Enhanced pan-genomic resources at the maize genetics and genomics database

Ethalinda K. Cannon 1, John L. Portwood II 1, Rita K. Hayford 1, Olivia C. Haley 1, Jack M. Gardiner 1, Carson M. Andorf 1, Margaret R. Woodhouse 1, 1

1USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA
2Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA
3Department of Computer Science, Iowa State University, Ames, IA 50011, USA

*Corresponding author: USDA-ARS, Corn Insects and Crop Genetics Research Unit, 819 Wallace Road, Ames, IA 50011, USA. Email: margaret.woodhouse@usda.gov

Pan-genomes, encompassing the entirety of genetic sequences found in a collection of genomes within a clade, are more useful than single reference genomes for studying species diversity. This is especially true for a species like Zea mays, which has a particularly diverse and complex genome. Presenting pan-genome data, analyses, and visualization is challenging, especially for a diverse species, but more so when pan-genomic data is linked to extensive gene model and gene data, including classical gene information, markers, insertions, expression and proteomic data, and protein structures as is the case at MaizeGDB. Here, we describe MaizeGDB’s expansion to include the genic subset of the Zea pan-genome in a pan-gene data center featuring the maize genomes hosted at MaizeGDB, and the outgroup teosinte Zea genomes from the Pan-Andropoganeae project. The new data center offers a variety of browsing and visualization tools, including sequence alignment visualization, gene trees and other tools, to explore pan-genes in Zea that were calculated by the pipeline Pandagma. Combined, these data will help maize researchers study the complexity and diversity of Zea, and to use the comparative functions to validate pan-gene relationships for a selected gene model.

Keywords: maize; pan-genome; database; genomes

Introduction

Research in the past decade has demonstrated that a single reference genome is not sufficient to represent species diversity (Yang et al. 2019). For instance, it was determined that about 10% of the human pan-genome is missing from the human reference genome (Sherman and Salzberg 2020), and two maize inbred lines might share only 79% of their genes (Morgante et al. 2005). Maize is a particularly diverse organism (Buckler et al. 2006); two maize lines are on average as diverged from one another as humans are from chimpanzees (Chen and Li 2001). Therefore, a significant amount of genetic information can be unaccounted for when relying on a single reference genome for studying species diversity. Because of this, researchers have come to rely on pan-genomes for their research within and across species.

A pan-genome comprises all the available genetic and genomic elements within a taxon. The first pan-genome was constructed in 2005 for six strains of the bacteria Streptococcus agalactiae using a whole-genome shotgun sequencing approach (Tettelin et al. 2005). Since then, ever-improving sequencing technologies such as long-read sequencing and optical map assembly methods have enabled the construction of complete, reference-quality assemblages for large and complex genomes such as maize. This has facilitated the quick and efficient sequencing of multiple genomes to generate pan-genomes in plants such as rice (Tettelin et al. 2005), tomato (Gao et al. 2019), and maize (Gao et al. 2019; Hufford et al. 2021). Pan-genomes contribute to basic research and breeding in multiple ways; for example, QTL, GWAS, genomic selection, and marker-assisted breeding are no longer limited by a single reference genome (Della Coletta et al. 2021). Zea mays ssp. mays (maize, corn) has been the world’s top production grain crop for over a decade (http://faostat.fao.org/), and is used extensively for livestock feed and biofuels (Lawrence and Walbot 2007; Carpita and McCann 2008; Penning et al. 2009; Bosch et al. 2011). Maize’s breeding history spans thousands of years, from traditional indigenous breeding practices (Romero Navarro et al. 2017; Aguirre-Liguori et al. 2019; Gonzalez-Segovia et al. 2019) to directed breeding, including the green revolution (Prasanna 2012; Andorf et al. 2019). For the last century, maize’s role as a historic biological model organism has led to discoveries that have advanced our understanding of biology across all organisms (Creighton and McClintock 1931; Beadle 1932; Rhoades 1984), particularly in plant biology, evolution, domestication, development, genetics, and genomics (Strahle and Scanlon 2009).

The Maize Genetics and Genomics Database (MaizeGDB—https://www.maizegdb.org) (Woodhouse et al. 2021, 2023) is the community database for maize researchers, providing data curatin and informatics resources to support maize genetics, genomics, and breeding research. MaizeGDB also serves as the maize research community’s support and outreach hub, facilitating collaboration and data sharing, and standardizing maize genetic and genomic nomenclature. MaizeGDB traces its lineage to the historical collection and sharing of maize research data, such as gene lists, composite genetic maps, and accessions for
use in research, ongoing since the early twentieth century (Rhoades 1984; Coe 2001). MaizeDB was established in 1991 to provide this growing volume of data via the Internet (Polacco et al. 2002; Polacco and Coe 2006). In 2003, MaizeDB became MaizeGDB when it expanded to include sequence data (Lawrence et al. 2004). MaizeGDB then shifted its primary focus to genomic data in 2008 with the release of the first complete B73 genome assembly (Lawrence et al. 2004; Schnable et al. 2009).

As of February 2024, MaizeGDB hosts 104 maize genome assemblies, including the representative reference genome for *Z. mays* ssp. mays, B73 (Schnable et al. 2009) and the important individual inbred lines PH207 (Hirsch et al. 2016), Mo17 (Yang et al. 2017; Sun et al. 2018), and W22 (Springer et al. 2018); a set of European flint lines (Haberer et al. 2020); sweet corn (Haberer et al. 2020; Hu et al. 2021); 12 founder inbred lines representing heterotic groups from both the US and China (Wang et al. 2023); and the set of 26 high-quality PacBio genome assemblies of the Nested Associated Mapping (NAM) population founder lines (Huffman et al. 2021), and nine genomes from the PanAndropogoneae Consortium (Stitzer et al. 2024) (under Toronto Agreement).

The NAM founder lines represent the broad diversity of domesticated maize (Yu et al. 2008), and the resulting 25 NAM populations have been key in elucidating maize flowering time (Buckler et al. 2009), leaf architecture (Buckler et al. 2009; Tian et al. 2011), disease resistance (Poland et al. 2011), and important agronomic traits (Wallace et al. 2014). The high-quality NAM founder assemblies, coupled with version 5 of the representative maize genome for B73, all assembled and annotated with the same methods, presented an ideal opportunity to carry out pan-maize genome analyses. To host the NAM founder genomes, their data, and results of pan-genome analyses, MaizeGDB was expanded and revised in 2021 to include pan-genomic data alongside information about individual gene models and in the genome browsers (Woodhouse et al. 2021). The maize community’s response to MaizeGDB’s pan-genomic approach was significant: a 3-fold increase in usage, particularly for the NAM genome browsers (Woodhouse et al. 2021, 2023).

*Zea mays* resides within the larger Andropogoneae clade, which includes the *Zea* outgroup *Tripsacum dactyloides*, an outgroup to *Tripsacum*, *Andropogon virginicus*, as well as many other grass species, including the grain crop Sorghum (Fig. 1). The genomes of many of these species within Andropogoneae were sequenced and assembled by the PanAnd Consortium (Stitzer et al. 2024) and released to MaizeGDB in 2023 (under Toronto Agreement). Quantitative estimates of evolutionary constraint across the Andropogoneae tribe are expected to improve predictions of fitness-related traits including yield in maize and Sorghum. Because convergent environmental adaptations across the Andropogoneae tribe could explain a substantial proportion of the variance in genotype by environment (GxE) (reference pending), their sequenced genomes will be a valuable resource to maize researchers.

Among the genome assemblies and annotations, the PanAnd Consortium has produced are eight high-quality accessions of teosintes, the wild relatives of domesticated maize. These include two *Zea diploperennis* genomes, *Zea nicaraguensis*, *Z. mays* subspecies *huehuetenangensis*, two *Z. mays* subspecies *parviglumis*, and two *Z. mays* subspecies *mexicana* genomes (under Toronto Agreement). These teosinte datasets will be particularly valuable for identifying domestication loci and for identifying useful genetic material to cross back into domesticated maize.

This paper describes the expansion of MaizeGDB’s pan-genome framework to include the PanAnd genomes described above; a new pan-gene interface that shows sequence alignment across pan-genes; phylogenetic trees of each pan-gene based on sequence similarity; and the integration of pan-gene macro and microsynteny visualization software.

**Materials and methods**

**Lifted gene models from the PanAnd genomes to other genomes on the gene model pages**

The gene model annotations of the teosinte genomes *Z. diploperennis* (*Zd-Gigi-REFERENCE-PanAnd-1.0, Zd-Momo-REFERENCE-PanAnd-1.0, Z. nicaraguensis* (*Zn-Pi165697-REFERENCE-PanAnd-1.0), *Z. mays* subspecies *huehuetenangensis* (*Zh-RIMHU001-REFERENCE-PanAnd-1.0, Z. mays* subspecies *parviglumis* (*Zv-TIL01-REFERENCE-PanAnd-1.0, Zv-TIL11-REFERENCE-PanAnd-1.0), and *Z. mays* subspecies *mexicana* (*Zx-TIL18-REFERENCE-PanAnd-1.0, *Zx-TIL25-REFERENCE-PanAnd-1.0), and the PanAnd teosinte outgroup *Andropogon virginicus* (*Av-Kellogg12878-REFERENCE-PanAnd-1.0) were reciprocally lifted to the genomes of version 5 of B73 and the NAM founder genomes (Hufford et al. 2021; Stitzer et al. 2024) as well as to each other using the liftoff software pipeline (Shumate and Salzberg 2020) as described in Shumate and Salzberg (2020) and Woodhouse et al. (2021). All genomes can be downloaded at MaizeGDB (the PanAnd genomes under Toronto Agreement).

**Creation of pan-genes for the pan-gene pages**

To present pan-gene data, summaries, and visualizations, a new data center, the pan-gene data center was created for MaizeGDB. This complements the existing gene data center, which shows information about individual gene models and associated loci, if any.

The Pandagama pan-gene analysis pipeline (https://github.com/legumeinfo/pandagama) was selected for construction of maize pan-genes. This pipeline takes advantage of commonly used software packages, along with additional filtering and analysis. The input is a set of annotations, including the GFF file describing the gene models, and CDS and protein FASTA files for each. The annotations are divided into “main” and “extra” sets. The main set establishes the pan-genes, then the gene models in the extra annotations are added to the pan-genes. For the *Zea* pan-genome, the main set of annotations included version 5 of the B73 representative genome and the NAM founders (Hufford et al. 2021), which represent the majority of domestic maize diversity, and the additional *Zea* species sequenced and annotated by the PanAnd project (Hufford et al. 2021, Stitzer et al. 2024). A full list of the annotations in both the “main” and “extra” sets is included in the configuration file Supplementary File 1, and pan-gene counts and feature averages can be found in Supplementary Table 1.

The Pandagama process consists of the following steps: attach positional information to gene model identifiers; search for pairwise homologies using MMSeq2 (Steinegger and Soding 2017; Hufford et al. 2021; Stitzer et al. 2024); filter by synteny with DAGchainer (Haas et al. 2004); cluster the results with mcl (Haas et al. 2004; Van Dongen 2008); add gene models not yet included by sequence similarity; align protein sequence for all gene models in each family with FAMSA (Deorowicz et al. 2016); build HMMs for each alignment; calculate gene trees with FastTree 2 (Price et al. 2010; Deorowicz et al. 2016); and finally, report statistics. Exemplar gene models are selected for each pan-gene. The use of exemplar gene models rather than calculating consensus sequence was selected because the consensus may be shorter
than the constituent genes in the pan-gene when portions of the alignment are represented by only a single gene or a few genes.

Results
The PanAnd genomes at MaizeGDB
Collections of genomes from specific projects, such as the NAM founders, the Chinese Academy of Agricultural Sciences (CAAS) founder inbred genomes (Wang et al. 2023), and the PanAnd genomes are collated and offered as Project Pages at MaizeGDB (Woodhouse et al. 2021; Wang et al. 2023). The B73, PanAnd, and NAM founder assemblies and annotations play a particularly important role in calculating pan-genes across Zea species. The most recent of these, the PanAnd genomes, are described here.

Information and data for the PanAnd genomes and annotations can be found by navigating to the home page and clicking on the PanAnd button (Fig. 1a). The PanAnd page (Fig. 2) is subdivided into three tabs: (A) The Project Details tab, which describes the project and lists the genomes within the project. Clicking on one of the genomes (e.g. Zv-TIL11-REFERENCE-PanAnd-1.0) from the Project Details tab takes the user to the Metadata tab for that genome (B), which describes the methods and statistics for the sequencing, assembly, and annotation of that genome. (C) is the Browser tab for each genome, which features a JBrowse instance of the genome and the gene model annotation at the top of the browser. In this example using Zv-TIL11-REFERENCE-PanAnd-1.0, the tracks for the lifted gene model annotations (see Materials and Methods) from four genomes are also selected, so that a user can compare gene model presence/absence across different genomes, and compare gene model structure. Clicking on any of the lifted gene models in these tracks will take the user to that gene model within that genome; this functionality is described in more detail in Woodhouse et al. (2021).

The MaizeGDB pan-gene data center
Pan-gene analysis methods are rapidly evolving, with several pipelines currently available. Consequently, conducting pan-gene analyses using different pipelines or even varying versions of the same pipeline can lead to inconsistent results. These discrepancies may arise due to factors such as the number of input genomes, the quality of gene model annotations, and other parameters. One goal of the MaizeGDB pan-gene data center is to allow researchers to determine how well a set of gene models within the MaizeGDB-calculated pan-gene dataset group together by providing multiple ways of examining the data.

While a pan-genome analysis considers the entirety of all sequences found across multiple individuals, a pan-gene analysis considers only the genic sequence (Sherman and Salzberg 2020). Early work on pan-genomes in bacteria (Tettelin et al. 2005) focused on the comparison of genes rather than the entire genome, which is suitable for gene-dense bacterial genomes; but in higher-order organisms, nongenic sequence typically makes up the majority of the genome. However, the calculation and presentation of full pan-genome data is a time- and resource-intensive endeavor, especially for a genome as large and complex as maize, so MaizeGDB’s initial foray into pan-genome data focuses on pan-gene analyses and representation.

Pan-genes are generally categorized as core, or present across all or nearly all genomes within a pan-gene analysis, and noncore (“dispensable” or “accessory”), when not present in all or most genomes (Tettelin et al. 2005; Sherman and Salzberg 2020). The exact definition of “most genomes” can vary, depending on the
research group and the quality of assemblies and annotations. Additionally, the terms “rare,” “orphan,” or “discrete” are frequently applied to describe pan-genes identified exclusively in a single genome within a pan-genome analysis. However, these labels can be easily confused with similar terminology used to describe genes that are unique to a single species in cross-phylogenetic comparisons.

The pan-gene data center can be accessed from the MaizeGDB home page (Fig. 1b). The data center home page (https://maizegdb.org/pan_gene_center/pan_gene) provides a variety of means for querying and downloading pan-gene data. The page is organized into six sections (Fig. 3), prefaced with the Introduction at the top of the page which briefly describes how the pan-genes were calculated (see Materials and Methods). The Simple Search section (Fig. 3a) allows researchers to search pan-genes by locus association, gene model ID, transcript ID, or protein ID. The Advanced Search (Fig. 3b) permits more detailed searches. Rather than establishing a definition for core, dispensable, and rare gene models for our pan-gene datasets, MaizeGDB has structured the Advanced Search so that researchers can set parameters to create their own definitions of these categories; find pan-genes present or absent in a subset of annotations, and other criteria. The parameters include: (1) the pan-gene analysis the user wishes to query (as of this writing, only a Zea pan-gene analysis is available; future versions will include pan-grass gene family analyses); (2) pan-genes with associated loci, proteins, or trait data; (3) a cutoff of at least X number of members, or fewer than X number of members; (4) a threshold of at least X% of annotations, or no more than X% of annotations; (5) a list of members from a menu of assemblies to include or exclude from the analysis. Researchers can limit the number of pan-genes returned by the search, with an upper limit of 2,000 records. An option to download the full search results is given. Note that the result table includes the pan-gene exemplar (the gene model selected as being the standard for a pan-gene, described in Materials and Methods), which may not be the gene model queried. Figure 3c indicates where a user can download pan-gene exemplar sequence for a set of gene models along with the full pan-gene data downloads.
Fig. 3. The pan-gene data center. a) Simple Search allows a user to search pan-genes by locus symbol, gene model ID, transcript ID, or protein ID. b) Advanced Search permits more detailed searches. Parameters include: 1) which pan-gene analysis the user wishes to query; 2) pan-genes with associated loci, proteins, or trait data (Zm00001eb067740 as an example); 3) have at least X number of members, or fewer than X number of members; 4) have at least X% of annotations, or no more than X% of annotations; 5) contain members from a menu of assemblies, or do not contain members from a menu of assemblies. A user can limit the number of pan-genes returned by the search, with an upper limit of 2,000 records. Note that the result table includes the pan-gene exemplar transcript for the pan-gene that contains Zm00001eb067740, which is Zm00026ab067840_T001 from the genome CML333. c) Downloads of all the pan-gene records from all analyses. d) Pan-gene distribution, i.e., the pan-gene size and the number of members per pan-gene size. e) A phylogenetic tree of all assemblies included in the pan-gene analysis. f) Definitions used to describe the pan-gene information at MaizeGDB.

Fig. 4. Pan-gene record page for Zm00001eb067740 https://maizegdb.org/pan_gene_center/pan_gene/Zm00001eb067740. The menu (center) has 11 drop-down tabs containing various types of data for Zm00001eb067740 and its pan-gene members. The pan-gene specific sections are highlighted. a) Analysis and statistics for the pan-gene analysis; b) all gene model members within the pan-gene; c) CDS and protein alignments of all members within the pan-gene performed by MSAviewer; CDS alignment is shown; d) a phylogenetic tree of all pan-gene members by IcyTree https://icytree.org/; e) mutation insertional data for the B73 gene model if present in the pan-gene; f) metabolic pathway information for the B73 gene model, if present; g) expression information for the B73 gene model, if present; h) protein information includes 3D structural visualization of the B73 gene model, if present; i) SNPs and traits for the B73v5 gene model, if present; j) function for the B73 gene model, if present; and k) CDS and protein sequence for the pan-gene exemplar Zm00026ab067840.
pan-gene data at MaizeGDB, describes the pan-gene distribution, i.e. the pan-gene size and the number of members per pan-gene size. Figure 3e shows the phylogenetic relationship among the genomes included in the pan-gene analysis calculated with OrthoFinder (Emms and Kelly 2019). Figure 3f includes the definitions used to describe the pan-gene information at MaizeGDB. Note that the pan-Zea pan-gene datasets are under Toronto Agreement at this time.

**The pan-gene record pages**

The pan-gene record page is divided into two or more tabs: one for the pan-gene record page, one for the Genomic Context Viewer (GCV) https://gcv.maizegdb.org/ (Cleary and Farmer 2018, 2023) which permits exploring macro and microsynteny among the genomes represented based on pan-genes, and a tab for each associated locus, if any (Fig. 4). For example, Fig. 4 shows the pan-gene record for gene model Zm00001eb067740.

**The pan-gene tab**

The pan-gene tab provides multiple ways of exploring the pan-gene and its members (Fig. 4). The information presented includes all data in MaizeGDB that is linked to gene models. The majority of these data are linked to versions 3–5 of the B73 genome assemblies and gene model annotations. Therefore, if the pan-gene has B73 as...
a member, all data linked to it is applied to all the members of a pan-gene.

The pan-gene tab is subdivided into 11 sections (Fig. 4 highlights the pan-gene-specific sections): a detailed analysis and statistics for the pan-gene analysis (Fig. 4a); all the gene model members within the pan-gene (Fig. 4b); CDS and protein alignments of all members within the pan-gene performed by MSAviewer (Yachdav et al. 2016), where the user can choose which color scheme to visualize the alignments (Fig. 4c); a phylogenetic tree of all pan-gene members drawn by IcyTree (Yachdav et al. 2016; Vaughan 2017); mutation insertional data for the B73 gene model if present in the pan-gene (Fig. 4e); metabolic pathway information for the B73 gene model, if present (Fig. 4f); expression information for the B73 gene model, if present (Fig. 4g); protein information includes 3D structural visualization of the B73 gene model, if present (Fig. 4h); SNPs and traits for the B73v5 gene model, if present (Fig. 4i); functions for the B73 gene model, if present (Fig. 4j); and CDS and protein sequence for the pan-gene exemplar. If a discrepancy is detected between the contents of the pan-gene and other evidence of gene model and locus relationships at MaizeGDB, an alert is shown at the top of the page (Fig. 4k).

**The Genome Context Viewer tab**

The GCV https://github.com/legumeinfo/gcv (Cleary and Farmer 2018, 2023) is a visualization tool that permits exploring macro- and microsynteny among the genomes represented in the pan-genomes (Fig. 5) at the level of pan-genes or gene families. The viewer is centered on the pan-gene exemplar, represented by the triangle outlined in bold (circled in Fig. 5a). Each triangle represents a gene model colored by its pan-gene within a window of collinearity for each genome (note: distance between gene models is indicated by the thickness of the line: thicker lines indicate greater distance). The direction of triangles indicates the forward or reverse strand. A more complete description of the use of GCV can be found at Cleary and Farmer (2023).
The locus tab(s)
If the pan-gene is associated with one or more loci, the locus tabs are labeled with the locus symbol (in this example, liguleless1) (Fig. 6). The locus information includes its name and synonyms, description, comments from references and personal communications, gene products, phenotypes, and other information. A link to the full MaizeGDB locus record page containing all information about that locus available in MaizeGDB, including genetic and historical data, is provided.

The pan-gene data center in combination with pan-genome tracks in MaizeGDB JBrowse instances can be used to explore the validity of pan-gene membership and specific gene model structural variant. An example is given in Fig. 7 where multiple views show that the CML277 gene model Zm00024ab34095 is missing several exons relative to other gene models within a pan-gene locus.

Conclusion
As pan-gene analysis and visualization is an active area of research and development, the pan-gene data center at MaizeGDB will continue to evolve. Future plans include adding pan-grass gene family analyses so users can connect gene models from the well-studied genomes, like the representative maize genome, B73, to grass gene models in order to study pan-grass relationships for a given gene. In this way, the addition of pan-gene and gene family data at MaizeGDB will help maize researchers study the complexity and diversity of Zea, and to use the comparative functions such as the CDS and protein alignment data, the gene tree data, and the GCV visualization tool to validate pan-gene relationships for a selected gene model.

Data availability
All pan-gene data is available at MaizeGDB, including bulk download files at this link https://download.maizegdb.org/Pan-genes/ Pan-Zea. Pandagma is available on GitHub (https://github.com/legumeinfo/pandagma), and the scripts used to process the Pandagma input and output files, including configuration files, are also available on GitHub (https://github.com/Maize-Genetics-and-Genomics-Database/pandagma_scripts/blob/main/config/ pan_zea.conf). Supplementary files are as follows: Supplementary File 1: Pandagma configuration file. Supplementary Table 1: assembly/annotations used in analysis along with gene counts, sizes, and percentage of gene models that were placed in pan-genes. Supplemental material available at GENETICS online.

Acknowledgements
The authors wish to thank Steven Cannon for his assistance with adapting Pandagma to work with Zea and grass genomes and annotations.

Funding
This research was supported by the United States Department of Agriculture, Agricultural Research Service, Project Number [5030-21000-068-00-D] to the Corn Insects and Crop Genetics Research Unit in Ames, Iowa. The work for this project was also conducted under a USDA-ARS nonassistance cooperative agreement (Accession 440229, University of Missouri). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and Employer.

Conflicts of interest
The authors declare no conflict of interest.

Literature cited
Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, et al. 2019. The tomato pan-


Editor: J. Blake