Multiple Maize Reference Genomes Impact the Identification of Variants by Genome-Wide Association Study in a Diverse Inbred Panel

Joseph L. Gage, Brieanne Vaillancourt, John P. Hamilton, Norma C. Manrique-Carpintero, Timothy J. Gustafson, Kerrie Barry, Anna Lipzen, William F. Tracy, Mark A. Mikel, Shawn M. Kaeppler,* C. Robin Buell,* and Natalia de Leon

J.L. Gage, W.F. Tracy, S.M. Kaeppler, N. de Leon, Dep. of Agronomy, Univ. of Wisconsin–Madison, 1575 Linden Drive, Madison, WI 53706; J.L. Gage (current address), USDA ARS, 538 Tower Road, Ithaca, NY 14853; S.M. Kaeppler, N. de Leon, Dep. of Energy Great Lakes Bioenergy Research Center, Univ. of Wisconsin–Madison, 1575 Linden Drive, Madison, WI 53706; S.M. Kaeppler, Wisconsin Crop Innovation Center, Univ. of Wisconsin–Madison, 8520 University Green, Middleton, WI 53562; B. Vaillancourt, J.P. Hamilton, N.C. Manrique-Carpintero, C.R. Buell, Dep. of Plant Biology, Dep. of Energy Great Lakes Bioenergy Research Center, and Plant Resilience Institute, Michigan State Univ., 612 Wilson Road, East Lansing, MI 48824; Timothy J. Gustafson, Monsanto Company, 7202 Portage Road, DeForest, WI 53532; K. Barry, A. Lipzen, Dep. of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598; Mark A. Mikel, Dep. of Crop Sciences, Univ. of Illinois, 1206 W. Gregory Dr., Urbana, IL 61801.

ABSTRACT Use of a single reference genome for genome-wide association studies (GWAS) limits the gene space represented to that of a single accession. This limitation can complicate identification and characterization of genes located within presence–absence variations (PAVs). In this study, we present the draft de novo genome assembly of 'PHJ89', an 'Oh43'-type inbred line of maize (*Zea mays* L.). From three separate reference genome assemblies ('B73', 'PH207', and PHJ89) that represent the predominant germplasm groups of maize, we generated three separate whole-seedling gene expression profiles and single nucleotide polymorphism (SNP) matrices from a panel of 942 diverse inbred lines. We identified 34,447 (B73), 39,672 (PH207), and 37,436 (PHJ89) transcripts that are not present in the respective reference genome assemblies. Genome-wide association studies were conducted in the 942 inbred panel with both the SNP and expression data values to map *Sugarcane mosaic virus* (SCMV) resistance. Highlighting the impact of alternative reference genomes in gene discovery, the GWAS results for SCMV resistance with expression values as a surrogate measure of PAV resulted in robust detection of the physical location of a known resistance gene when the B73 reference that contains the gene was used, but not the PH207 reference. This study provides the valuable resource of the Oh43-type PHJ89 genome assembly as well as SNP and expression data for 942 individuals generated from three different reference genomes.

Abbreviations: AUDPC, area under the disease progress curve; B73v4, version 4 of the B73 reference genome; ePAV, presence or absence of expression; FPKM, fragments per kb of exon model per million mapped reads; GWAS, genome-wide association studies; PAV, presence–absence variations; RTAs, representative transcript assemblies; SCMV, *Sugarcane mosaic virus*; SNP, single nucleotide polymorphism; WiDiv-942, expanded version of the Wisconsin diversity panel.

CORE IDEAS

- We present the draft de novo genome assembly of the 'Oh43'-type maize inbred 'PHJ89'.
- Expression and single nucleotide polymorphism data were generated separately from three different references.
- Genome-wide association studies' candidate gene predictions vary depending on the reference used.

Pangenomes, originally described in the bacterial species *Streptococcus agalactiae* (Tettelin et al., 2005), are becoming increasingly recognized for their prevalence in plant species. Consisting of core and dispensable genes that are present in all and some individuals, respectively, the pangenome is a description of all genic content for a given species. Descriptions of microbial and phytoplankton pangenomes (Tettelin et al., 2005; Donati

Citation: Gage, J.L., B. Vaillancourt, J.P. Hamilton, N.C. Manrique-Carpintero, T.J. Gustafson, K. Barry, A. Lipzen, W.F. Tracy, M.A. Mikel, S.M. Kaeppler, C.R. Buell, and N. de Leon. 2018. Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. Plant Genome 12:180069. doi: 10.3835/plantgenome2018.09.0069

Received 17 Sept. 2018. Accepted 4 Feb. 2019.
*Corresponding authors (smkaeppl@wisc.edu; buell@msu.edu).

et al., 2010; Read et al., 2013; Liu et al., 2014; Zhou et al., 2014) have motivated the generation of pangenomes of plant species such as maize (Hirsch et al., 2014), rice (*Oryza sativa* L.) (Schatz et al., 2014), *Brassica rapa* L. (Lin et al., 2014), soybean [*Glycine max* (L.) Merr.] (Lam et al., 2010), *Brachypodium distachyon* (L.) P.Beauv. (Gordon et al., 2017), and bread wheat (*Triticum aestivum* L.) (Montenegro et al., 2017).

On a genome-wide scale, PAVs have been shown to be involved in an array of agronomically relevant traits in maize such as plant architecture, flowering, and disease resistance (Walker et al., 1995; Chia et al., 2012; Lu et al., 2015) and are hypothesized to contribute to heterosis (Springer et al., 2009). Standard methods for GWAS rely on associating SNPs with phenotypic variability. This can complicate the detection of PAVs associated with phenotypes, because of a lack of SNPs in the PAV region for individuals with the absence allele. In addition, phenotypes associated with regions that are absent in the reference genome may be mapped to genomic regions in linkage disequilibrium with the PAV; however, associations will not be identified if there is an absence of SNPs in linkage disequilibrium with the PAV.

Maize displays considerable variation in genome content between individuals. A study of 2.8 Mb of sequence in two inbred lines, B73 and 'Mo17', found that less than half of the sequence was collinear between the two and that more than one-third of genes were unique to one inbred (Brunner et al., 2005). Further genome-wide characterization of the differences between B73 and Mo17 identified thousands of PAV sequences present in B73 and not Mo17, along with hundreds of copy-number variants (Springer et al., 2009). Subsequent studies have identified substantial amounts of sequence that do not align to the B73 reference, which is estimated to represent about 70% of the low-copy sequence in the maize pangenome (Gore et al., 2009). Previous characterization of the maize pangenome by transcriptome sequencing (pantranscriptome) of 503 diverse maize inbred lines identified more than 8500 high confidence transcripts that are not present in the B73 reference genome assembly, almost half of which were supported by BLAST alignments to other species (Hirsch et al., 2014). These studies underscore the importance of evaluating the maize pangenome as well as considering the implications of using a single inbred as the reference genome when aligning sequences and calling genotypes.

The first maize inbred line sequenced, B73 (Schnable et al., 2009), is a representative of a group of germplasm referred to as the Stiff Stalks, named after the Iowa Stiff Stalk Synthetic population, from which many founding Stiff Stalk lines were derived (Reif et al., 2005). Inbred Stiff Stalk lines are frequently crossed with germplasm from other groups, such as the Lancasters and Iodents (Duvick et al., 2004; Reif et al., 2005), to create vigorous and high-yielding hybrids. Heterosis in the offspring of crosses between these germplasm groups may be driven partly by genes that display single-parent expression (Baldauf et al., 2018) or by gene content differences between the inbred parents (Springer et al., 2009). PH207, a representative of the Iodent group that is present in the pedigree of hundreds of modern Iodent lines (Mikel, 2011), was recently sequenced de novo by Hirsch et al. (2016). Several other inbred reference genomes have been made public recently, including 'W22' (Springer et al., 2018), 'Mo17' (Sun et al., 2018), 'CML247' (Lu et al., 2015), 'EP1', and 'F7' (Unterseer et al., 2017). Generating more de novo genome sequences of individuals representing other groups of maize germplasm will be useful for identifying genotype–phenotype associations located on PAVs and for future studies of heterosis.

In this study, we present the draft de novo genome assembly of PHJ89, an Oh43-type inbred line developed by Pioneer Hi-Bred, Inc, Johnston, IA. We also use RNA sequencing data of an expanded version of the Wisconsin diversity panel (Hirsch et al., 2014, Mazaheri et al., 2019), called the WiDiv-942, to generate individual pantranscriptomes in parallel from three reference genomes: version 4 of the B73 reference genome (B73v4) (Jiao et al., 2017), PH207 (Hirsch et al., 2016), and PHJ89 (this study). We separately created SNP and whole-seedling gene expression datasets for all 942 members of the WiDiv-942 panel from each of the three reference genomes and their pantranscriptomes. To demonstrate the utility of these datasets, we used SNPs and gene expression values as explanatory variables in GWAS for SCMV resistance, for which a causative gene is known to reside within a PAV (Liu et al., 2017).

MATERIALS & METHODS

Generation of a Reference Genome Sequence for PHJ89, an Inbred Representative of Oh43-Type Germplasm

PHJ89 Genome Sequencing and Assembly

The Oh43-type ex- Plant Variety Protection inbred line PHJ89 is identified in the US National Plant Germplasm System as PI 548798 and was formerly protected under US Plant Variety Protection number 9100092. Two separate DNA isolations were performed with leaf tissue from 10 individual plants at the V3 stage with cetyl trimethyl ammonium bromide (Murray and Thompson, 1980; Saghai-Maroof et al., 1984). One DNA isolation was used to make three mate-pair libraries at the University of Illinois Biotechnology Center with three separate size selections of 2 to 4, 6 to 8, and 12 to 15 kb. The libraries were constructed with the Nextera Mate Pair Library Sample Prep kit (Illumina, San Diego, CA), followed by the TruSeq DNA Sample Prep kit (Illumina). The mate-pair libraries were sequenced at the University of Illinois on the Illumina HiSeq 2500 to 160 nt in paired-end mode. The second DNA isolation was used to make two whole-genome sequencing libraries at the Joint Genome Institute, Walnut Creek, CA; the library sizes were ~400 and ~600 bp. These two libraries were sequenced on the Illumina HiSeq 2500 to 251 nt in paired-end mode at the Joint Genome Institute.

Paired-end libraries were processed with Cutadapt (version 1.9.1; Martin, 2011) to remove adapters and low-quality sequence with the parameters -q 10 -m 200. Cleaned paired-end reads were error-corrected with the ALLPATHS-LG (version 52400; Gnerre et al., 2011) standalone error correction pipeline with the default parameters. Error-corrected reads from the 400 bp library were then merged with FLASH (version 1.2.11; Magoc and Salzberg, 2011) with the parameters -r 250 -f 500 -s 150. Mate-pair libraries were first processed with NextClip (version 1.3.1; Leggett et al., 2014) with a minimum retained read length of 31 nt; reads containing the junction adaptor were concatenated and cleaned with Cutadapt (-q 10 -m 31). The processed reads (Supplemental File S1) were then assembled with ABySS (version 1.9.0; Simpson et al., 2009) with a kmer length of 127, a mate-pair read minimum alignment length of 31 bp, and a minimum unitig length of 500 bp. To assess genome quality, genomic reads were cleaned with Cutadapt (version 1.18; Martin 2011) with the parameters -n 2 -m 200 -u 1 -q 10 and were aligned to the PHJ89 genome assembly with BWA-MEM (version 0.7.17; Li 2013) with the default parameters. Read alignment counts were obtained with SAMtools (version 1.9; Li et al., 2009).

PHJ89 RNA-Sequencing and Genome-Guided Transcript Assembly

Two RNA-Seq libraries from whole seedlings at the V1 stage, including roots, were made and sequenced by the Joint Genome Institute, Walnut Creek, CA. The library (identifier AYOWB) was sequenced on the Illumina HiSeq 2500 to 151 nt in paired-end mode, whereas the library (identifier ZWGA) was sequenced on the Illumina HiSeq 2500 to 150 nt in paired end mode. RNA-Seq reads were cleaned with Cutadapt (version 1.14, Martin 2011) with the parameters -m 100 -q 10 (Supplemental File S2). Cleaned reads were aligned to the PHJ89 assembly with TopHat2 (version 2.1.1; Kim et al., 2013) in strand-specific mode with a maximum intron size of 10 kb. The resulting RNA-Seq alignments were then assembled into transcripts with Trinity (version 2.1.1; Grabherr et al., 2011) in strand-specific, genome-guided mode with a maximum intron size of 10 kb and a minimum contig length of 500 bp.

PHJ89 Genome Annotation

A custom repeat library was created with RepeatModeler (version 1.0.8; http://www.repeatmasker.org/, accessed 9 Apr. 2019) using scaffolds greater than 50 kb. Protein coding genes were removed from the custom repeat library with ProtExcluder (version 1.1; Campbell et al., 2014) and a curated database of plant protein coding genes. The final custom repeat library was compiled by combining the filtered custom repeat library, a curated maize repeat library (v5a; maizesequence.org, accessed 9 Apr. 2019), and Viridiplantae repeats from Repbase version 2015-08-07 (https://www.girinst.org/repbase/, accessed 16 Apr. 2019). The assembly was masked with RepeatMasker (version 4.0.6; Chen, 2004) with the final custom repeat library. To annotate the gene models, Augustus (version 3.1; Stanke et al., 2006) was trained with the RNA-Seq alignments from the ZWGA library and the gene models predicted on the repeat-masked assembly with any scaffolds less than 1 kb removed. Gene models were then improved with PASA2 [version 2.0.2 https://github.com/PASApipeline/PASApipeline/wiki, accessed 9 Apr. 2019 (Haas et al., 2003)], with the genome-guided transcript assemblies as evidence, to create the working set of gene models. High-confidence gene models were identified by removing gene models that lacked expression evidence or a Pfam domain, were partial, or had an internal stop codon. Functional annotation was transitively assigned to the working model set by sequence similarity to the *Arabidopsis thaliana* (L.) Heynh. proteome (TAIR10; Arabidopsis.org, accessed 9 Apr. 2019), Swiss-Prot (Boeckmann et al., 2003), and Pfam version 29 (Punta et al., 2012).

Orthologous and paralogous gene families were constructed with the software OrthoFinder version 1.1.5 (Emms and Kelly, 2015) with the annotated proteomes of the B73v4, PH207, and PHJ89 genomes with and without the three respective predicted proteomes of the representative transcript assemblies (RTAs) (described in the following section). Reciprocal best hits were identified by searching the B73v4 representative proteins against the PH207 and PHJ89 representative proteins with NCBI BLAST+ BLASTP (version 2.8.1; Altschul et al., 1990) with an e-value cutoff of 1×10^{-10}. To define the potential gene function of the pangenome, proteins from the representative transcripts from *A. thaliana*, *B. distachyon*, *O. sativa*, *Setaria viridis* (L.) P.Beauv., *Sorghum bicolor* (L.) Moench., and *Vitis vinifera* L. were downloaded from Phytozome version 12 and searched against the annotated proteomes of B73v4, PH207, and PHJ89 and the three respective predicted proteomes of the RTAs with BLAST version 2.2.26. Pfam domains (version 31.0) were identified with HMMER version 3.1b1 (http://hmmer.org/, accessed 9 Apr. 2019; Eddy 1998).

Pantranscriptome Analyses

RNA-Seq Libraries and Read Processing

Whole-seedling RNA-Seq reads from a total of 959 inbreds from two previous studies (Hirsch et al., 2014; Mazaheri et al., 2019) were used to identify the maize pantranscriptome (Supplemental File S3). FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accessed 9 Apr. 2019) was used to assess read quality. Reads were treated as single-ended throughout the entire quality and mapping pipeline. Reads were first mapped to spike-in sequences (ERCC RNA Spike-In Mix, ThermoFisher Scientific, Waltham, MA) and the UniVec database (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/, accessed 9 Apr. 2019) with Bowtie version 0.12.7 (Langmead et al., 2009) with the default parameters; if libraries did not have spike-in added during library preparation and had > 1% of sequences mapped to spike-in/UniVec, then the

sample was removed. Likewise, libraries with spike-in added during library preparation which had >5% spike-in or UniVec were removed. Adapters and low-quality sequences were removed from reads with Cutadapt version 1.8 (Martin, 2011) with the parameters -n 5, -q 20, 20, and -m 30. If >20% of the reads were removed by Cutadapt, the library was discarded. For consistency in read length across all samples, reads were trimmed to 100 nt. PolyA/T tails were removed with Cutadapt version 1.8 (Martin, 2011) with the following parameters: -m 30, -n 4, and -O 20. If a library had ≤5 million reads after all cleaning, the library was removed. After all library filtering and quality control, a total of 959 inbreds proceeded through the downstream components of the pipeline.

RNA-Seq Alignments to Three Reference Genomes

Cleaned reads were mapped to the B73v4 (Zm-B73-REFERENCE-GRAMENE-4.0-; Jiao et al., 2017), PH207 (Zm-PH207-REFERENCE NS-UIUC UMN-1.0-; Hirsch et al., 2016), and PHJ89 version 1 reference sequences (this study), without the mitochondrial and plastid sequences with TopHat2 (version 2.0.14; Kim et al., 2013) and Bowtie2 (version 2.2.3; Langmead and Salzberg, 2012) with the following parameters: -i 5, -I 60000, and –no-novel-indels.

De Novo Assembly of Representative Transcript Assemblies

Unmapped reads from alignment of the 959 inbred RNA-Seq datasets to each of the three respective reference genomes (Supplemental File S4) were separated into unmapped pairs and unmapped singletons and normalized with the in silico read normalization utility provided by Trinity version 2.2.0 (Grabherr et al., 2011) allowing a maximum kmer coverage of 30. For the B73v4-derived novel transcripts, Trinity version 2.2.0 was run with a minimum kmer count of 2, a minimum contig length of 500, and a group pairs distance of 500 with 55,044,661 pairs and 181,395,509 singletons. For the novel PH207-derived transcripts, the same Trinity parameters were used with 55,101,707 pairs and 178,754,708 singletons; for the novel PHJ89-derived transcripts, 55,234,686 pairs and 180,843,326 singletons were used with the same parameters. To remove assembled transcripts of high similarity, the longest isoform per gene was extracted with the Perl script provided by Trinity: get_longest_isoform_seq_per_trinity_gene.pl. Transcripts were then aligned to the cognate reference genome and the organellar genomes with GMAP (version 2012–04–21; Wu and Watanabe, 2005). To remove transcripts that represented the alleles of annotated genes in each of the cognate reference genomes, a transcript was discarded if it aligned to the cognate reference or organellar genome at ≥85% identity and coverage. For the three reference genomes, 66,756 (B73v4), 62,211 (PH207), and 64,188 (PHJ89) transcripts were removed.

Contaminants were removed from the GMAP-filtered transcripts with BLAST+ version 2.50 (Camacho et al., 2009) with an e-value cutoff of 1×10^{-5} and the following BLAST programs or modes and databases: megablastn Spike-in (ThermoFisher Scientific), megablastn UniVec (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/, accessed 9 Apr. 2019), blastn-short Illumina adapters (oligonucleotide sequences, Illumina), megablastn NCBI nt (a partly nonredundant nucleotide sequence database; ftp://ftp.ncbi.nlm.nih.gov/blast/db/, accessed 9 Apr. 2019), and blastx NCBI nr (a nonredundant protein sequence database; ftp://ftp.ncbi.nlm.nih.gov/blast/db/, accessed 9 Apr. 2019). For the spike-in and UniVec search, transcripts were removed if the percent identity was ≥95% and either 50% of the query or 50% of the subject was covered. All transcripts with a hit to the Illumina adapters were removed. For the NCBI nt search, transcripts with ≥95% identity, ≥50% coverage of the subject or query, and a hit to a non-Viridiplantae were removed. Transcripts with ≥50% identity, ≥50% coverage of the query or subject, and non-Viridiplantae were removed on the basis of the NCBI nr search. The contaminant searches resulted in removal of 49,045 (B73v4), 49,658 (PH207), and 49,449 (PHJ89) transcripts. To reduce redundancy of the novel transcripts, CD-HIT version 4.6 (Fu et al., 2012; Li and Godzik, 2006) was run with a sequence identity threshold of 95%, resulting in a final set of 34,447 (B73v4), 39,672 (PH207), and 37,436 (PHJ89) RTAs.

To determine whether the RTAs identified by the three reference genomes represented diverged alleles in other maize inbreds, we aligned each set of RTAs against B73v4 (Jiao et al., 2017), CML247 (Lu et al., 2015), F7 (Unterseer et al., 2017), EP1 (Unterseer et al., 2017), PH207 (Hirsch et al., 2016), and PHJ89 (this study) using GMAP (v2012–04–21; Wu and Watanabe, 2005). If an RTA aligned with ≥55% coverage and ≥85% identity, it was considered present in the genome.

Expression Abundance

Cleaned reads were aligned to the B73v4, PH207, and PHJ89 genomes plus their respective set of RTAs with TopHat2 (version 2.0.14; Kim et al., 2013) and Bowtie2 (version 2.2.3; Langmead and Salzberg, 2012) with the following settings: -i 5, and -I 60000. Cufflinks version 2.2.1 (Trapnell et al., 2010) was used to calculate normalized gene expression with the parameters: -I 60000, -b, and -G. Cufflinks was run twice, once with quantification against the reference with the respective genome general feature format and another with the RTA GFF (Datasets 1–6). For B73v4, 28 genes with problems in their GFF format were removed. For additional quality control, a Pearson's pairwise complete observations correlation (R^2) matrix was created with pairwise comparisons across all inbreds from expression levels.

Single Nucleotide Polymorphisms

Single nucleotide polymorphisms were called on both the reference genome and its derived RTAs with SAMTools (version 0.1.18; Li et al., 2009). A mpileup file was created with SAMTools mpileup with BAQ computation disabled, indel calling disabled, and a mapQ filter of 50.

First, an allele call was filtered on the basis of the base quality; if the quality was ≥20, the call was retained. The genotype of an individual at a position was called if there were at least five reads covering the position, the frequency of the allele at the position was >5% in the individual, and the allele was called at least twice. If there was more than one allele that passed this filter (making the position heterozygous), the genotype for that individual was called missing data. On a population basis, a position was retained if there was at least one inbred that had a call different from the reference and <80% missing data at a reference position (genotype calls in at least 192 inbreds) and <95% missing data at an RTA position (genotype calls in at least 48 inbreds). Additional quality control was performed by evaluating the inbreds' relationships in an unweighted pair group method with an arithmetic mean tree and Roger's genetic distances (Rogers, 1972). The tree was created in PowerMarker version 3.25 (Liu and Muse, 2005) and viewed in FigTree version 1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/, accessed 9 Apr. 2019). Based on their location in the tree and known pedigree information, the inbred line Mo8W was removed from the SNP and fragments per kb of exon model per million mapped reads (FPKM) matrix files. An additional 16 inbred lines were also removed from the SNP and FPKM files, resulting in a final set of 942 inbreds, referred to as the WiDiv-942 (Mazaheri et al., 2019), which were analyzed in this study.

Association Mapping

The B73- and PH207-aligned SNP sets were filtered to contain only diallelic SNPs (899,784 and 846,088, respectively), which were imputed with fastPHASE version 1.4 (Scheet and Stephens, 2006). The default parameters were used, with the -H flag set to a negative number to suppress phasing. The RTA SNPs (74,741 in the B73-aligned SNP set and 112,999 in the PH207-aligned SNP set) and SNPs on unaligned scaffolds (5384 for B73 and 12,126 for PH207) were excluded from imputation because they had an unknown physical location but were subsequently combined with the imputed SNPs to form the final SNP sets for GWAS. PHJ89 SNPs were not used for GWAS because their scaffolds were not assembled into chromosomes, complicating imputation.

Phenotypic data for GWAS were best linear unbiased predictors of the area under the disease progress curve (AUDPC) for SCMV in 578 individuals from the WiDiv-942 (Gustafson et al., 2018). Genome-wide association studies were performed using the gwas() function from the rrBLUP package (Endelman, 2011) in R (R Core Team, 2016), with zero principal components, a minor allele frequency threshold of 1%, and a kinship matrix computed from 10,000 SNPs randomly selected from across the 10 chromosomes. Genome-wide association studies were performed using rrBLUP specifically because the gwas() function does not naively impute missing data. The RTAs have large quantities of missing data, some of which reflect true absence from the individuals without an allele call and, as such, should not be imputed.

Chromosomal genes and RTAs were assigned a binary score of zero or one, indicating whether the lower bound of the FPKM confidence interval computed by Cufflinks (Trapnell et al., 2010) was zero or greater than zero, respectively. This matrix was used for GWAS, with the binary expression scores acting as the explanatory variables in place of SNPs. Binary expression GWAS was performed with the same parameters as the SNP GWAS described above, with the exception of filtering for minor allele frequency. The kinship matrix was the same as the one used for SNP GWAS.

RESULTS & DISCUSSION

De Novo Genome Assembly and Annotation of PHJ89

The PHJ89 genome was assembled with 68.5× coverage of two paired-end libraries coupled with three mate-pair libraries (2, 6, and 12 kb), resulting in a total assembly of 2.3 Gb with an NG50 scaffold size (i.e., 50% of the estimated genome size is in a scaffold of this size or greater) of 18.7 kb and a maximum scaffold length of 416.5 kb (Supplemental File S5). We checked assembly quality by aligning the paired-end reads back to the genome; after filtering out reads with alignments with a mapping quality less than 30 yet retaining reads that multimapping to capture the high repetitive sequence content in maize, 83.5% of the genomic reads aligned to the PHJ89 genome assembly. Genome assessment with BUSCO (version 2.0; Simao et al., 2015) revealed that 96% of the Embryophyta orthologs were present in the assembly [complete, 93.8% (single-copy, 8.2%; duplicated, 5.6%), fragmented, 2.2%, missing, 4.0%, number of gene groups searched, 1440], indicating a high level of completeness in the genic regions in the assembly. We also assessed the representation of the genic space of the PHJ89 assembly with cleaned RNA-Seq reads; of the two RNA-Seq libraries, 91.6 and 94.0% aligned to the genome, with 86.0 and 90.8% of all read pairs aligning (Supplemental File S2). When we used a custom repeat library, 84.9% of the genome was masked as putative repetitive sequence (Supplemental File S6). As the majority of small contigs were fragments of repetitive sequences, small contigs less than 1 kb were filtered from the assembly prior to annotation resulting in a filtered assembly of 1.6 Gb, spanning 151,993 scaffolds with a N50 scaffold size (50% of the assembled genome is in a scaffold of this size or greater) of 35.7 kb (Supplemental File S5). Structural annotation of the filtered assembly yielded a working set of 67,743 loci encoding 77,057 gene models (Supplemental File S7). The working set was filtered to remove low-confidence models, defined as those lacking transcript or protein alignment evidence, resulting in a high-confidence gene model set containing 29,670 loci encoding 38,560 gene models (Supplemental File S7). As only whole-seedling RNA-Seq data were available for PHJ89 annotation, the PHJ89 high-confidence gene set was smaller than those

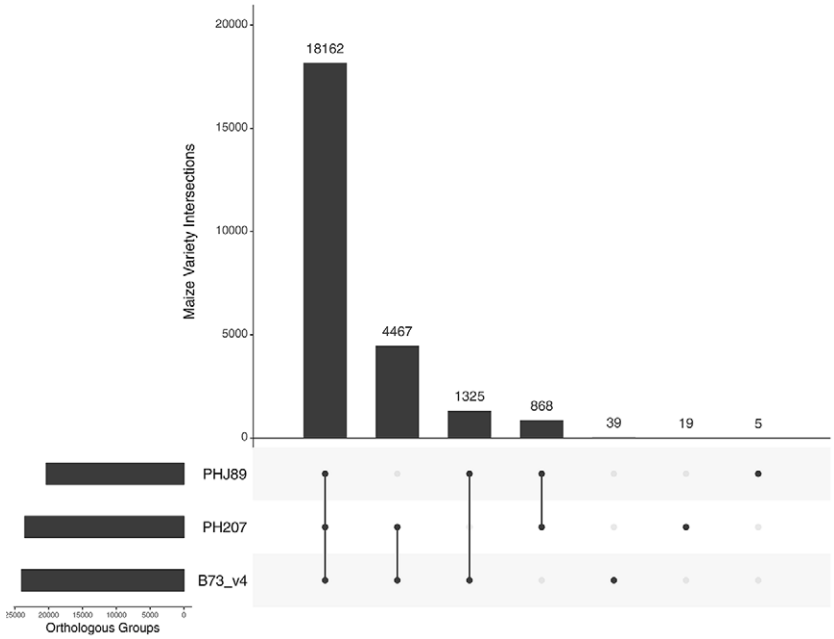

Fig. 1. Clustering of the predicted proteomes of the three major maize germplasm groups (B73, PH207, and PHJ89) with OrthoFinder (Emms and Kelly, 2015). The numbers at the top of each bar represent the number of clusters; the number of genes can be found in Supplemental File S8. The results were plotted with UpsetR (Conway et al., 2017).

of B73v4 or PH207 because of the limited survey of expression across the genome. For the high-confidence gene models, gene length and number of exons per model were substantially higher than in the working set; as a consequence, 85.8% of the high-confidence gene models were assigned a putative function.

The PHJ89 genome is the first de novo genome sequence of an inbred line from the Oh43 germplasm group. The Oh43 group is one of the major germplasm pools represented in modern commercial maize breeding programs (Mikel, 2011). Other groups, the Stiff Stalks, Lancasters, and Iodents, are represented by the previously sequenced lines B73 (Schnable et al., 2009; Jiao et al., 2017), Mo17 (Sun et al., 2018), and PH207 (Hirsch et al., 2016), respectively. By sequencing PHJ89, we were able to produce a draft reference sequence that is representative of a germplasm pool that is highly heterotic in crosses to both Iodent and Stiff Stalk type lines. Like PH207, PHJ89 is a commercial inbred line released from Plant Variety Protection and, as such, it represents more recently developed and highly selected material than older, foundational breeding lines and research lines like W22. To understand the differences in protein coding potential among these three germplasm groups, we performed clustering of the predicted proteomes of B73, PH207, and PHJ89 with OrthoFinder (Emms and Kelly, 2015), which clusters orthologous and paralogous groups, and by using a reciprocal best hits approach. Orthofinder yielded 18,162 paralogous groups containing 70,273 proteins from all three inbred proteomes (Fig. 1; Supplemental File S8). In addition to paralogous groups unique to two inbreds (B73 + PH207, B73 + PHJ89, or PH207 + PHJ89), each inbred had inbred-specific paralogous groups and singletons. In total, 7920, 9955, and 4929 proteins were unique to B73, PH207, and PHJ89, respectively, highlighting the diversity among these three germplasm groups. The small size of the PHJ89 scaffolds makes syntenic alignments among the three genomes challenging; therefore, we used reciprocal best hits to define the true homologs and examined the distribution of these putative homologs in the B73 genome (Fig. 2). Interestingly, only 44.4% (17,553) of the B73v4 representative models had a reciprocal best hit with both PH207 and PHJ89, whereas 16.1% (6375) and 8.4% (3331) of the B73v4 representative models had a reciprocal best hit with PH207 and PHJ89, respectively. These reciprocal best hits were dispersed throughout the genome; however, blocks of reciprocal best hits unique to one inbred were apparent (Fig. 2)

Generation of a Pantranscriptome

Previous efforts to generate a pantranscriptome in maize used a single reference genome, namely B73 (Hirsch et al., 2014). With access to multiple reference genomes that represented three different heterotic groups, we assessed how the choice of a reference genome affected the discovery of novel or dispensable transcripts. By using the three reference genomes, B73 (representative of the Stiff Stalk heterotic group), PH207 (representative of the Iodent heterotic group), and PHJ89 (representative of the Oh43 type heterotic group), we cataloged the maize pantranscriptome. In total, we identified 34,447, 39,672, and 37,436 RTAs that were not present in the B73, PH207, and PHJ89 reference genomes, respectively (Supplemental File S4). On the basis of the alignment of the RTAs to six publicly available reference genomes [the three discussed here (B73, PH207, and PHJ89) plus three additional publicly available genome

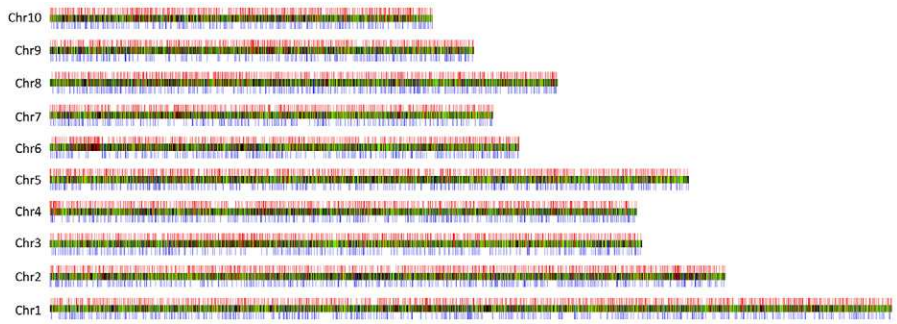

Fig. 2. The maize inpred line B73's reciprocal best hits with PHJ89 and PH207. Each bar shows the B73 genes in order along each chromosome. Each blue tick represents a gene that only has a PHJ89 reciprocal best hit; likewise, each red tick represents a gene with only a PH207 reciprocal best hit. Green indicates the B73 gene at that location has a reciprocal best hit with PH207 and PHJ879. Black indicates the B73 gene does not have a reciprocal best hit with either PH207 or PHJ89.

assemblies: the European inbreds EP1 and F7 (Unterseer et al., 2017) and the tropical inbred CML247 (Lu et al., 2015)], the majority of these RTAs (71–74%) aligned to at least one of the other maize genomes (Fig. 3), confirming that the RTAs represented divergent alleles of annotated maize genes and were not artifacts of our computational pipeline. However, an ample number of RTAs remained unaligned to even a single reference genome, suggesting that even with six reference genomes, we have yet to capture all genes within the maize pangenome.

Examination of the distribution of the RTAs across the 942 inbreds revealed that the majority were restricted in expression to less than 100 inbreds (Fig. 4), suggesting that a large number of dispensable genes were present in a limited number of inbreds (a line was considered to express a gene if the lower bound of the FPKM confidence interval was greater than 0). The RTAs were constructed from unaligned RNA-Seq reads from all inbreds that did not align to the reference genome and, as a consequence, represented a hybrid assembly (Hansey et al., 2012). In addition, the RTAs were filtered to remove alleles or close paralogs with ≥85% identity and 85% coverage with the cognate reference genome. To assess presence or absence for annotated genes and RTAs in the 942 inbreds, we used expression abundance estimations based on confidence levels as output from Cufflinks. Thus, as reads were permitted to align to both the reference genome and the RTAs, it is possible that the genes and RTAs with low levels of expression and/or with paralogs or more distant alleles in the reference genome or RTAs will be binned as not present.

Access to a pantranscriptome derived from this larger diversity panel allowed the examination of additional features of the RTAs. One enigma of grass genomes, including maize, is the bimodal distribution of genic GC content (Carels and Bernardi, 2000). Distributions of GC content in both the annotated B73 protein-coding genes and the B73-derived RTAs revealed that the RTAs had a unimodal GC content distribution, with a peak slightly lower than that of the low-GC annotated B73 genes (Fig. 5A). This suggests that dispensable genes are not a large component of high-GC genes in maize.

On average, RTA transcript length increases as the frequency of the RTA within the WiDiv-942 increases (Fig. 5B). This is expected, given the two factors that positively impact representation of a transcript in our RTA dataset: first, how widely distributed a dispensable gene is; second, how highly expressed it is, which will impact the probability of capturing the near- or full-length transcript. The mean expression level of annotated genes and RTAs increases with the number of inbreds in which they were identified. As shown in Fig. 5C and 5D, although the mean expression was lower for RTAs relative to annotated B73 genes, expression was correlated with the number of inbreds in which a gene or RTA was present. Similar patterns were observed for both the mean expression level in B73 and in all inbred lines. Thus, on average, a gene or RTA that is expressed by a larger number of inbreds will have higher expression than a gene or RTA restricted to a smaller subset of inbreds.

Access to the transcriptome and predicted proteome of nearly 1000 inbreds provides an unprecedented opportunity to define the extent of the maize pangenome. We clustered the predicted proteomes from the three annotated maize reference genomes with the predicted proteomes from all three sets of RTAs with the program Orthofinder (Emms and Kelly, 2015) that groups genes into orthologous and paralogous groups. Clustering of the predicted proteins from either the annotated genomes or the RTA datasets revealed a core set of 76,246 proteins within 6532 paralogous groups that contained at least one protein from B73, PH207, PHJ89, and an RTA; 92 proteins unique to B73; 39 unique to PH207; 13 unique to PHJ89; and 26,861 proteins within 9736 paralogous groups that were not present in any of the three reference genomes (Fig. 6).

Alignment of the annotated proteins of the maize reference genomes and RTA datasets to six different species' protein-annotated genomes in Phytozome version 12 *(A. thaliana*, *B. distachyon*, *O. sativa*, *S. viridis*, *S. bicolor*, and *V. vinifera*), and to the Pfam database was performed to determine the extent of known gene function in the annotated genes and dispensable gene sets (Supplemental File S9). With an E-value threshold of 1

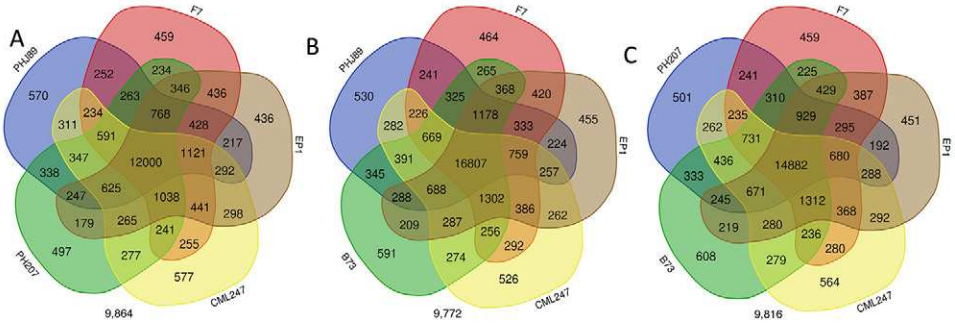

Fig. 3. Venn diagram displaying the representation of (A) B73 representative transcript assemblies (RTAs),(B) PH207 RTAs, and (C) PHJ89 RTAs at a coverage of ≥55% and identity of ≥85% in other maize inbred genomes (B73, CML247, F7, EP1, PH207, and PHJ89). Out of the RTAs, 71% (24,583) of B73, 75% (29,900) of PH207, and 74% (27,620) of PHJ89 RTAs were represented in at least one of the five genomes. Indicated below each Venn diagram is the number of RTAs that were not represented in one of the five genomes.

$\times 10^{-6}$, 94.3, 91.7, and 94.4% of the annotated proteins in the B73, PH207, and PHJ89 reference genomes, respectively, aligned to a protein sequence or a Pfam domain. In contrast, for the three RTA datasets, the proportion of proteins with an alignment to these core proteomes or Pfam ranged between 38.5 and 40.5%, suggesting that dispensable genes represent novel protein sequences. It is possible that a portion of the RTAs may be artifacts. However, it was shown in a comparison of the B73 and PH207 genome (Hirsch et al., 2016) that a substantial number of PAVs were caused by partial deletions of the gene in the reciprocal genome and, as a consequence, some of our RTAs represented partial deletions of genes that were not present in the cognate reference genomes. Furthermore, it has been shown in a direct comparison of the B73 vs. the PH207 genome (Hirsch et al., 2016), as well as a comparison of the B73 and Mo17 genome (Sun et al., 2018), that some genes are unique to an accession. In addition, sequence similarity is dependent on length; the predicted peptides from the RTAs are substantially shorter on average than the peptides predicted from genes within the genome. The B73-, PH207-, and PHJ89-derived RTAs were, on average, 199, 160, and 181 amino acids shorter than their respective genome-derived peptides. The impact of sequence length on the ability to detect sequence similarity is further demonstrated by the observation that the B73-, PH207-, and PHJ89-derived RTAs that aligned to a related sequence were, on average, 79, 78, and 84 amino acids longer than the RTAs that did not align to a related sequence. With 942 inbreds included in this study, we are likely to have captured a significant amount of the diversity in maize and thus the RTAs reported here provide a robust representation of dispensable genes in the maize pantranscriptome.

The effect of Using SNPs or Expression Data on GWAS Results

In addition to generating expression profiles, we also generated SNP datasets separately from each of the three reference genomes plus their respective RTAs. As an example of how using expression profiles can aid in identifying genes associated with traits of interest, GWAS of AUDPC for SCMV were performed with either SNPs or a binary coding of presence or absence of expression (ePAV) as the explanatory variables. We chose SCMV resistance as an appropriate GWAS example because a major resistance gene, *ZmTrxh*, has been identified (Liu et al., 2017) and is located at 24.03 Mb on chromosome 6 of the B73v4 reference genome. The gene is within a previously identified PAV (Springer et al., 2009; Gustafson et al., 2018). Because a causative resistance gene is known and is located within a PAV, SCMV resistance is an ideal candidate for comparing SNP and ePAV GWAS methods for identifying phenotype–genotype associations that lie within PAV regions.

The results from B73-aligned SNP GWAS reveal the strongest association (–log10(p) = 21.7) between AUDPC

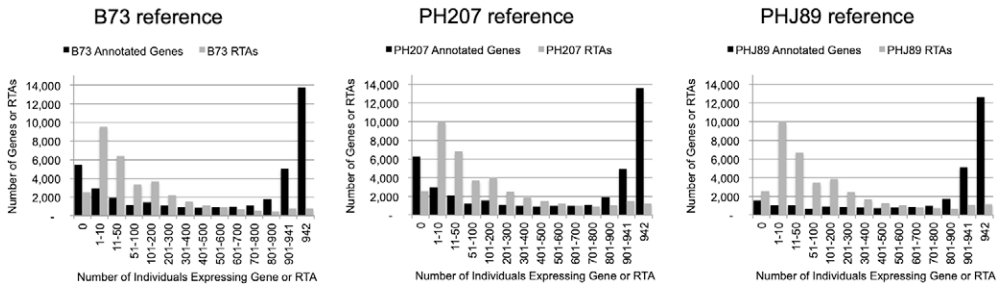

Fig. 4. Distributions for each reference maize genome of the number of individual inbreds expressing reference annotated protein-coding genes (black) or representative transcript assemblies (RTAs) (gray). RTAs are generally expressed by a restricted number of individuals, whereas a large proportion of reference-annotated protein-coding genes are expressed by all or most individuals. Individuals were considered to express a gene or RTA if the lower bound of the fragments per kb of exon model per million mapped reads (FPKM) confidence interval was greater than zero.

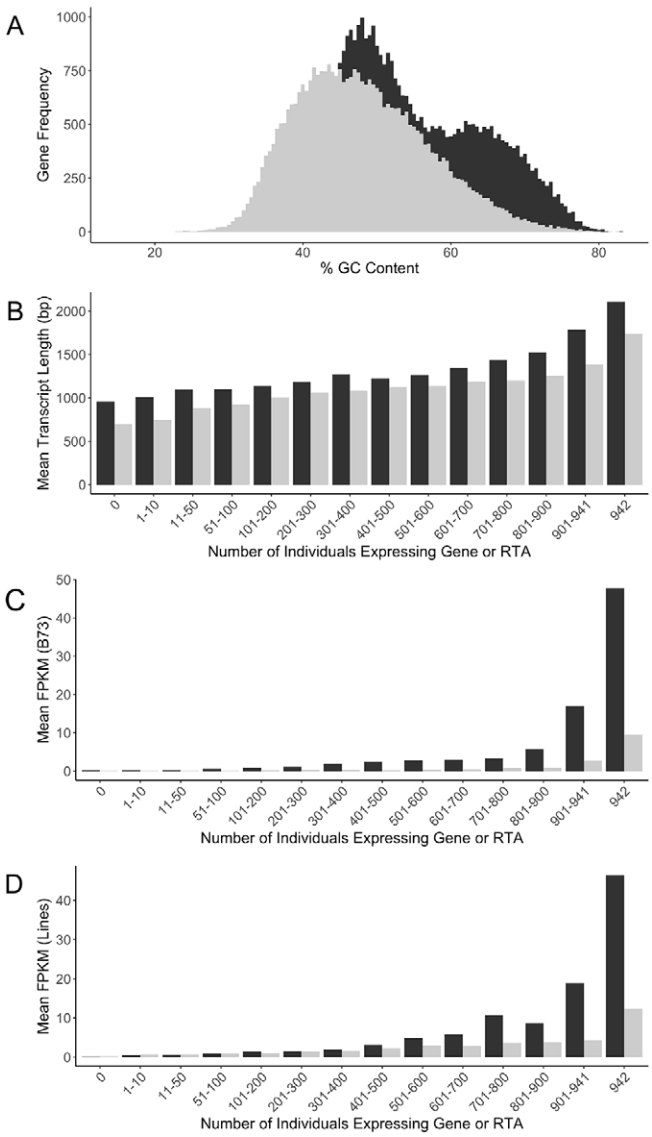

Fig. 5. Metrics of the representative transcript assemblies (RTAs) derived from analysis of 942 maize inbreds relative to the B73 genome were examined. All analyses performed with B73-reference expression values. B73 annotated genes, black; RTAs, gray. (A) Percentage of GC content of RTAs versus annotated genes in B73. (B) Transcript length distribution relative to the frequency at which an annotated gene or RTA was present within the expanded version of the Wisconsin diversity panel (WiDiv-942). (C) Mean fragments per kb exon model (or transcripts) per million mapped reads (FPKM) of annotated genes in B73 and RTAs based on their frequency within the WiDiv-942. (D) Mean FPKM in all inbred lines of annotated genes and RTAs based on their frequency within the WiDiv-942.

and SNP rs6_15535118. This is the most significant SNP on a peak that spans from 15.5 to 16.2Mb on chromosome 6, approximately 8.5 Mb away from *ZmTrxh* (Fig. 7). Efforts to identify genes conferring resistance to SCMV by means of genetic mapping have been reported since the early 2000s (Dußle et al., 2000) and a number of publications since then culminated in the identification of *ZmTrxh* in 2017 (Zhang et al., 2003; Uzarowska et al., 2009; Tao et al., 2013; Liu et al., 2017; Gustafson et al., 2018). The 17-yr process of identifying ZmTrxh was complicated by its location

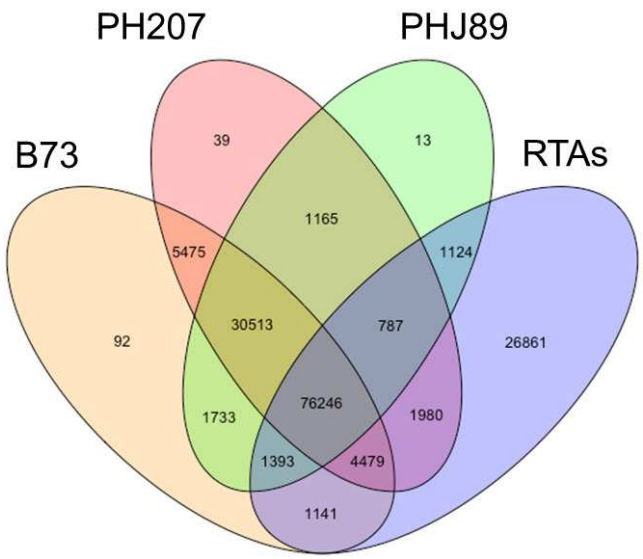

Fig. 6. Venn diagram of paralogous groups identified in the predicted proteomes of the three maize genome assembly annotation datasets (B73, PH207, and PHJ89) and the three sets of representative transcript assemblies generated in this study by OrthoFinder (Emms and Kelly, 2015).

within a PAV (Springer et al., 2009; Tao et al., 2013). Traditional SNP-based GWAS is unreliable for identifying associations within PAV regions because imputation can force allele calls onto individuals that have an absence allele and therefore do not carry SNPs in the region of interest. Additionally, causative SNPs that are not located in the coding sequence or untranslated regions of expressed genes will not be detected in GWAS of SNPs from RNA-Seq. Similar to SNPs within PAVs, SNPs within genes that are not expressed by a particular individual may also suffer from poor imputation accuracy. When ePAV is used as a proxy for PAV allele at a particular gene, GWAS should identify the genes for which PAV is associated with the phenotype of interest. Performing B73-aligned ePAV GWAS on the WiDiv-942 resulted in direct identification of *ZmTrxh* (*Zm00001d035390*) as the gene most significantly associated with AUDPC (–log10(p) = 46.7) (Fig. 7). Association between SCMV resistance and the presence–absence allele in the region containing *ZmTrxh* has previously been demonstrated (Gustafson et al., 2018); ePAV GWAS was able to replicate this result because of lack of expression in lines with the absence allele in the region containing *ZmTrxh*. This demonstrates the utility of expression data for identifying genes by GWAS that are located within PAVs, which are pervasive in maize and enriched for associations with plant architecture, flowering, and disease resistance (Springer et al., 2009; Chia et al., 2012; Hirsch et al., 2014; Lu et al., 2015). This ePAV approach assumes that expression is a reliable proxy for PAV. The presence of a gene does not necessarily imply its expression, so both PAV and variability for whether the gene is expressed or not will affect the ultimate association between that gene and the trait of interest. A similar approach that used expression values as the explanatory variable in GWAS has

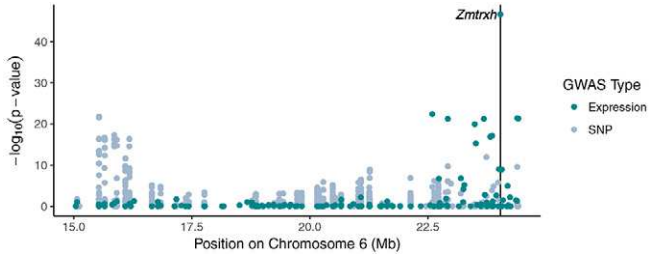

Fig. 7. Results from a genome-wide association study (GWAS) of *Sugarcane mosaic virus* (SCMV) resistance between 15 and 25 Mb on chromosome 6 of the maize inbred line B73. The most significant single nucleotide polymorphism (SNP) association is approximately 8 Mb away from a gene known to confer SCMV resistance, *ZmTrxh*, which is located within a presence–absence variant (PAV). The GWAS using the presence or absence of gene expression as the explanatory value, rather than SNP alleles, correctly identified *ZmTrxh* as the most significantly associated gene. Both GWAS were performed using SNP and fragments per kb of exon model per million mapped reads (FPKM) data from the B73 reference.

been used previously in maize to identify a gene associated with vegetative phase change that was not otherwise identified by SNP GWAS (Hirsch et al., 2014). In that case, the use of expression values for GWAS was not motivated by suspected PAV in genomic regions of interest but rather by the hypothesis that transcript abundance could explain phenotypic variation for vegetative phase change (Hirsch et al., 2014). In this study, we demonstrate that GWAS with expression values can also be useful for identifying associations with PAV regions.

Impact of Using Different Reference Genomes on GWAS Results

In addition to differences between SNPs and expression values called on a single reference genome, differences in SNP or expression data called against different reference genomes can also impact GWAS results. If *ZmTrxh* had not been present in B73, efforts to fine-map it would have been challenging. As an example of how reference choice can impact ePAV results, we performed a comparison of B73- and PH207-aligned ePAV GWAS results for SCMV resistance to demonstrate the variability of mapping results between or among reference genomes. As stated above, B73-aligned ePAV for AUDPC results in direct identification of *ZmTrxh* on chromosome 6 as the most significantly associated gene. The most significant PH207-aligned ePAV GWAS result was an RTA, locus_DN90808_c0_g1_i3. When the sequence of locus_DN90808_c0_g1_i3 was mapped to the B73 reference with GMAP, it aligned with 99.6% identity and 89.2% coverage to the region between 24,034,204 and 24,035,381 bp on chromosome 6, a nearly perfect overlap with *ZmTrxh*. This result indicates that PH207-aligned ePAV GWAS identified the same gene as B73-aligned ePAV GWAS, but the PH207-aligned results could not provide any indication of the physical location of the identified gene unless it cross-referenced the B73 genome. The PH207 reference contains a gap in the region where *ZmTrxh* was expected to be, but alignment of the raw PH207 reads (Hirsch et al., 2016) against the *ZmTrxh* coding sequence indicated that *ZmTrxh* was present in PH207 but was not included in the final assembly.

CONCLUSIONS

This study presented the *de novo* genome sequence of the first Oh43-type inbred line, PHJ89, which complements the Stiff Stalk line B73, the Lancaster line Mo17, and the Iodent line PH207 as representatives of several major germplasm groups in North American temperate maize. Our focus in sequencing the PHJ89 genome was to identify the gene space of an Oh43-type inbred, for which short-read Illumina sequences and associated assembly software performed well, as shown by the high BUSCO score (93.8% complete; 2.2% fragmented). Future efforts to generate chromosome-scale assemblies will require additional sequencing datasets and approaches such as Hi-C (Belton et al. 2012). In this study, we used the B73, PH207, and PHJ89 reference genome sequences to create parallel pantranscriptomes and SNP datasets for 942 diverse inbred lines. The expression profiles revealed the more dispensable nature of RTAs relative to annotated genes, regardless of reference genome. Finally, using SCMV resistance as an example, we demonstrated the utility of both expression profiles and genomic datasets created with different reference genomes for discovering GWAS associations that exist on PAVs or in unassembled regions of the genome. These results reinforce the dynamic and complex nature of the maize genome and provide resources for further exploring genetic diversity in maize.

Data Availability

Raw sequence data can be found in the NCBI Sequence Read Archive under BioProject PRJNA189400, PRJNA437324, and PRJNA448931 (see Supplemental Files S1, Supplemental File S2, and Supplemental File S3). Large data files have been deposited at the Dryad Digital Repository (doi:10.5061/dryad.dk22g4h). The repository includes: (i) B73-derived RTAs (Fasta file), (ii) PH207-derived RTAs (Fasta file), (iii) PHJ89-derived RTAs (Fasta file), (iv) B73-derived SNPs, (v) PH207-derived SNPs, (vi) PHJ89-derived SNPs, (vii) the PHJ89 genome assembly, (viii) the PHJ89 annotated gene set, and (ix) FPKM matrices for cognate genomes and the cognate genomes plus RTAs.

Supplemental Information

Supplemental File S1: Metrics of the reads used in the PHJ89 genome assembly.

Supplemental File S2: RNA-sequencing libraries used in the annotation of PHJ89.

Supplemental File S3: RNA sequencing data and statistics for all 959 maize inbred lines.

Supplemental File S4: Novel transcript discovery from a diverse inbred panel with three reference genome sequences.

Supplemental File S5: Assembly metrics of the PHJ89 genome.

Supplemental File S6: Repetitive sequence statistics of the PHJ89 genome assembly.

Supplemental File S7: Annotation of gene models in the PHJ89 genome assembly.

Supplemental File S8: Summary of OrthoFinder paralogous groups between B73v4, PH207, and PHJ89.

Supplemental File S9: Number of reference genome (B73, PH207, and PHJ89) and RTA proteins that aligned to any protein in six species and/or encoded a Pfam domain.

Conflict of Interest Disclosure

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

JG is supported by the National Research Initiative for Agriculture and Food Research Initiative Competitive Grants Program grant No. # 2012-67013-19460 from the USDA National Institute of Food and Agriculture. This work was funded by the Department of Energy (DOE) Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494). The work conducted by the US DOE Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the US DOE under Contract No. DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

JLG, BV, JPH, NCM-C, and CRB performed the data analysis. TJG and WFT provided the SCMV data. KB, AL, and MAM provided the plant materials and performed the sequencing. CRB, SMK, and NdL conceived of and designed the study. All authors contributed to manuscript writing, editing, and approval.

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Meyers, and D.J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Baldauf, J.A., C. Marcon, A. Lithio, L. Vedder, L. Altrogge, H.P. Piepho, et al. 2018. Single-parent expression is a general mechanism driving extensive complementation of non-syntenic genes in maize hybrids. Curr. Biol. 28:431–437. doi:10.1016/j.cub.2017.12.027

Belton, .J.M., R.P. McCord, J.H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker. 2012. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods 58(3):268–276. doi:10.1016/j.ymeth.2012.05.001

Boeckmann, B., A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365–370. doi:10.1093/nar/gkg095

Brunner, S., K. Fengler, M. Morgante, S. Tingey, and A. Rafalski. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17:343–360. doi:10.1105/tpc.104.025627

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, et al. 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10:421. doi:10.1186/1471-2105-10-421

Campbell, M.S., M. Law, C. Holt, J.C. Stein, G.D. Moghe, D.E. Hufnagel, et al. 2014. MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164:513–524. doi:10.1104/pp.113.230144

Carels, N., and G. Bernardi. 2000. Two classes of genes in plants. Genetics 154:1819–1825.

Chen, N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Prot. Bioinf. 4.10.1–4.10.14. doi:10.1002/0471250953.bi0410s05

Chia, J.M., C. Song, P.J. Bradbury, D. Costich, N. de Leon, J. Doebley, et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44:803–807. doi:10.1038/ng.2313

Conway, J.R., A. Lex, and N. Gehlenborg. 2017. UpSetR: An R package for the visualization of intersecting sets and their properties. Bioinformatics 33: 2938–2940. doi:10.1093/bioinformatics/btx364.

Donati, C., N.L. Hiller, H. Tettelin, A. Muzzi, N.J. Croucher, S.V. Angiuoli, et al. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. 11:R107. doi:10.1186/gb-2010-11-10-r107

Dußle, C.M., A.E. Melchinger, L. Kuntze, A. Stork, and T. Lübberstedt. 2000. Molecular mapping and gene action of *Scm1* and *Scm2*, two major QTL contributing to SCMV resistance in maize. Plant Breed. 119:299–303. doi:10.1046/j.1439-0523.2000.00509.x

Duvick, D.N., J.S.C. Smith, and M. Cooper. 2004. Long-term selection in a commercial hybrid maize breeding program. Plant Breed. Rev. 24:109–151.

Eddy, S.R. 1998. Profile hidden Markov models. Bioinformatics 14:755–763. doi:10.1093/bioinformatics/14.9.755

Emms, D.M., and S. Kelly. 2015. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157. doi:10.1186/s13059-015-0721-2

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255. doi:10.3835/plantgenome2011.08.0024

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. doi:10.1093/bioinformatics/bts565

Gnerre, S., I. Maccallum, D. Przybylski, F.J. Ribeiro, J.N. Burton, B.J. Walker et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U. S. A. 108: 1513–1518. doi:10.1073/pnas.1017351108.

Gordon, S.P., B. Contreras-Moreira, D.P. Woods, D.L. Des Marais, D. Burgess, S. Shu, et al. 2017. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat. Commun. 8:2184. doi:10.1038/s41467-017-02292-8

Gore, M.A., J.-M. Chia, R.J. Elshire, Q. Sun, E.S. Ersoz, B.L. Hurwitz, et al. 2009. A first-generation haplotype map of maize. Science 326:1115–1117. doi:10.1126/science.1177837

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652. doi:10.1038/nbt.1883

Gustafson, T.J., N. de Leon, S.M. Kaeppler, and W.F. Tracy. 2018. Genetic analysis of *Sugarcane mosaic virus* resistance in the Wisconsin Diversity Panel of maize. Crop Sci. 58:1853–1865. doi:10.2135/cropsci2017.11.0675

Hansey, C.N., B. Vaillancourt, R.S. Sekhon, N. De Leon, S.M. Kaeppler, and C.R. Buell. 2012. Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. PLoS One (3):e33071. doi:10.1371/journal.pone.0033071

Haas, B.J., A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith, Jr., L.I. Hannick, et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31:5654–5666. doi:10.1093/nar/gkg770

Hirsch, C.N., J.M. Foerster, J.M. Johnson, R.S. Sekhon, G. Muttoni, B. Vaillancourt, et al. 2014. Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26:121–135. doi:10.1105/tpc.113.119982

Hirsch, C.N., C.D. Hirsch, A.B. Brohammer, M.J. Bowman, I. Soifer, O. Barad, et al. 2016. Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in Maize. Plant Cell 28:2700–2714. doi:10.1105/tpc.16.00353

Jiao, Y., P. Peluso, J. Shi, T. Liang, M.C. Stitzer, B. Wang, et al. 2017. Improved maize reference genome with single-molecule technologies. Nature 546:524–527. doi:10.1038/nature22971

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S.L. Salzberg. 2013. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36. doi:10.1186/gb-2013-14-4-r36

Lam, H.M., X. Xu, X. Liu, W. Chen, G. Yang, F.L. Wong, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42:1053–1059. doi:10.1038/ng.715

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359. doi:10.1038/nmeth.1923

Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25. doi:10.1186/gb-2009-10-3-r25

Leggett, R.M., B.J. Clavijo, L. Clissold, M.D. Clark, and M. Caccamo. 2014. NextClip: An analysis and read preparation tool for Nextera long mate pair libraries. Bioinformatics 30:566–568. doi:10.1093/bioinformatics/btt702

Li, W. and A. Godzik. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. doi:10.1093/bioinformatics/btl158

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. https://arxiv.org/abs/1303.3997 (accessed 16 Apr. 2019).

Lin, K., N. Zhang, E.I. Severing, H. Nijveen, F. Cheng, R.G.F. Visser, et al. 2014. Beyond genomic variation- comparison and functional annotation of three *Brassica rapa* genomes: A turnip, a rapid cycling and a Chinese cabbage. BMC Genomics 15:250. doi:10.1186/1471-2164-15-250

Liu, F., Y. Zhu, Y. Yi, N. Lu, B. Zhu, and Y. Hu. 2014. Comparative genomic analysis of *Acinetobacter baumannii* clinical isolates reveals extensive genomic variation and diverse antibiotic resistance determinants. BMC Genomics 15:1163. doi:10.1186/1471-2164-15-1163

Liu, K., and S.V. Muse. 2005. PowerMarker: An integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129. doi:10.1093/bioinformatics/bti282

Liu, Q., H. Liu, Y. Gong, Y. Tao, L. Jiang, W. Zuo, et al. 2017. An atypical thioredoxin imparts early resistance to *Sugarcane mosaic virus* in maize. Mol. Plant 10:483–497. doi:10.1016/j.molp.2017.02.002

Lu, F., M.C. Romay, J.C. Glaubitz, P.J. Bradbury, R.J. Elshire, T. Wang, et al. 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. Nat. Commun. 6. 6914. doi:10.1038/ncomms7914

Magoc, T., and S.L. Salzberg. 2011. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963. doi:10.1093/bioinformatics/btr507

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17(1): 10–12. doi:10.14806/ej.17.1.200

Mazaheri, M., M. Heckwolf, B. Vaillancourt, J. Gage, B. Burdo, S. Heckwolf, et al. 2019. Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol. 19:45. doi:10.1186/s12870-019-1653-x

Mikel, M.A. 2011. Genetic composition of contemporary U.S. commercial dent corn germplasm. Crop Sci. 51:592–599. doi:10.2135/cropsci2010.06.0332

Montenegro, J.D., A.A. Golicz, P.E. Bayer, B. Hurgobin, H.T. Lee, C.K.K. Chan, et al. 2017. The pangenome of hexaploid bread wheat. Plant J. 90:1007–1013. doi:10.1111/tpj.13515

Murray, M.G., and W.F. Thompson. 1980. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 10:4321–4325. doi:10.1093/nar/8.19.4321

Punta, M., P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40:D290–D301. doi:10.1093/nar/gkr1065

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Read, B.A., J. Kegel, M.J. Klute, A. Kuo, S.C. Lefebvre, F. Maumus, et al. 2013. Pan genome of the phytoplankton *Emiliania underpins* its global distribution. Nature 499:209–213. doi:10.1038/nature12221

Reif, J.C., A.R. Hailauer, and A.E. Melchinger. 2005. Heterosis and heterotic patterns in maize. Maydica 50:215–223.

Rogers, J.S. 1972. Measures of genetic similarity and genetic distance. Studies Genet. VII:145–153.

Saghai-Maroof, M.A., K.M. Soliman, R.A. Jorgensen, and R.W. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA 81:8014–8018. doi:10.1073/pnas.81.24.8014

Schatz, M.C., L.G. Maron, J.C. Stein, A.H. Wences, J. Gurtowski, E. Biggers, et al. 2014. Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. Genome Biol. 15:506. doi:10.1186/s13059-014-0506-z

Scheet, P., and M. Stephens. 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. doi:10.1086/502802

Schnable, P., D. Ware, R. Fulton, and J. Stein. 2009. The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112–1115. doi:10.1126/science.1178534

Simao, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, and E.M. Zdobnov. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. doi:10.1093/bioinformatics/btv351

Simpson, J.T., K. Wong, S.D. Jackman, J.E. Schein, S.J. Jones, and I. Birol. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res. 19:1117–1123. doi:10.1101/gr.089532.108

Springer, N.M., S.N. Anderson, C.M. Andorf, K.R. Ahern, F. Bai, O. Barad, et al. 2018. The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat. Genet. 50:1282–1288. doi:10.1038/s41588-018-0158-0

Springer, N.M., K. Ying, Y. Fu, T. Ji, C.T. Yeh, Y. Jia, et al. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 5. doi:10.1371/journal.pgen.1000734

Stanke, M., O. Schoffmann, B. Morgenstern, and S. Waack. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62. doi:10.1186/1471-2105-7-62

Sun, S., Y. Zhou, J. Chen, J. Shi, H. Zhao, H. Zhao, et al. 2018. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat. Genet. 50:1289–1295. doi:10.1038/s41588-018-0182-0

Tao, Y., L. Jiang, Q. Liu, Y. Zhang, R. Zhang, C.R. Ingvardsen, et al. 2013. Combined linkage and association mapping reveals candidates for *Scmv1*, a major locus involved in resistance to sugarcane mosaic virus (SCMV) in maize. BMC Plant Biol. 13:162. doi:10.1186/1471-2229-13-162

Tettelin, H., V. Masignani, M.J. Cieslewicz, C. Donati, D. Medini, N.L. Ward, 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. doi:10.1073/pnas.0506758102

Trapnell, C., B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28:511–515. doi:10.1038/nbt.1621

Unterseer, S., M.A. Seidel, E. Bauer, G. Haberer, F. Hochholdinger, N. Opitz, C. Marcon, K. Baruch, M. Spannagl, K.F.X. Mayer, and C.-C. Schön. 2017. European Flint reference sequences complement the maize pan-genome. bioRxiv. doi:10.1101/103747

Uzarowska, A., G. Dionisio, B. Sarholz, H.P. Piepho, M. Xu, C.R. Ingvardsen, et al. 2009. Validation of candidate genes putatively associated with resistance to SCMV and MDMV in maize (*Zea mays* L.) by expression profiling. BMC Plant Biol. 9:15. doi:10.1186/1471-2229-9-15

Walker, E.L., T.P. Robbins, T.E. Bureau, J. Kermicle, and S.L. Dellaporta. 1995. Transposon-mediated chromosomal rearrangements and gene duplications in the formation of the maize R-r complex. EMBO J. 14:2350–2363. doi:10.1002/j.1460-2075.1995.tb07230.x

Wu, T.D., and C.K. Watanabe. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875. doi:10.1093/bioinformatics/bti310

Zhang, S.H., X.H. Li, Z.H. Wang, M.L. George, D. Jeffers, F.G. Wang, et al. 2003. QTL mapping for resistance to SCMV in Chinese maize germplasm. Maydica 48:307–312.

Zhou, Y., C.A.D. Burnham, T. Hink, L. Chen, N. Shaikh, A. Wollam, et al. 2014. Phenotypic and genotypic analysis of *Clostridium difficile* isolates: A single-center study. J. Clin. Microbiol. 52:4260–4266. doi:10.1128/JCM.02115-14