been characterized by qualitative genetic analyses. Cloned genes in maize involved in the juvenile-to-adult vegetative transition include *Corngrass1* (Chuck et al., 2007) and the AP2-like transcription factor *Glossy15*, which is a miR172 target (Moose and Sisco, 1996; Lauter et al., 2005). The signal transduction pathway regulating the juvenile-to-adult vegetative phase change includes common components among flowering plants (Wang et al., 2011). The autonomous pathway that involves miR156, miR172, and *SQUAMOSA PROMOTER BINDING-LIKE* genes and regulates both the vegetative and floral transitions is likely conserved in flowering plants (Poethig, 2009). Maize genes and noncoding regions involved in the timing of the vegetative-to-floral transition include *Vegetative to generative transition1*, *Zea CENTRORAPIALIS8*, and *indeterminate1* (Salvi et al., 2007; Lazakis et al., 2011). Substantial natural phenotypic variation has been reported for these traits in maize (Chardon et al., 2004; Salvi et al., 2007; Buckler et al., 2009; Chen et al., 2012; Xu et al., 2012). However, to date, the genetic architecture of these important developmental transitions has not been explored in the context of a species pan-genome.

Completed studies of the maize pan-genome have been limited to a few inbred lines and likely did not reveal the breadth of the collective diversity in maize. The goal of this study was to expand our knowledge of the maize pan-genome, assess the extent of variation across a large set of diverse inbred lines, determine the restricted or unrestricted nature of the maize pan-genome, and associate pan-genes with genetic variation and phenotypic diversity for traits important to crop adaptation.

RESULTS

Transcript Assembly and Characterization

We used whole-seedling transcriptome sequencing (RNA sequencing [RNA-seq]) on a panel of 503 diverse maize inbred lines representative of the major U.S. grain heterotic groups (Stiff Stalk Synthetic and Non-Stiff Stalk Synthetic), sweet maize, and popcorn, as well as exotic maize lines (Supplemental Data Set 1; Hansey et al., 2010) to characterize the maize pan-transcriptome as a proxy for the maize pan-genome. Seedling tissue was selected to maximize discovery of transcripts as it represents diverse organ types and has a higher number of expressed genes relative to other tissues (Sekhon et al., 2011). RNA-seq reads were mapped to the B73 reference genome sequence for SNP discovery and expression analysis of reference genes, while reads that did not map to the B73 reference genome sequence (3.2 billion total; Supplemental Data Set 1) were used for de novo assembly and identification of novel transcripts. A joint assembly of unmapped reads from all 503 inbred lines yielded 31,398 loci composed of 102,017 assembled transcripts with an N50 contig size of 911 bp (Supplemental Figure 1). Translations of the assembled transcripts were aligned to rice proteins. For the majority of the Oases loci that had a match to the rice proteins, the top match for all transcripts within the locus were to a single rice gene model and in most cases to a single protein, suggesting the isoforms encoded identical proteins (Supplemental Figure 2). To eliminate redundancy among the transcripts, for each locus, the representative transcript assembly (RTA), defined as the longest transcript within a locus, was used for further analysis.

The 31,398 RTAs were aligned to the B73 reference sequence to identify alleles and close paralogs, and 22,354 RTAs with alignments of >85% coverage and 85% identity were removed (Figure 1A). After removing an additional 363 potentially contaminant RTAs, 8681 high confidence RTAs that were absent

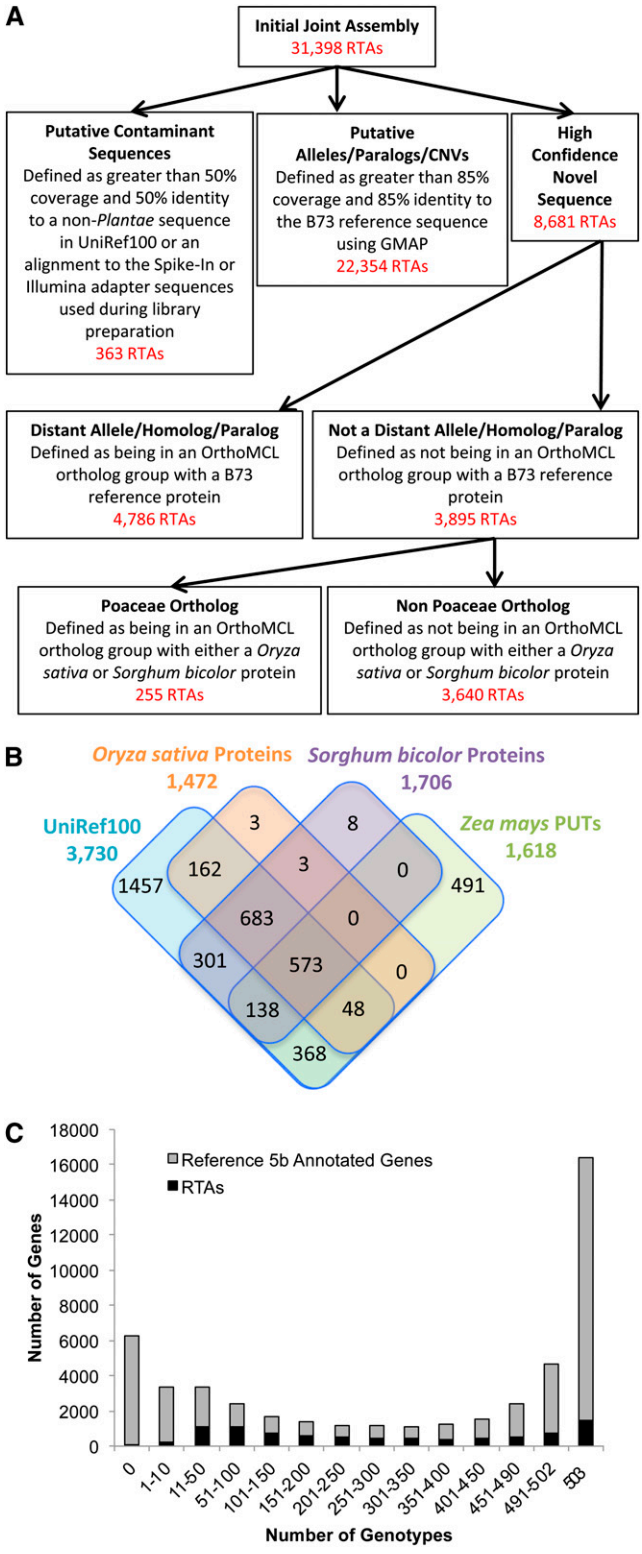

Figure 1. Flowchart, Support Statistics, and Gene Expression Distribution for the Joint Assembly across 503 Diverse Maize Inbred Lines.

from the B73 reference sequence were retained (Supplemental Data Set 2). Nearly 50% (4235) of these RTAs were supported by BLAST alignments to rice and sorghum (*Sorghum bicolor*) proteins (Ouyang et al., 2007; Paterson et al., 2009), the UniRef100 database (Suzek et al., 2007), and/or maize PlantGDB-assembled unique transcripts (Duvick et al., 2008) (Figure 1B; Supplemental Data Set 2), suggesting that these are not contaminants or artifacts generated from the sequencing and assembly process. Additionally, there was minimal evidence of transcripts being split across multiple RTAs as BLAST alignments to rice and sorghum identified only 228 RTAs that could be collapsed into 104 transcripts (Supplemental Table 1). Gene Ontology (GO) slim associations most frequently observed within the 8681 RTAs were cellular processes, membrane associated, and nucleotide binding domains (Supplemental Table 2). Due to the conservative alignment criteria used, as well as the fact that only seedling transcripts were surveyed, the 8681 RTAs identified are most likely an underrepresentation of the complete maize pan-genome. Furthermore, while the 503 inbred lines evaluated in this study are quite diverse (Supplemental Figure 3), they likely represent a fraction of the global maize diversity with respect to rare alleles in open pollinated populations and landraces.

Transcript Abundance Profiles in the Maize Pan-Genome

In addition to genomic level PAV, transcript level PAV has been observed in maize and is hypothesized to be important for phenotypic variation, including heterosis (Stupar and Springer, 2006; Swanson-Wagner et al., 2006; Lai et al., 2010; Hansey et al., 2012). To evaluate the extent of transcriptome level variation in maize, we estimated transcript abundance for the 503 inbred lines in the context of the maize pan-genome including the B73 reference gene models and RTAs. Plants for this experiment were grown in a controlled greenhouse environment. Replicates of B73 and Mo17 were used to assess any potential environmental effects. High average pairwise Pearson correlations of expression values between 31 B73 replicates (R^2 = 0.93) and 20 Mo17 replicates (R^2 = 0.95) indicate consistent environmental conditions throughout the experiment.

(A) Flowchart describing annotation of the RTAs from the joint assembly as putative contaminant sequences, putative alleles/homologs/paralogs/orthologs, or novel sequences.
(B) Support of the filtered RTAs. RTAs were searched against the UniRef100 database requiring a minimum E-value of 1e-5 and a minimum of 50% coverage and 50% identity using WU BLASTX, the *O. sativa* v7 proteins, and the *S. bicolor* v1 proteins requiring a minimum E-value of 1e-10 and a minimum of 70% coverage and 70% identity using WU BLASTX, and the maize PlantGDB-assembled unique transcripts (PUTs) version 171a requiring a minimum E-value of 1e-10 and a minimum of 85% coverage and 85% identity using WU BLAST.
(C) Distribution of gene expression in the maize seedling pan-transcriptome using a quantitative presence/absence classification. Genes were considered expressed if the fragments per kilobase of exon model per million fragments mapped 95% low confidence interval boundary as defined by Cufflinks was greater than zero.

Pairwise Pearson correlation coefficients of expression between samples were high (average R^2 = 0.89), indicating the overall seedling transcriptomes for these inbred lines were very similar. However, variation in transcript abundances and PAV among the inbred lines was apparent (Figure 1C). A total of 14,968 annotated B73 reference genes and 1425 RTAs were expressed in every line including B73 (essential/core transcripts), while 25,510 transcripts (18,327 annotated B73 reference genes and 7183 RTAs) were present in only a subset of lines representing potential dispensable transcripts or PAV in the genome. Of the 8681 novel RTAs, 4341 showed expression in the B73 reference inbred (1425 were expressed in all 503 individuals and 2916 were expressed in a subset of the individuals that included B73). The B73 reference sequence is a true draft sequence (Schnable et al., 2009), and the incomplete nature of the reference sequence is reflected in these results. Additionally, while samples were consistently harvested at the V1 developmental stage, it is possible that in some instances the observed PAVs reflect differences in developmental rate across the inbred lines.

Using a set of 10 diverse inbred lines, real-time PCR (RT-PCR) and PCR was performed to validate the computationally predicted PAV for 24 RTAs at the transcriptome and genome levels, respectively. RT-PCR results supported 81.1% of the transcriptome PAV predictions for the 18 primer pairs where at least one RT-PCR or PCR band was observed (Supplemental Data Set 3), with transcript sizes consistent with the computational prediction. Of the 18.9% nonconcordant predictions, 8.3% had computationally predicted expression, yet no band was observed, while 10.6% that were not predicted to be expressed had a RT-PCR band present. Technical limitations, such as primers designed in nonconserved regions of the transcripts due to alternative splicing, sequence divergence, and/or a lower dynamic range of RT-PCR compared with RNA-seq can explain the lack of complete concordance between the computational predictions and experimental validation. We hypothesized that the PAV observed at the transcriptome level for these RTAs in some cases may be due to genomic level PAV. Indeed, eight out of the 10 RTAs with transcript level PAV also showed genomic level PAV (Supplemental Data Set 3).

SNPs in the Maize Pan-Genome

A total of 1,628,790 SNPs were identified using RNA-seq reads mapped to the reference genome and the RTAs. To minimize sequence errors, at least two reads per individual were required, and an allele had to be present at a 5% frequency or higher in the population to be included in the final set. Of the ~1.6 million SNPs, 485,179 had <75% missing data and are hereafter referred to as the "working SNP set." These SNPs were present in 23,906 (60.6%) of the annotated B73 reference genes and 4429 (51.0%) of the RTAs, with a maximum of 173 SNPs in a single gene. Genetic distance clustering using the working SNP set distinctly clustered inbred lines within the pedigree-based subgroups (Supplemental Figure 3), which further validates the overall accuracy of the SNP set.

Linkage Disequilibrium Mapping of Representative Transcript Assemblies

Physical locations of the RTAs in the reference sequence were determined using a linkage disequilibrium (LD) mapping approach for RTAs with one or more SNPs in the working SNP set (Figure 2A). To test the accuracy of this approach, the analysis was performed using SNPs in a known B73 reference gene against all other SNPs anchored in the reference genome sequence. The gene (GRMZM2G057778) was mapped to an interval within two genes of its actual position in the reference sequence assembly (Supplemental Figure 4). Using stringent criteria, 76.7% (3396) of the RTAs with at least one SNP were LD mapped to a single physical position in the B73 reference genome. The LD-mapped RTAs were distributed throughout the genome with a lower frequency in pericentromeric regions (Figure 2B; Supplemental Data Set 4). The interval for the position of the LD-mapped RTAs ranged from 1 bp to 10.7 Mb with an average interval size of 2.0 Mb (Supplemental Figure 5). The average length of the 3396 LD-mapped RTAs was slightly larger than the remaining 5285 RTAs, with an average length of 1111 and 837 bp, respectively (Supplemental Figure 6). The average fragments per kilobase of exon model per million fragments mapped (FPKM) value across the 503 inbred lines was also slightly higher in the LD-mapped RTAs (average FPKM 5.9) versus the non-LD-mapped RTAs (average FPKM 1.3), both of which were lower than the average observed in the reference genes, which was 13.3 (Supplemental Figure 7). However, there was no bias in the distribution of RTAs regardless of whether they were expressed in all 503 inbred lines (core genes), in a subset of the lines including B73 (dispensable genes), or in a subset of the lines excluding B73 (dispensable genes). A total of 1033 RTAs with at least one SNP could not be LD mapped to a single location, attributable to the possibilities that the RTAs may be located in different positions across the inbred lines, the only SNP used for mapping within an RTA could have high amounts of missing data (up to 75%), or the genes adjacent to an RTA location could lack SNPs in the working SNP set and rapid decay in LD would prevent mapping of the RTA (Chia et al., 2012).

Syntenic relationships were used to provide further evidence for the placement of the RTAs on the physical map. Figure 2C shows an example of synteny of an RTA with three other grass species. RTA_10140 is located in a syntenic block consistent with its mapped location via LD mapping. Of the 8681 RTAs, 5041 were classified into orthologous groups based on protein similarity to *Poaceae* protein sequences (Figure 1A; Supplemental Data Set 5), with multiple examples of RTAs located at expected positions based on synteny. Many of the RTAs (4786) were in orthologous groups that also included a B73 protein. These RTAs are putative distant alleles (<85% coverage and 85% identity at the nucleotide level) or homologs/paralogs, with some showing evidence of possible tandem duplication based on the LD mapping.

Evaluation of the Restricted/Unrestricted Nature of the Maize Genome

In bacteria, the number of individuals/strains required to capture the full pan-genome varies between species with as few as four strains needed for some species (closed genome) and an infinite number of strains necessary for others (open genome) (Tettelin et al., 2005). The origin of dispensable genes in bacterial species and plants are likely mechanistically distinct, with those in bacterial species primarily originating from lateral gene transfer, and those in plants originating from gene loss, gene/whole-genome duplication

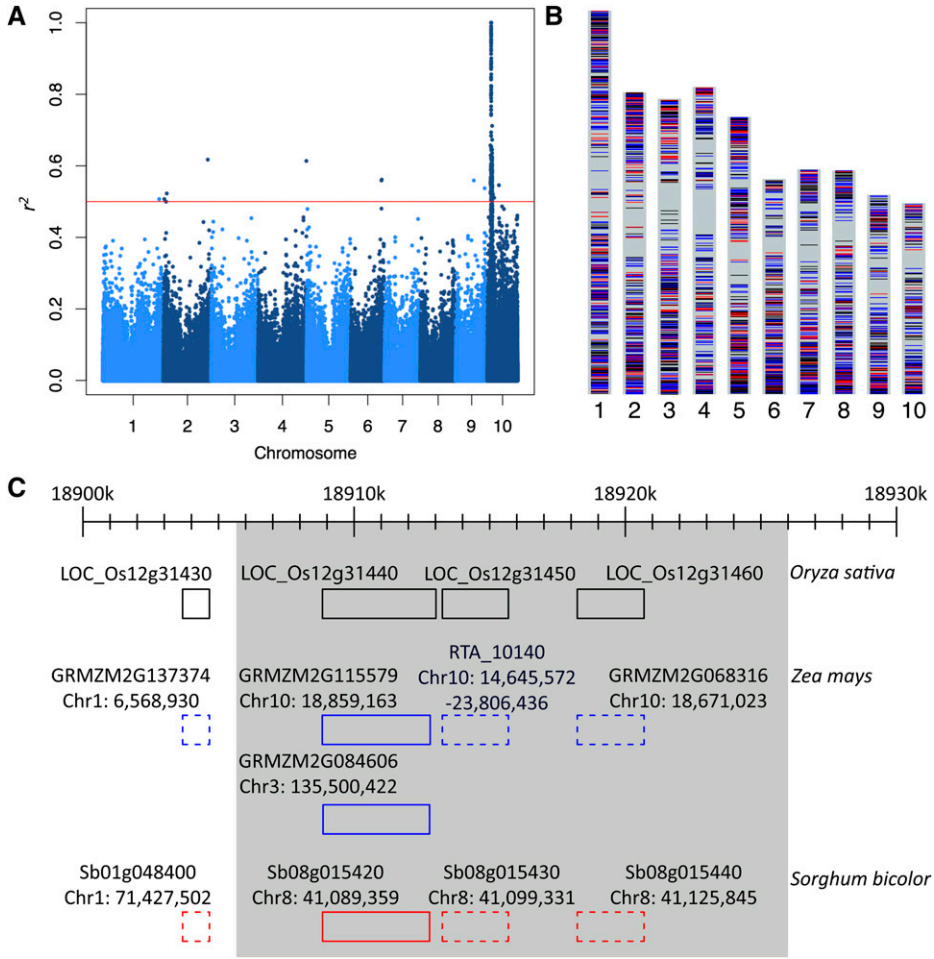

Figure 2. LD Mapping of RTAs.

(A) Pairwise r^2 values between the 458,259 SNPs anchored to the B73 maize v2 reference sequence and the 13 SNPs on RTA_10140.
(B) Pictorial representation of the maize chromosomes with the position of each of the 3396 LD-mapped RTAs. Colors of lines signify RTAs that are expressed in every inbred line (black), are not expressed in every inbred line but are expressed in B73 (blue), or are not expressed in every inbred line and are not expressed in B73 (red).
(C) Synteny analysis for RTA_10140 relative to the rice chromosome 12 sequence. Solid boxes show orthologous genes identified using EXONERATE with a minimum threshold of 70% identity over 70% of the length of the rice CDS sequence, and dashed boxes indicate orthologous genes identified using OrthoMCL with CDS sequences from rice v7, maize v2, and sorghum v1, and the 8681 RTAs.

and divergence, and transposon-mediated mechanisms. To assess if the maize genome is restricted (containing a discrete number of sequences) or unrestricted (containing an infinite number of sequences) with respect to the pan-genome, we used transcriptome assemblies from individual inbred lines as a proxy for their genomic content. Individual assemblies were generated for 366 inbred lines that had greater than three million unmapped reads (Supplemental Data Set 1), using only three million reads per individual assembly. On average, 62.5% of the reads from each line were assembled with an average of 8173 loci per line and an average transcript length of 454 bp. Using the RTA from each locus, alleles/paralogs/copy number variants (2,227,111 RTAs) and contaminant sequences (12,245) were filtered as described above yielding 747,977 total RTAs.

Stepwise addition of inbred lines from $n = 2$ to $n = 366$ in independent clustering runs showed a plateau in the total number of orthologous groups and singletons (24,129) in the collective dispensable transcriptome (Figure 3). These results demonstrate that within the context of this set of maize germplasm, the maize pan-genome, as estimated by the seedling transcriptome, is restricted. Further sampling of seedling RNA within this germplasm base will likely result in very minimal gene discovery after ~350 inbred lines. Thus, for the germplasm included in this study, the sampling depth was sufficient for the complete seedling pan-transcriptome. However, additional genes may still be discovered from the transcriptome of other tissues and different germplasm sources or whole-genome resequencing efforts.

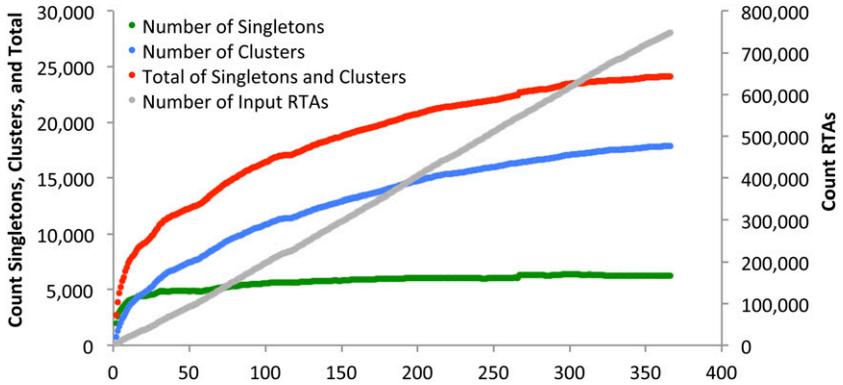

Figure 3. Evaluation of the Restricted/Unrestricted Nature of the Maize Pan-Genome.

Sequence-based clustering showing the number of RTAs plateaus at ~350 inbred lines. Clusters are defined as OrthoMCL groups with at least two RTAs.

Genetic Dissection of Traits Important for Fitness and Adaptation

We evaluated the utility of variants discovered from the RNA-seq reads to assess traits important for fitness and adaptation. The timing of developmental progression affects fitness of individuals within a species, thereby allowing adaptation to new environments. We explored natural variation for the juvenile-to-adult vegetative phase change and flowering time in our diversity panel to understand the relationship between genetic variation in the maize pan-genome and phenotypic diversity for traits important in crop adaptation and improvement.

Vegetative phase change was scored in 424 of the 503 inbred lines by identifying the last leaf with epicuticular wax. We observed significant natural variation for the last leaf with epicuticular wax, ranging from leaf 3.45 to 13.4 (Supplemental Figure 8 and Supplemental Data Set 6). The estimated entry-mean based heritability for last leaf with epicuticular wax was 0.53 based on phenotypic data measured across two years with two replications per year. Interestingly, the duration of juvenility was not correlated with flowering time.

A genome-wide association study (GWAS) was performed using the "GWAS SNP set" of 451,066 SNPs described in Methods, using a mixed linear model accounting for both population structure (P) and familial relatedness (K) (Yu et al., 2006). After a simple*M* correction for multiple testing (Gao et al., 2008), a significant association was detected on chromosome 9 (Figure 4A). The most significant SNPs were located within *glossy15*, an AP2-like transcription factor responsible for the expression of juvenile vegetative traits (Moose and Sisco, 1996). Twenty-nine SNPs in our GWAS SNP set were located within the *glossy15* gene model; nine of those were within the translated region and were evaluated for detrimental amino acid substitutions on protein function using a SIFT analysis (Kumar et al., 2009). Four SNPs caused nonsynonymous amino acid changes, including the most significant SNP detected in GWAS (Supplemental Table 3); however, none of the substitutions were predicted to alter protein function. In addition, none of the SNPs in our data set were located within the miRNA172 binding site of *glossy15* mRNA. This quantitative trait locus (QTL) explained 9.4% of the phenotypic variation and had an additive effect of ~0.48 leaves, relative to the minor allele.

Transcript abundance variation is an important component of phenotypic variation. Additionally, there is evidence that the developmental transition between the juvenile and adult vegetative phase is determined very early in development (Lauter et al., 2005; Chuck et al., 2007). As such, it is likely that transcript abundance variation at the seedling stage can partly explain phenotypic variation for this trait. To test this, GWAS was also performed with transcript abundance as the independent variable for all of the reference genes and RTAs for last leaf with epicuticular wax. The association analysis was done using the same mixed model as described above. One significant association was detected on chromosome 2 in a gene annotated as encoding a nuclear transcription factor Y subunit A-10 (GRMZM2G096016; Figure 4B). Using transcript PAV as the independent variable, this gene was the second most significant genome-wide. However, the P value did not exceed the significance threshold. Additionally, PCR analysis of 10 inbred lines revealed that this gene was indeed present in all 10 inbred lines.

Two measures of flowering time were available for 409 of the inbreds, growing degree days (GDDs) to silk and GDDs to pollen shed (Supplemental Data Set 6). Tropical lines were not included in this analysis, as they do not flower in Wisconsin. A Pearson correlation of 0.965 was observed between GDDs to pollen shed and GDDs to silk.

GWAS was performed for the flowering time traits using the GWAS SNP set and the parameters described above. Significant associations were detected for flowering time traits on chromosomes 1, 3, 4, 5, and 6 (Figure 5A; Supplemental Figure 9A and Supplemental Table 4). Association mapping was also conducted using transcript abundance as the independent variable, and significant associations were detected on chromosomes 1 and 3 (Figure 5B; Supplemental Figure 9B). The most significant SNPs on chromosome 3, explaining over 7% of the phenotypic variation, were in a CBS domain–containing protein (GRMZM2G171622). In a previous analysis of the maize nested association mapping population, a significant association in this region was also detected with a bootstrap posterior probability of 0.42 (Chen et al.,

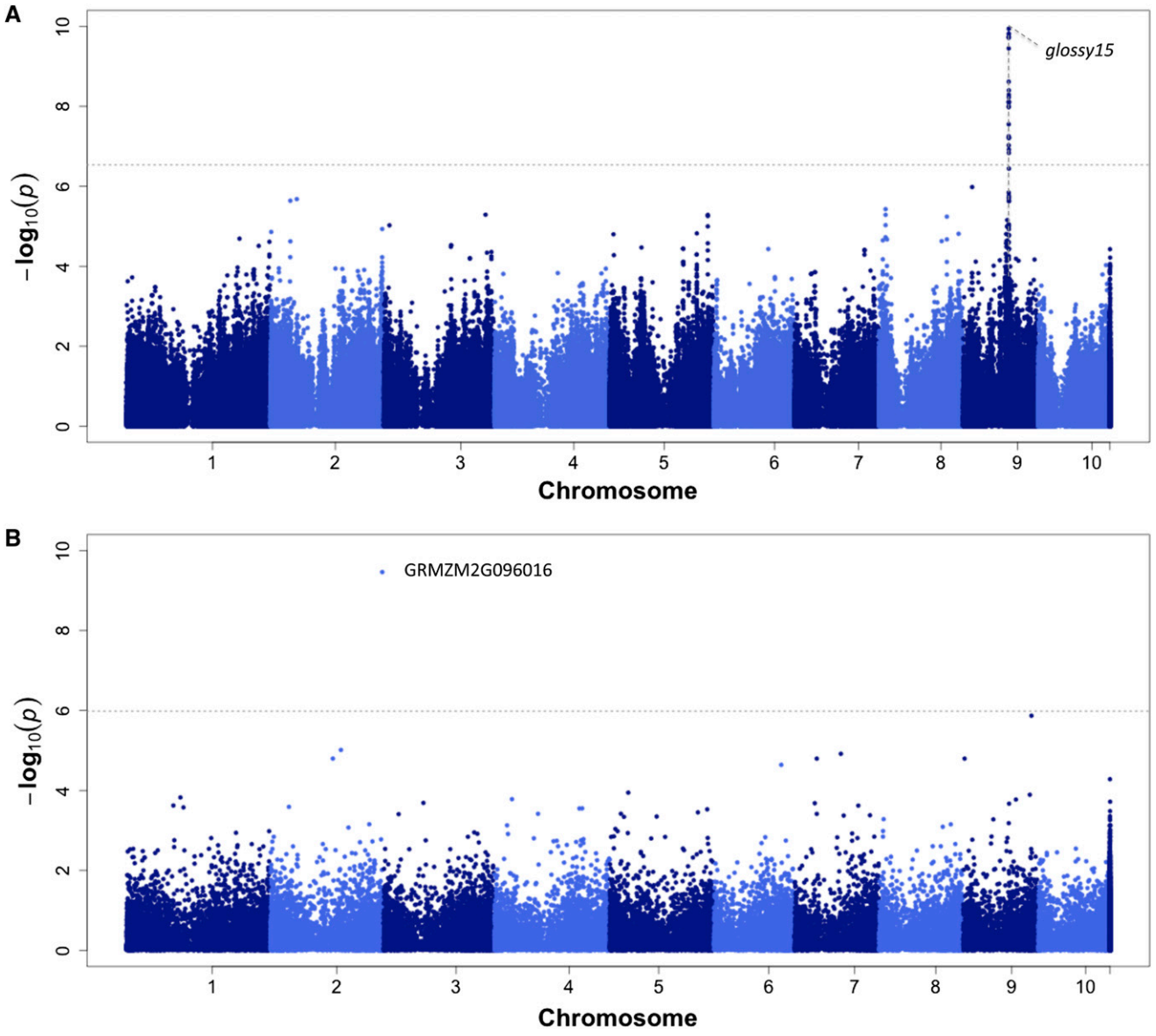

Figure 4. GWAS for Vegetative Phase Change Measured as Last Leaf with Epicuticular Wax.

(A) Manhattan plot of GWAS results using SNP markers. Significance threshold (horizontal dashed line) was set using the simple*M* method (2.7×10^{-7}). Significant SNPs were located in *glossy15*.

(B) Manhattan plot of GWAS results using transcript abundance as the independent variable. GRMZM2G096016 encodes a nuclear transcription factor Y subunit A-10 gene. Significance threshold was set using Bonferroni correction (1.04×10^{-6}).

2012). The gene adjacent to GRMZM2G171622 is annotated as encoding a MADS box transcription factor (GRMZM2G171650 [*zmm22*]) and contained only a single polymorphism within our data set and showed limited variation (minor allele frequency = 0.18). This SNP within *zmm22* was in high LD with SNPs found in the CBS domain–containing protein (Figure 5C), and *zmm22* was significant in the association analysis based on transcript abundance. Interestingly, *zmm22* has previously been shown to be important in maize cultivar improvement (Zhao et al., 2011), and this study provides evidence of its potential role in flowering time.

The determination of the timing of developmental phase changes, including flowering time, can occur very early in development, with known flowering time genes such as Zm-*Rap2.7* being expressed as early as 24 d after germination (Sekhon et al., 2011, 2013). Interestingly, 91.7% of the genes expressed in B73 V1 pooled leaves (FPKM > 5) were also expressed in the V13 immature tassel (Sekhon et al., 2011, 2013), lending further support to the notion that the seedling transcriptome can be relevant to traits measured later in development. In fact, based on seedling transcript abundance, we identified a MADS box

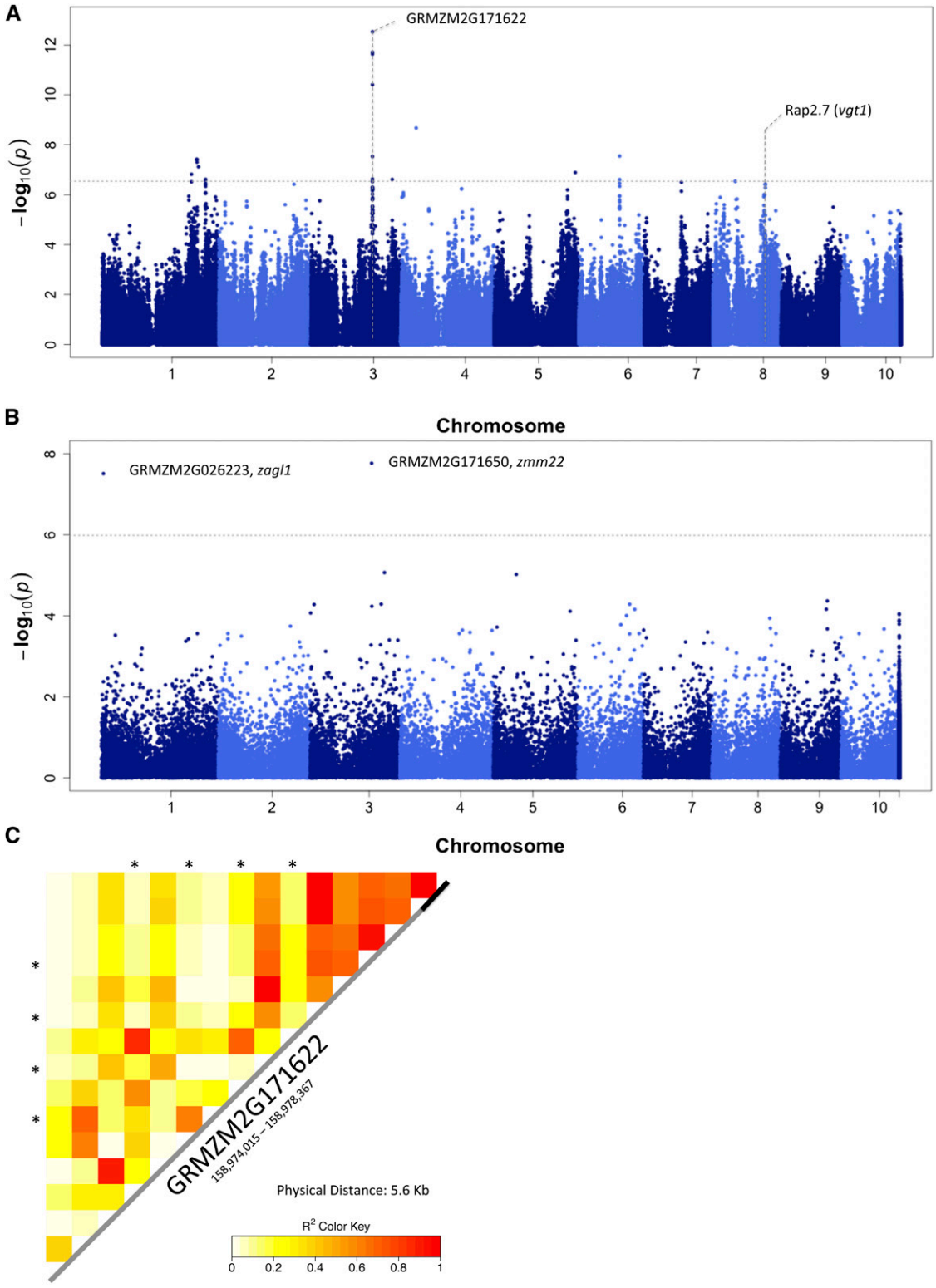

Figure 5. GWAS for GDDs to Pollen Shed.

(A) Manhattan plot of GWAS results using SNP markers. Genome-wide significance threshold (horizontal dashed line) was set using the simple*M* method (2.7×10^{-7}). GRMZM2G171622 encodes a CBS domain–containing protein.

transcription factor gene (GRMZM2G026223) on chromosome 1 that was associated with flowering time in our population. This MADS box transcription factor gene, *zagl1*, has previously been shown to have undergone selection during maize domestication (Zhao et al., 2011). Only one polymorphism was found in *zagl1* within our data set; therefore, it presents very limited sequence variation (minor allele frequency = 0.08), consistent with the lack of diversity expected under a domestication selection sweep. These results demonstrate the value of using transcript abundance in GWAS, especially for the identification of genes with limited SNP variation. No significant associations were identified using transcript PAV as the independent variable for either flowering time trait.

Zm-*Rap2.7*, located on chromosome 8, and the noncoding region 70 kb upstream (*vgt1*) have been shown to be involved in flowering time control (Salvi et al., 2007). SNPs identified within Zm-*Rap2.7* were not significant in our diversity panel at the genome-wide multiple testing threshold but were significant when tested as a candidate gene (P value = 3.47×10^{-5}).

DISCUSSION

In many plant species, genome duplication events and expansion and contraction of repetitive elements within the genome have fueled evolutionary diversity in some parts of the genome while other portions of the genome remain conserved, leading to the core and dispensable genomes. A previous comparison across four randomly chosen genomic regions between the maize inbred lines B73 and Mo17 revealed that only 50% of the sequences were present in both genotypes (Brunner et al., 2005; Morgante et al., 2007). Similarly, for the 8681 RTAs that were identified in this study, ~83% lacked sequence support at the transcriptome level for presence in all 503 lines.

Comparative studies across plant species have identified a large number of lineage-specific genes between genera or higher taxonomic orders (Bertioli et al., 2009; Cheung et al., 2009). By expanding the maize pan-genome using RNA-seq transcriptome data, we identified many syntelogs between rice, sorghum, and maize that were previously thought to be absent in maize based on the reference sequence of the single inbred line B73. It is known that the reference sequence for B73 is incomplete (Schnable et al., 2009; Lai et al., 2010; Hansey et al., 2012); thus, some of the sequences could eventually be identified through gap filling of the reference sequence. However, ~50% of the RTAs did not have evidence of being present in the B73 inbred line. The genome of maize, an ancient tetraploid, is thought to be composed of two distinct subgenomes, one of which (maize2) appears to have experienced greater gene loss (Schnable et al., 2011). The RTAs identified in this study represent some of the genes thought to be absent in one or both of the maize subgenomes.

Many studies have explored the relationship between genome-level variants and phenotypic traits through QTL mapping and GWAS. However, these studies have generally failed to explain the totality of phenotypic diversity, as genome-level variants typically can explain only a portion of the observed phenotypic diversity. In addition to identifying SNPs, RNA-seq allowed us to identify useful transcriptome variation in the context of the pan-genome. Using this robust data set, which included SNPs and transcriptome-level variation in the context of the maize pan-genome, we were able to increase our understanding of the genetics underlying adaptation and evolution by studying the progression of development in maize.

For juvenile-to-adult vegetative transition, at the genome level, we identified a highly significant gene (*glossy15*) that has previously been characterized as a qualitative mutation important for natural variation for this developmental progression (Moose and Sisco, 1996; Lauter et al., 2005). In addition, we also identified a gene (GRMZM2G096016) that would not have been discovered in traditional GWAS studies focused solely on SNP variation, as the genetic marker used to identify it was a transcript abundance marker.

For the flowering time traits, we found considerable overlap between our GWAS results and previous studies. The flowering time QTL on chromosome 3 detected in the maize nested association mapping population overlapped in position with significant genes in this study (Buckler et al., 2009), and with the population size and LD of this diversity panel, we were able to obtain greater genic resolution at this QTL. Additionally, two MADS box transcription factors, *zmm22* and *zagl1*, were identified as significantly associated with variation for flowering time. *zmm22* encodes a StMADS-11-like transcription factor, and this clade of proteins act as a repressor of flowering in several species including wheat (*Triticum aestivum*) and rice (Kane et al., 2005; Sentoku et al., 2005; Kikuchi et al., 2008). In addition, MADS box genes have been shown to be targets of selection during domestication and cultivar improvement (Zhao et al., 2011). *zagl1* was previously shown to have a reduction of genetic variation in maize landraces compared with teosinte, providing evidence of selection during domestication, while *zmm22* had genetic variation in both teosinte and maize landraces but decreased genetic variation in cultivated inbreds, demonstrating evidence of improvement selection on this gene (Briggs et al., 2007).

This study greatly expands our understanding of the dynamic nature of the maize pan-genome and demonstrates that a substantial portion of the variation underlying adaptive traits may lie outside a single reference genome sequence for a species. The identification of both known and previously unknown genes

Figure 5. (continued).

(B) Manhattan plot of GWAS results using gene expression level as the independent variable for GDDs to pollen shed. Significance threshold was set using Bonferroni correction (1.04×10^{-6}).

(C) LD heat map between the most significant gene on chromosome 3 based on SNP markers for GDDs to pollen shed, GRMZM2G171622, and a likely candidate gene, GRMZM2G171650. Asterisks indicate significant SNPs identified through GWAS.

associated with adaptive traits through traditional GWAS with SNP markers demonstrated the utility this RNA-seq–based SNP set for dissecting complex phenotypic traits. Additionally, using RNA-seq, we were able to identify genes associated with adaptive traits through transcript abundance and transcript PAV that would otherwise not be identified. The associated transcriptome PAVs did not translate to genomic PAVs in this study; however, there are many reasons that transcriptome PAV could be observed beyond deletion of an entire gene, such as epialleles (Makarevitch et al., 2007; Eichten et al., 2011) and variation outside of the gene model. While genomic-level PAVs were not identified in this study, it does not mean they are not important for phenotypic diversity. Indeed, PAV is an extreme form of copy number variation, and copy number variation of a *MATE1* gene was recently shown to be associated with aluminum tolerance (Maron et al., 2013). The results in this study highlight the importance of evaluating multiple levels of diversity when examining the genetic architecture of complex traits in plants.

METHODS

Plant Materials

A set of 503 diverse maize inbred lines that included 465 lines from the Wisconsin Diversity Set (Hansey et al., 2010) was evaluated (Supplemental Data Set 1). Plants were grown under greenhouse conditions (27°C/24°C day/night and 16 h light/8 h dark) with six plants per pot (30-cm top diameter, 28-cm height, 14.5 liters/volume) in Metro-Mix 300 (Sun Gro Horticulture) with no additional fertilizer. Whole-seedling tissue including roots at the V1 stage (Abendroth et al., 2011) from three plants per inbred line was pooled. RNA was isolated using Trizol (Invitrogen) and purified with the Qiagen RNeasy MinElute Cleanup kit. For obtaining DNA, seedling leaf tissue from 5 to 10 plants was bulked, and DNA was extracted using the cetyl(trimethyl)ammonium bromide method (Saghai-Maroof et al., 1984).

RNA-Seq Library Construction and Sequencing

Individual RNA-seq libraries were prepared for each of the genotypes in this study across seven plates. Libraries on the first plate were manually prepared without the use of spike-in sequences (Supplemental Data Set 1). Libraries on the subsequent plates were prepared using robotics and spike-in sequences were added to every other well using the ERCC RNA Spike-In Mix (Ambion) to assess the rate of cross contamination during library preparation.

For each library, polyadenylated RNA was isolated from 10 μg of total RNA using Dynabeads mRNA isolation kit (Invitrogen). This isolation was done twice to ensure the samples were free of rRNA contamination. The purified RNA was fragmented using RNA fragmentation reagents (Ambion) at 70°C for 3 min, targeting fragments ranging between 200 and 300 bp. The fragmented RNA was purified using Ampure SPRI beads (Agencourt). Reverse transcription was completed using SuperScript II reverse transcription (Invitrogen) with an initial annealing of random hexamer (Fermentas) at 65°C for 5 min, followed by an incubation of 42°C for 50 min and an inactivation step at 70°C for 10 min, and cDNA was purified with Ampure SPRI beads. Second-strand synthesis was performed using a deoxynucleotide triphosphate mix wherein dTTP was replaced by dUTP at 16°C for 2 h. Double-stranded cDNA fragments were purified using Ampure SPRI beads, blunt-ended, A tailed, and ligated with Truseq adaptors using the Illumina DNA sample prep kit. Adaptor-ligated DNA was purified using Ampure SPRI beads. To remove second-strand cDNA, dUTP was digested using AmpErase UNG (Applied Biosystems), and the digested cDNA was cleaned with Ampure SPRI beads, amplified for 10 PCR cycles with the Illumina Truseq primers, and then cleaned with Ampure SPRI beads. Sequencing was done on the Illumina HiSeq at the Joint Genome Institute (Walnut Creek, CA) to generate 100 nucleotide paired-end reads.

For the Oh43 sequence reads, ~5 μg of total RNA was used for library construction. The Oh43 RNA-seq library was not strand specific and did not contain spike-in sequences. Polyadenylated RNA purification, RNA fragmentation, cDNA synthesis, and PCR amplification was performed according to the Illumina RNA-seq protocol. Sequencing was performed on the Illumina HiSeq at the Research Technology Support Facility at Michigan State University to generate 50-nucleotide single-end reads.

RNA-Seq and Sample Quality Control Analysis

For each sequence library, read quality was evaluated using the FastQC software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Libraries with a fail flag such as low base quality scores were removed. Additionally, any library with less than five million reads was excluded. To determine the proportion of spike-in and vector sequences in each library, reads were mapped to the spike-in sequences and the UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) using Bowtie version 0.12.7 (Langmead et al., 2009) in the quality aware mode with the default parameters. Libraries that did not have spike-in sequences added were removed if >1% of the reads aligned to the spike-in sequences. For those libraries with spike-in sequences added, if >5% of the reads mapped, the library was removed from downstream analyses. For those libraries that passed this level of quality control, all reads that had an alignment to the spike-in/UniVec sequences were removed for future analysis. After this filtering, any libraries with less than five million reads were also removed from future analysis.

To quantify expression levels and identify SNPs, sequence reads for each library were mapped to the version 2 maize B73 reference sequence assembly (AGPv2, http://ftp.maizesequence.org/) (Schnable et al., 2009) using Bowtie version 0.12.7 (Langmead et al., 2009) and TopHat version 1.4.1 (Trapnell et al., 2009). Gene annotation was not provided during the read mapping step. For expression quantification, reads were mapped with a minimum intron size of 5 bp, a maximum intron size of 60,000 bp, and insertion/deletion detection disabled. Normalized gene expression levels were determined using Cufflinks version 1.3.0 (Trapnell et al., 2010) and the 5b filtered gene set (http://ftp.maizesequence.org/). Pearson correlation coefficients were calculated between each pair of genotypes using expression levels. A relatively high correlation was expected because the same tissue was harvested across the genotypes; thus, genotypes with R^2 values <0.5 across samples were removed. Additionally, any genotype with <60% reads mapping to the B73 reference sequence were removed from future analysis.

For SNP identification and genotyping, sequence reads from all individuals were first cleaned using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) with the fastx_clipper program requiring a minimum length of 30 bp. After clipping, reads were mapped with a minimum intron size of 5 bp, a maximum intron size of 60,000 bp, insertion/deletion detection disabled, and requiring a single unique hit. Alignments for reads that mapped uniquely were processed using the sort, index, and pileup programs within SAMtools version 0.1.12a (Li et al., 2009). To determine the genotype of an individual at a given position, a minimum of five reads was required. Additionally, the reads had to support only one allele, where support was defined as at least 5% of the reads and at least two reads. A locus was considered polymorphic if at least two alleles had >5% allele frequency. Finally, loci were removed that had >50% missing data (203,512 retained SNPs). If an individual had >40% missing data, that individual was removed from future analysis. Using the filtered SNPs, pairwise Roger's genetic distances (Rogers, 1972) were calculated, and an unweighted pair group method with arithmetic mean (UPGMA) tree was generated in PowerMarker version 3.25 (Liu and Muse, 2005) using default parameters and visualized in FigTree v1.3.1

(http://tree.bio.ed.ac.uk/software/figtree/). Integrity of the samples was evaluated based on manual comparison of relative positions on the tree with known pedigree relationships. RNA-seq reads were generated for 546 genotypes; the 503 genotypes discussed in this article included those that passed rigorous quality control filtering as described above.

Novel Transcript Assembly and Characterization

For the inbred lines that passed the quality control analysis described above, reads that could not be aligned to the B73 reference genome sequence using the parameters described for expression quantification were further cleaned using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Adapter sequences were removed using the fastx_clipper program, followed by artifact filtering with fastx_artifacts_filter and removal of low quality sequences with fastq_quality_trimmer requiring a minimum quality score of 20. A minimum sequence length of 30 nucleotides was required for all of the programs. From each of the 503 genotypes, 400,000 cleaned reads (201,200,000 total reads) were assembled using an assembly pipeline consisting of Velvet version 1.2.07 (Zerbino and Birney, 2008) and Oases version 0.2.08 (Schulz et al., 2012). The preliminary assembly generated by Velvet was used as the input for Oases. A kmer of 27 and a minimum transcript length of 500 bp was used.

RTAs were mapped to AGPv2 using the splice site aware aligner GMAP version 2012-04-21 (Wu and Watanabe, 2005). Any RTA with an alignment >85% coverage and 85% identity was removed from future analysis. RTAs were searched against the spike-in sequences and Illumina adapter sequences using WU BLASTN (Altschul et al., 1990) requiring a minimum E-value of 1e-10, and any RTAs with an alignment to the database were removed from future analysis. Additionally, RTAs were filtered against the UniRef100 database release 2012_07 (Suzek et al., 2007) using WU BLASTX (Altschul et al., 1990; Gish and States, 1993) using the shortqueryok option, requiring a minimum E-value of 1e-5, a seed word length of 4, a neighborhood word score of 1000, and a minimum of 50% coverage and 50% identity. Any RTA with a best hit to a non-*Plantae* sequence was removed from downstream analyses.

For assembly quality analysis and annotation, the remaining RTAs were searched against the *O. sativa* version 7 proteins (Ouyang et al., 2007) and the *S. bicolor* version 1.0 proteins (Paterson et al., 2009) using WU BLASTX (Altschul et al., 1990; Gish and States, 1993), and the maize PlantGDB assembled unique transcripts version 171a (Duvick et al., 2008) using WU BLASTN (Altschul et al., 1990). For all BLAST searches, a minimum E-value of 1e-10 was used and only the best hits were considered. For the BLASTX searches, a minimum coverage and identity of 70% was used. For the search against the maize PUTs, a minimum coverage and identity of 85% were used. Functional annotations were assigned based on these BLAST analyses (Supplemental Data Set 2). To assign GO slim terms (Gene Ontology Consortium, 2010), the RTAs were translated using ESTScan version 3.0.3 (Iseli et al., 1999) using the default parameters. InterProScan version 4.8 (Zdobnov and Apweiler, 2001) with the following parameters was used to determine GO terms: –cli –nocrc –iprlookup –goterms –appl hmmpfam –altjobs –seqtype p –format raw. The map2slim script within the Perl GO::Parser module was used to reduce the GO terms to GO slim terms.

Transcript Abundance Profile Analysis in the Maize Pan-Genome

Sequence reads for each library were mapped to AGPv2 plus the 8681 unfiltered RTAs using Bowtie version 0.12.7 (Langmead et al., 2009) and TopHat version 1.4.1 (Trapnell et al., 2009), and normalized gene expression levels were determined using Cufflinks version 1.3.0 (Trapnell et al., 2010) with the parameters described for RNA-seq and sample quality control analysis. To characterize transcript PAV, sequence reads were mapped requiring a unique alignment. A gene/RTA was defined as expressed if the FPKM low confidence interval as described by Cufflinks (Trapnell et al., 2010) was greater than zero.

Confirmation of RTA and Presence Absence Variation

Transcript assembly, expression PAV, and genomic PAV for 24 random RTAs across 10 inbred lines (B14-SSS, B37-SSS, B73-SSS, C103-NSS, Mo17-NSS PHN11-NSS, CML 322–exotic, NC358-exotic, P39–sweet maize, and HP301-popcorn) were evaluated using RT-PCR and PCR. The 24 RTAs were computationally predicted to have expression in 1 to 10 of the lines and each had a minimum length of 750 bp. For this confirmation, RTAs were sampled from the complete set of 8681 RTAs. One microgram of the same total RNA used for the RNA-seq library construction was reverse transcribed into cDNA using the SuperScript III first-strand synthesis system (Invitrogen) according to the manufacturer's protocol, and ~70 ng cDNA was used per RT-PCR reaction. Genomic PAV was evaluated in the 24 RTAs for the same 10 inbred lines using 50 ng of genomic DNA per PCR reaction. The thermocycler program for RT-PCR and PCR included 95°C for 4.0 min of initial denaturation, followed by 30 cycles of 95°C for 1.0 min, 55°C for 1.0 min, and 72°C for 1.5 min, followed by 72°C for 7 min of final extension. Primer sequences are available in Supplemental Data Set 3.

SNP Analysis in the Expanded Maize Pan-Genome

Reads were cleaned and mapped, and the genotype scores at each SNP position were determined as described in the RNA-seq and sample quality control analysis section above. Loci with >75% missing data were removed. Pairwise Roger's genetic distances (Rogers, 1972) were calculated, and a UPGMA was generated in PowerMarker version 3.25 (Liu and Muse, 2005) using default parameters and visualized in FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

LD Mapping of Representative Transcript Assemblies

LD mapping was used to determine the position of the RTAs relative to the AGPv2 genome sequence. Pairwise r^2 values were calculated between each of the SNPs in AGPv2 (458,259) and RTA (26,920) sequences using PLINK version 1.07 (Purcell et al., 2007). An all-versus-all BLAST with the RTAs, rice version 7 CDS sequences (Ouyang et al., 2007), the sorghum version 1.0 CDS sequences (Paterson et al., 2009), and the maize version 2 CDS sequences (Schnable et al., 2009) was performed using WU TBLASTX (Altschul et al., 1990) requiring a minimum E-value of 1e-10 and allowing up to 500 hits per sequence. Orthologous gene families were identified by OrthoMCL version 1.4 (Li et al., 2003; Chen et al., 2007) in mode 4 with default parameters.

Evaluation of the Restricted/Unrestricted Nature of the Maize Genome

Reads that could not be aligned to AGPv2 using the parameters described for expression quantification were used to generate individual assemblies for each inbred line that had a minimum of three million unmapped reads after cleaning using the methods described for the joint assembly ($n = 366$; Supplemental Data Set 1). Velvet version 1.2.07 (Zerbino and Birney, 2008) and Oases version 0.2.08 (Schulz et al., 2012) were used for the assembly with a kmer of 27 and minimum transcript size of 250 bp. RTAs were filtered as described above for the joint assembly; a total of 747,977 RTAs were retained. An all-versus-all blast was performed using WU BLASTN (Altschul et al., 1990) requiring a minimum E-value of 1e-10 and allowing up to 5000 hits per sequence. Sequence-based clustering of the individual assemblies was performed using OrthoMCL version 1.4 (Li et al., 2003; Chen et al., 2007) in mode 4 with default parameters. Stepwise addition of lines from $n = 2$ to $n = 365$ lines was then performed

using OrthoMCL version 1.4 (Li et al., 2003; Chen et al., 2007) in mode 5 with default parameters.

Association Mapping Analysis

Last leaf with epicuticular wax, GDDs to silk, and GDDs to pollen shed were measured on plants grown in a randomized complete block design with two replications across two years at the Arlington Agriculture Research Station in Wisconsin. To keep an accurate leaf count for measuring, the last leaf with epicuticular wax a hole was punched in the 5th leaf at the V7 stage and a collar was placed around the stalk between leaf 8 and 9 at the V10 stage. The last leaf showing epicuticular wax was scored at the V14 stage. Days to pollen and days to silk were scored when 50% or greater of the plants in a plot had visible pollen shed or silk emergence, respectively. A linear model was used for phenotypic analysis, $Y_{ijk} \sim \mu + Genotype_i + Rep(Year)_{j(k)} + Year_k + Year_k \times Genotype_i + e_{ijk}$, where Y was the phenotypic value of the ith genotype (Genotype) in the jth replicate (Rep) within the kth year (Year), Year × Genotype is the interaction term between these two variables, and μ was the overall mean. All effects, except the overall mean, were considered random.

From the 485,179 high confidence SNPs identified in the "LD Mapping of Representative Transcript Assemblies" section above, nonbiallelic positions were filtered. Additionally SNPs in non-RTA sequences with >30% missing data were removed. For the remaining 438,222 non-RTA SNPs, the population-based haplotype clustering algorithm of Scheet and Stephens, implemented via fastPHASE software version 1.4.0 (Scheet and Stephens, 2006), was used to impute missing genotype scores. Default settings for all parameters of the algorithm were used to impute the missing genotypes. For the RTA SNPs, SNPs with <60% missing data were retained (12,844 SNPs). Imputation was not performed on these SNPs due to the large LD window size that several of the RTAs were mapped to. The collective set of imputed reference SNPs (438,222 SNPs) and filtered RTA SNPs (12,844 SNPs) comprise a total of 451,066 SNPs referred to as "GWAS SNP set" was used for GWAS analysis.

GWAS was performed with 424 and 409 of the 503 original inbred lines for last leaf with epicuticular wax and the flowering time traits, respectively, using a previously proposed mixed linear model (Yu et al., 2006), which is as follows: $y = X\beta + Wm + Pv + Zu + e$, where y is a vector of phenotypic observations, β is a vector of fixed effects other than the SNP under testing (year and rep effects), m is a vector of SNP (fixed effect), v is the vector of population effects (fixed effect), u is a vector of polygene background random effect (proportion of the breeding values not accounted for by the SNP marker), and e is a vector of residual effects. P is an incidence matrix of principal component scores (eigenvectors) of marker-allele frequencies (Patterson et al., 2006), and X, W, and Z are incidence matrices of ones and zeros relating y to β, m, and u, respectively. The covariance of u is equal to KVA, where K is the kinship matrix that was estimated with a random set of SNPs according to the VanRaden method (VanRaden, 2008) and VA is the additive variance estimated using restricted maximum likelihood. The kinship matrix estimations and compressed mixed linear model (Zhang et al., 2010) were performed with the GAPIT R package (Lipka et al., 2012). Quality of the GWAS model fit was evaluated with QQ plots. To account for multiple testing, without being overly conservative and control for the Type-II error rate, we used the simple*M* method (Gao et al., 2008). Due to the linkage observed between SNP markers, the total number of markers being tested does not reflect the number of independent tests. The simple*M* approach applies a Bonferroni correction to the actual number of independent tests, or the effective number of independent tests (*M*eff), by considering the LD between each pair of markers and applying principal component analysis to obtain the eigenvalues. The simple*M* method has been shown to be an effective way to control the experiment-wise error rate in GWAS (Gao et al., 2010; Johnson et al., 2010). In this study, the *M*eff was 172,470 (equal to the number of eigenvalues necessary to explain 99.0% of the variance). LD could not be accurately measured for the SNPs in the RTAs, and for these SNPs, the simple*M* method was not applied. Including the total SNPs in the RTAs (12,844) and the 172,470 effective tests in the reference sequence SNPs, the Bonferroni threshold considering independent tests was 0.05/185,314 [i.e., 2.7×10^{-7} (α_e = 0.05)]. For the GWAS using gene expression or PAV as the independent variable, the test for each gene was considered independent, resulting in 48,137 total tests and a Bonferroni threshold of 1.04×10^{-6}.

Glossy15 protein sequences for each of the 503 inbred lines were generated based on SNPs within the GWAS SNP set. The SIFT BLink algorithm was used to evaluate damaging amino acid changes on protein function (Kumar et al., 2009). The cloned W64A sequence was used as the functional sequence input (GenBank accession number GI:1732031; Moose and Sisco, 1996).

Accession Numbers

Sequence data from this article can be found in the Sequence Read Archive and the National Center for Biotechnology Information under accession number PRJNA189400. Multifasta files of joint and individual transcript assembly sequences, gene expression values, SNP genotype scores for the filtered SNP set, imputed GWAS SNP set genotype scores, and GWAS results are available for download from the Dryad Digital Repository (http://doi.org/10.5061/dryad.r73c5).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Assembled Transcript Size Distribution for the Joint Assembly across 503 Diverse Maize Inbred Lines.

Supplemental Figure 2. Number of Rice Proteins and Genes per Oases Defined Locus.

Supplemental Figure 3. Unweighted Pair Group Method with Arithmetic Mean Tree of 503 Diverse Maize Inbred Lines Based on 485,179 Working Set SNP Markers.

Supplemental Figure 4. Linkage Disequilibrium Mapping of an Example B73 Reference Gene to the B73 Reference Genome Sequence.

Supplemental Figure 5. Linkage Disequilibrium Mapping Interval Size Distribution.

Supplemental Figure 6. Distribution of Mapped versus Unmapped Representative Transcript Assemblies.

Supplemental Figure 7. Relationship between Reference Gene and Representative Transcript Assembly Size and Average Gene Expression across 503 Inbred Lines.

Supplemental Figure 8. Natural Variation for Vegetative and Floral Transition Traits Measured on a Set of Diverse Inbred Lines.

Supplemental Figure 9. Manhattan Plots of Genome-Wide Association Analysis Results for Growing Degree Days to Silk.

Supplemental Table 1. Joint Representative Transcript Assemblies with Evidence of Transcript Splitting.

Supplemental Table 2. Gene Ontology Slim Terms for the 8681 Filtered Representative Transcript Assemblies.

Supplemental Table 3. Single Nucleotide Polymorphisms in the Genome-Wide Association Study SNP Set That Causes Amino Acid Changes in the *glossy15* Protein.

Supplemental Table 4. Summary of Growing Degree Days to Pollen Shed Genome-Wide Association Study Results.

Supplemental Data Set 1. Read Number, Mapping Information, and Individual Assembly Status for the 503 Maize Inbred Lines Included in This Study.

Supplemental Data Set 2. Annotation of the 8681 Joint Assembly Representative Transcript Assemblies.

Supplemental Data Set 3. Primers Used for Confirmation of the Joint Assembly Representative Transcript Assemblies and Transcriptome and Genome Presence/Absence Variation.

Supplemental Data Set 4. Linkage Disequilibrium Mapping of Representative Transcript Assemblies.

Supplemental Data Set 5. OrthoMCL Results from Analysis of *Zea mays* v2, *Oryza sativa* v7, and *Sorghum bicolor* v1 CDS Sequences and the Joint Assembly Representative Transcript.

Supplemental Data Set 6. Phenotypic Data for 424 Inbred Lines for Last Leaf with Epicuticular Wax and 409 Lines for Growing Degree Days to Pollen Shed and Growing Degree Days to Silk.

ACKNOWLEDGMENTS

This work was funded by the Department of Energy Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

C.N.H., C.R.B., N.D.L., and S.M.K. designed the research. B.V., C.N.H., E.L., J.M.F., K.B., M.A.P., and R.S.S. performed research. B.V., C.N.H., F.P., G.M., J.M.F., and J.M.J. analyzed data. C.N.H., C.R.B., J.M.F., N.D.L., and S.M.K. wrote the article.

Received October 27, 2013; revised January 3, 2014; accepted January 9, 2014; published January 31, 2014.

REFERENCES

Abendroth, L.J., Elmore, R.W., Boyer, M.J., and Marlay, S.K. (2011). Corn Growth and Development. (Ames, IA: Iowa State University Extension).

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

Bertioli, D.J., et al. (2009). An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. BMC Genomics **10:** 45.

Briggs, W.H., McMullen, M.D., Gaut, B.S., and Doebley, J. (2007). Linkage mapping of domestication loci in a large maize teosinte backcross resource. Genetics **177:** 1915–1928.

Brunner, S., Fengler, K., Morgante, M., Tingey, S., and Rafalski, A. (2005). Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell **17:** 343–360.

Buckler, E.S., et al. (2009). The genetic architecture of maize flowering time. Science **325:** 714–718.

Cao, J., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. **43:** 956–963.

Chardon, F., Virlon, B., Moreau, L., Falque, M., Joets, J., Decousset, L., Murigneux, A., and Charcosset, A. (2004). Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. Genetics **168:** 2169–2185.

Chen, C., DeClerck, G., Tian, F., Spooner, W., McCouch, S., and Buckler, E. (2012). PICARA, an analytical pipeline providing probabilistic inference about a priori candidates genes underlying genome-wide association QTL in plants. PLoS ONE **7:** e46596.

Chen, F., Mackey, A.J., Vermunt, J.K., and Roos, D.S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE **2:** e383.

Cheung, F., Trick, M., Drou, N., Lim, Y.P., Park, J.Y., Kwon, S.J., Kim, J.A., Scott, R., Pires, J.C., Paterson, A.H., Town, C., and Bancroft, I. (2009). Comparative analysis between homoeologous genome segments of *Brassica napus* and its progenitor species reveals extensive sequence-level divergence. Plant Cell **21:** 1912–1928.

Chia, J.M., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. **44:** 803–807.

Chuck, G., Cigan, A.M., Saeteurn, K., and Hake, S. (2007). The heterochronic maize mutant Corngrass1 results from overexpression of a tandem microRNA. Nat. Genet. **39:** 544–549.

Clark, R.M., et al. (2007). Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science **317:** 338–342.

de Leon, N., and Coors, J.G. (2002). Twenty-four cycles of mass selection for prolificacy in the golden glow maize population. Crop Sci. **42:** 325–333.

Dudley, J.W., and Lambert, R.J. (1992). Ninety generations of selection for oil and protein in maize. Maydica **37:** 1–7.

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M.D., Lawrence, C.J., Lushbough, C., and Brendel, V. (2008). PlantGDB: A resource for comparative plant genomics. Nucleic Acids Res. **36:** D959–D965.

Eichten, S.R., et al. (2011). Heritable epigenetic variation among maize inbreds. PLoS Genet. **7:** e1002372.

Gan, X., et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature **477:** 419–423.

Gao, X., Becker, L.C., Becker, D.M., Starmer, J.D., and Province, M.A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. Genet. Epidemiol. **34:** 100–105.

Gao, X., Starmer, J., and Martin, E.R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet. Epidemiol. **32:** 361–369.

Gene Ontology Consortium (2010). The Gene Ontology in 2010: Extensions and refinements. Nucleic Acids Res. **38:** D331–D335.

Gish, W., and States, D.J. (1993). Identification of protein coding regions by database similarity search. Nat. Genet. **3:** 266–272.

Gore, M.A., Chia, J.M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S., Ross-Ibarra, J., Ware, D.H., and Buckler, E.S. (2009). A first-generation haplotype map of maize. Science **326:** 1115–1117.

Hansey, C.N., Johnson, J.M., Sekhon, R.S., Kaeppler, S.M., and de Leon, N. (2010). Genetic diversity of a maize association population with restricted phenology. Crop Sci. **51:** 704–715.

Hansey, C.N., Vaillancourt, B., Sekhon, R.S., de Leon, N., Kaeppler, S. M., and Buell, C.R. (2012). Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. PLoS ONE **7:** e33071.

Hogg, J.S., Hu, F.Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J.C., and Ehrlich, G.D. (2007). Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol. **8:** R103.

Iseli, C., Jongeneel, C.V., and Bucher, P. (1999). ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 138–148.

Johnson, R.C., Nelson, G.W., Troyer, J.L., Lautenberger, J.A., Kessing, B.D., Winkler, C.A., and O'Brien, S.J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics **11:** 724.

Kahlke, T., Goesmann, A., Hjerde, E., Willassen, N.P., and Haugen, P. (2012). Unique core genomes of the bacterial family vibrionaceae: Insights into niche adaptation and speciation. BMC Genomics **13:** 179.

Kane, N.A., Danyluk, J., Tardif, G., Ouellet, F., Laliberté, J.F., Limin, A.E., Fowler, D.B., and Sarhan, F. (2005). TaVRT-2, a member of the StMADS-11 clade of flowering repressors, is regulated by vernalization and photoperiod in wheat. Plant Physiol. **138:** 2354–2363.

Kikuchi, R., Sage-Ono, K., Kamada, H., Handa, H., and Ono, M. (2008). PnMADS1, encoding an StMADS11-clade protein, acts as a repressor of flowering in *Pharbitis nil*. Physiol. Plant. **133:** 786–793.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. **4:** 1073–1081.

Lai, J., et al. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. Nat. Genet. **42:** 1027–1030.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10:** R25.

Lauter, N., Kampani, A., Carlson, S., Goebel, M., and Moose, S.P. (2005). MicroRNA172 down-regulates glossy15 to promote vegetative phase change in maize. Proc. Natl. Acad. Sci. USA **102:** 9412–9417.

Lazakis, C.M., Coneva, V., and Colasanti, J. (2011). ZCN8 encodes a potential orthologue of Arabidopsis FT florigen that integrates both endogenous and photoperiod flowering signals in maize. J. Exp. Bot. **62:** 4833–4842.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics **25:** 2078–2079.

Li, L., and Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. **13:** 2178–2189.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., and Zhang, Z. (2012). GAPIT: Genome association and prediction integrated tool. Bioinformatics **28:** 2397–2399.

Liu, K., and Muse, S.V. (2005). PowerMarker: An integrated analysis environment for genetic marker analysis. Bioinformatics **21:** 2128–2129.

Makarevitch, I., Stupar, R.M., Iniguez, A.L., Haun, W.J., Barbazuk, W.B., Kaeppler, S.M., and Springer, N.M. (2007). Natural variation for alleles under epigenetic control by the maize chromomethylase zmet2. Genetics **177:** 749–760.

Maron, L.G., et al. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc. Natl. Acad. Sci. USA **110:** 5241–5246.

McNally, K.L., et al. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc. Natl. Acad. Sci. USA **106:** 12273–12278.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. Curr. Opin. Genet. Dev. **15:** 589–594.

Moose, S.P., and Sisco, P.H. (1996). Glossy15, an APETALA2-like gene from maize that regulates leaf epidermal cell identity. Genes Dev. **10:** 3018–3027.

Morgante, M., De Paoli, E., and Radovic, S. (2007). Transposable elements and the plant pan-genomes. Curr. Opin. Plant Biol. **10:** 149–155.

Odhiambo, M.O., and Compton, W.A. (1987). Twenty cycles of divergent mass selection for seed size in Corn1. Crop Sci. **27:** 1113–1116.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res. **18:** 2024–2033.

Ouyang, S., et al. (2007). The TIGR Rice Genome Annotation Resource: Improvements and new features. Nucleic Acids Res. **35:** D883–D887.

Paterson, A.H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. Nature **457:** 551–556.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. **2:** e190.

Poethig, R.S. (2009). Small RNAs and developmental timing in plants. Curr. Opin. Genet. Dev. **19:** 374–378.

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., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. **81:** 559–575.

Rogers, J.S. (1972). Measure of genetic similarity and genetic distance. Studies in Genomics. VII. Univ. Tex. Publ. **7213:** 145–153.

Russell, W.K. (2006). Registration of KLS_30 and KSS_30 populations of maize. Crop Sci. **46:** 1405–1406.

Saghai-Maroof, M.A., Soliman, K.M., Jorgensen, R.A., and Allard, R.W. (1984). Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA **81:** 8014–8018.

Salvi, S., et al. (2007). Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. USA **104:** 11376–11381.

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. **78:** 629–644.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA **108:** 4069–4074.

Schnable, P.S., et al. (2009). The B73 maize genome: Complexity, diversity, and dynamics. Science **326:** 1112–1115.

Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics **28:** 1086–1092.

Sekhon, R.S., Briskine, R., Hirsch, C.N., Myers, C.L., Springer, N.M., Buell, C.R., de Leon, N., and Kaeppler, S.M. (2013). Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. PLoS ONE **8:** e61005.

Sekhon, R.S., Lin, H., Childs, K.L., Hansey, C.N., Buell, C.R., de Leon, N., and Kaeppler, S.M. (2011). Genome-wide atlas of transcription during maize development. Plant J. **66:** 553–563.

Sentoku, N., Kato, H., Kitano, H., and Imai, R. (2005). OsMADS22, an STMADS11-like MADS-box gene of rice, is expressed in non-vegetative tissues and its ectopic expression induces spikelet meristem indeterminacy. Mol. Genet. Genomics **273:** 1–9.

Springer, N.M., et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. **5:** e1000734.

Stupar, R.M., and Springer, N.M. (2006). Cis-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. Genetics **173:** 2199–2210.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: Comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23:** 1282–1288.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D., and Springer, N.M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. **20:** 1689–1699.

Swanson-Wagner, R.A., Jia, Y., DeCook, R., Borsuk, L.A., Nettleton, D., and Schnable, P.S. (2006). All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. Proc. Natl. Acad. Sci. USA **103:** 6805–6810.

Tettelin, H., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA **102:** 13950–13955.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: The bacterial pan-genome. Curr. Opin. Microbiol. **11:** 472–477.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics **25:** 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28:** 511–515.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. **91:** 4414–4423.

Wang, J.W., Park, M.Y., Wang, L.J., Koo, Y., Chen, X.Y., Weigel, D., and Poethig, R.S. (2011). miRNA control of vegetative phase change in trees. PLoS Genet. **7:** e1002012.

Weigel, D., and Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol. **10:** 107.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21:** 1859–1875.

Xu, J., Liu, Y., Liu, J., Cao, M., Wang, J., Lan, H., Xu, Y., Lu, Y., Pan, G., and Rong, T. (2012). The genetic architecture of flowering time and photoperiod sensitivity in maize as revealed by QTL review and meta analysis. J. Integr. Plant Biol. **54:** 358–373.

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., Kresovich, S., and Buckler, E.S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38:** 203–208.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan—An integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:** 847–848.

Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:** 821–829.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., and Buckler, E.S. (2010). Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. **42:** 355–360.

Zhao, Q., Weber, A.L., McMullen, M.D., Guill, K., and Doebley, J. (2011). MADS-box genes of maize: Frequent targets of selection during domestication. Genet. Res. **93:** 65–75.