OXFORD GENETICS

GENETICS, 2025, 232(4), iyaf266
https://doi.org/10.1093/genetics/iyaf266
Advance Access Publication Date: 19 January 2026
Knowledgebase and Database Resources

SorghumBase: a knowledgebase for sorghum genomics, phenomics, and stakeholder engagement

Nicholas Gladman ,[1,2] Andrew Olson ,[2] Sunita Kumari,[2] Sharon Wei ,[2] Kapeel Chougule,[2] Zhenyuan Lu,[2] Marcela K. Tello-Ruiz ,[2] Peter Van Buren,[2] Vivek Kumar,[2] Lifang Zhang,[2] Audra Olson,[2] Catherine Kim,[2] Janeen Braynen,[2] Chad Hayes,[3,4] Zhanguo Xin ,[4] Robert Klein,[5] William Rooney,[6] Nicholas Provart ,[7] Asher Pasha ,[7] Abigail O'Meara,[2] Nadia Shakoor,[8] Todd P. Michael,[9] Melanie Harrison,[10] Doreen Ware[1,2,*]

[1]Robert W. Holley Center for Agriculture and Health, US Department of Agriculture-Agricultural Research Service, Cornell University, Ithaca, NY 14853, United States
[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, United States
[3]Institute of Genomics for Crop Abiotic Stress Tolerance, Department of Plant and Soil Science, Texas Tech University, Lubbock, TX 79409, United States
[4]Plant Stress and Germplasm Development Unit, US Department of Agriculture-Agricultural Research Service, Cropping Systems Research Laboratory, Lubbock, TX 79415, United States
[5]Southern Plains Agricultural Research Center, US Department of Agriculture-Agricultural Research Service, College Station, TX 77845, United States
[6]Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843-2474, United States
[7]Department of Cell & Systems Biology/Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON M5S 3B2, Canada
[8]Donald Danforth Plant Science Center, St. Louis, MO 63132, United States
[9]The Plant Molecular and Cellular Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, United States
[10]Plant Genetic Resources Conservation Unit, US Department of Agriculture-Agricultural Research Service, Griffin, GA 30223, United States

*Corresponding author: Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, United States. Email: Doreen.Ware@usda.gov

Centralizing valuable community data and resources into a user-friendly interface and accessible repository has become essential for agricultural science; embracing Findable Accessible, Interoperable, and Reusable (FAIR) principles is now standard for effective databases. SorghumBase (https://www.sorghumbase.org) is a knowledgebase designed for the sorghum research community. The SorghumBase team curates genomic, transcriptomic, variation, and phenotypic information and aggregates community events, providing rich visualizations and bulk data access. The modular framework of the database is built with open-access software to yield a robust, modifiable, and sustainable data infrastructure. Release 9 of SorghumBase includes: (i) 88 sorghum reference genomes and an updated pan-gene index, (ii) over 100 million variants have been mapped onto the 2 genomes, BTx623 and Tx2783, (iii) assignment of 41 million Reference Cluster SNP identifiers (rsIDs) from BTx623 across the pan-genome, (iv) updated gene search homology, gene expression, and germplasm visualizations and features, (v) added and standardized 234 phenotypic data from 40 community-generated GWAS studies and 148 traits from the Sorghum QTL Atlas (Oz Sorghum), (vi) improved news, funding, and a research content management system for community access and interaction, (vii) outreach materials including training documents and videos, and (viii) community engagement initiatives through training and working groups. SorghumBase serves as a hub for sorghum data and stakeholder engagement while promoting community standards to drive research and multi-omics breeding approaches.

Keywords: sorghum; genomics; pan-genomes; phenomics; variation; germplasm; breeding; resource; knowledgebase

Introduction

SorghumBase functions as a central hub for genomic, phenotypic, and community resource information for the sorghum stakeholders and is funded by the United States Department of Agriculture (USDA). For a quick use-case overview for the knowledgebase, users can access this training video (https://youtu.be/XL881C0KVi4). Originally designed to support and lead stewardship and sharing of the emergent sorghum genomic and genetic data, SorghumBase has accelerated the accumulation of genomic, genetic, and phenomic information generated by the sorghum research community as well as the SorghumBase team (Figs. 1 and 2a). Through the promotion of community standards and adherence to the Findable Accessible, Interoperable, and Reusable (FAIR) principles (Wilkinson et al. 2016) of data management, SorghumBase focuses on improving management of genomic and phenotypic data sets. This includes participation and leadership roles in the AgBioData Consortium working groups (https://www.agbiodata.org/), engagement with Breeding Insight (https://breedinginsight.org/) to establish field-base and phenotypic data standards, and coordinating with the Germplasm Resource Information Network (GRIN) (Byrne et al. 2018) from the U.S. National Plant Germplasm System (https://www.grin-global.org/) to host germplasm information and ordering infrastructure on SorghumBase. Ultimately, this works toward the SorghumBase team's efforts of biocuration of community data, open access to hosted content, and information sharing with the sorghum research and breeding community. The primary initiative for SorghumBase is the cultivation of sorghum reference genomes, variation information, and recent phenotypic content. The portal uses Ensembl and Gramene open source software (Kersey et al.

Received on 05 September 2025; accepted on 03 December 2025
Published by Oxford University Press on behalf of The Genetics Society of America 2026.

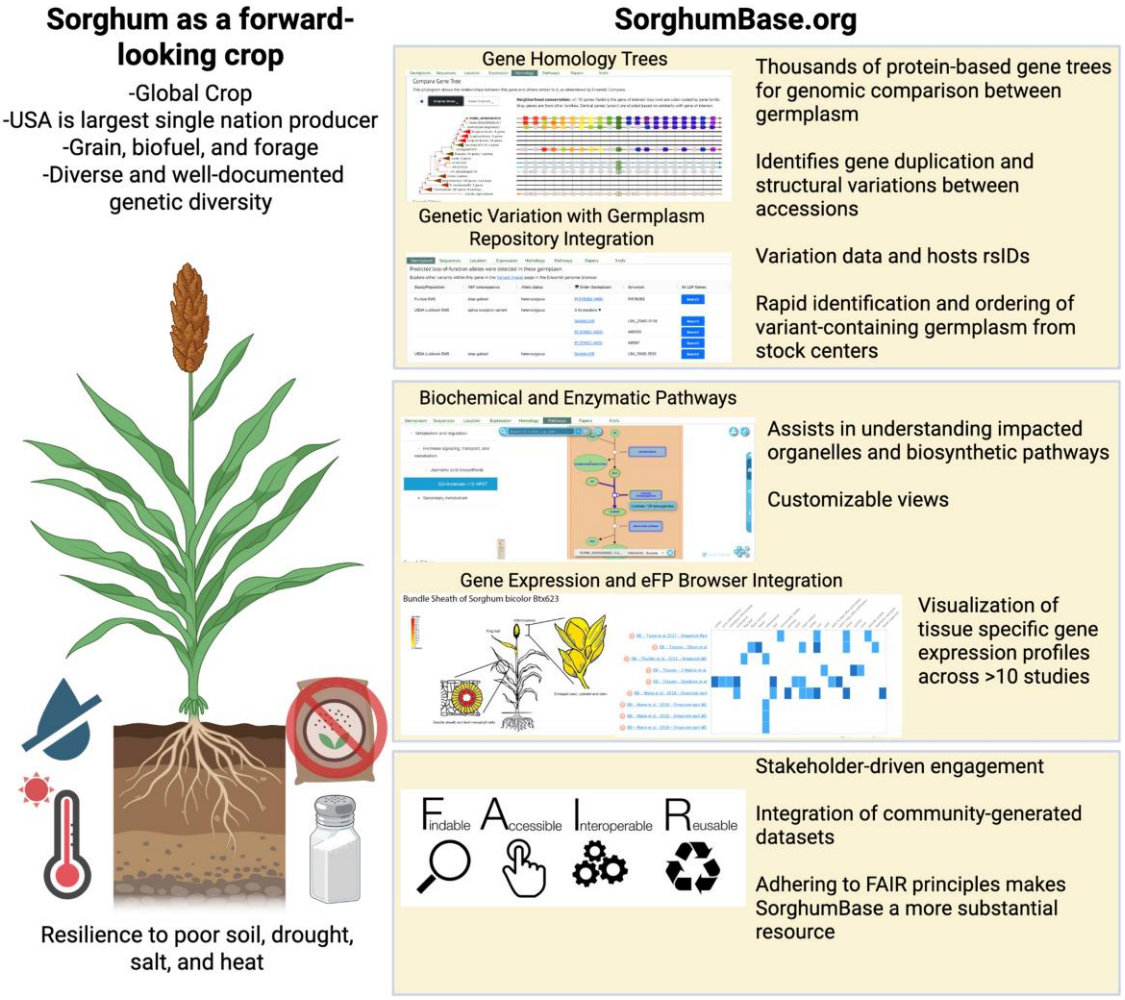

Fig. 1. SorghumBase overview. Sorghum is a globally important C4 crop with extensive genomic diversity represented across multiple germplasm repositories. Renowned for its resilience to salt, drought, and heat stress, sorghum also serves as a monocot model for abiotic stress research. SorghumBase is a comprehensive knowledgebase designed to support sorghum researchers and breeders through an integrated suite of tools and resources. These include gene tree–based homology and germplasm views that present the genetic diversity and neighborhood context of all annotated gene models across the sorghum pan-genome. The platform also enables rapid identification of germplasm carrying PTVs, which can be directly ordered from germplasm repositories. Additional features include detailed biochemical and enzymatic pathway views along with gene expression data (with eFP browser visualizations). Guided by FAIR principles, SorghumBase enhances interoperability with other community databases and resources, providing a USDA-funded platform that serves the needs of the entire sorghum research and breeding community. FAIR image taken from Wikimedia Commons. Figure created with Biorender.

2018; Tello-Ruiz et al. 2021) and is built on 25 years of development in data models, workflows, robust visualizations, and application programming interfaces (APIs).

Community engagement and site features

Working groups and community feedback

To continuously build off the initial SorghumBase release (Gladman et al. 2022), several community-oriented working groups were established to prioritize the needs of stakeholders. The Genome Working Group advised on germplasm targets for constructing reference genome assemblies based on agricultural traits (e.g. disease resistance) and breeding value. This resulted in the creation of nine genome assemblies by the SorghumBase team (Fig. 2b). The Community Marker Panel Working Group partnered with several private and public stakeholders to develop a commercial mid-density marker panel whose service was provided by a United States-based producer. The Genetic Variation and Phenotyping Working Group applied standard ontologies to community-generated Genome-Wide Association Studies (GWAS) and Quantitative Trait Locus (QTL) data for incorporation into site views; additionally, creating interoperability with the Sorghum OZ QTL Atlas (Mace et al. 2019). This working group also advanced the adoption and assignment of Reference Cluster SNP identifiers (rsIDs) to variants that were subsequently mapped across the sorghum pan-genome. The SorghumBase User Working Group comprised a small panel of sorghum experts from public and private institutions to generate granular feedback on the SorghumBase platform as well as community targets. Multiple members of the SorghumBase team also participate in various roles, including chairperson positions, on several AgBioData Consortium working groups in order to foster standardized languages, practices, and their ultimate adoption within the agricultural research and database community (Harper et al. 2018; Deng et al. 2023; Cannon et al. 2025; Marrano et al. 2025).

Community features and training

To drive broad stakeholder engagement with the website, SorghumBase expanded news and featured research content with the goal of providing a singular resource that a diverse

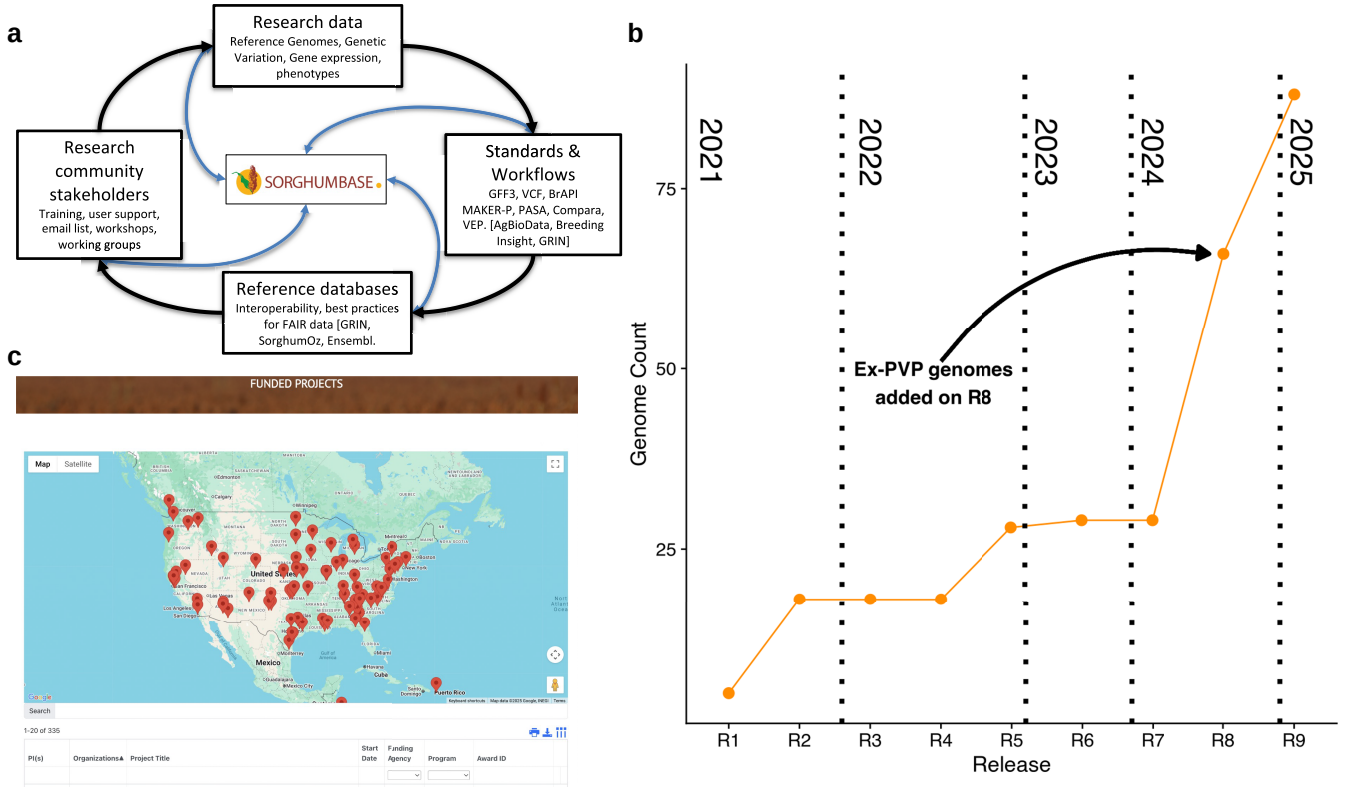

Fig. 2. SorghumBase content for increasing stakeholder value. a) The "virtuous cycle" of the SorghumBase mission that adheres to FAIR principles: hosting sorghum genomic and phenomic data; championing and implementing standards and workflows for data; increasing user value through external database interoperability; and plentiful user outreach and training activities and materials. The black arrows symbolize how each mission block feeds into each other while the blue arrows indicate how they are integrated into SorghumBase while also feeding back into community productivity. b) Displays the growth in the number of hosted genomes by database release (x axis) and year (dotted vertical lines), highlighting the rapid growth after R7 due to the incorporation of ex-PVP lines and the SorghumBase team generating their own reference assemblies. c) The Funded Projects page of SorghumBase that shows the physical location of funded programs with a full sortable table for users to identify various sorghum research projects and their principal investigators.

sorghum community could find useful. The content management system we employ uses WordPress to easily generate web pages and add content all while maintaining customizable design options. In addition to experimental data and metadata curation, we continue to curate all publications that are useful to sorghum stakeholders in the Publications section as well as create several research highlights every month that feature a handful of articles the SorghumBase team feels are particularly relevant to sorghum breeders and researchers. These research highlights also can include information, additional images, and context from the research group that authored the study. Currently over 1,500 publications and 150 research highlights exist and are updated monthly. News and Events pages on SorghumBase contain past, current, and future happenings that are important to domestic and international sorghum researchers and breeders. The Funded Projects page was created to facilitate identifying and collaborating with other sorghum researchers: it lists all publicly funded projects pertinent to sorghum stakeholders. This feature allows users to search and find the location, program description, principal investigators, and contact information for each project curated by the SorghumBase team (Fig. 2c). All of these community engagement initiatives and increased features have been added on top of numerous video and written training materials that are available on the website as well as YouTube (https://www.youtube.com/@sorghumbase9338). This includes our User Guides and Release Notes that cover new content for each site release. Additional training and outreach occur through our continual presence at major scientific conferences throughout the year where the SorghumBase team gives workshops and updates to site functionality and community projects.

Data types and user interface

Genomes, gene trees, and Sorghum pan-genome

The foundation of SorghumBase has been its genomic content and user-friendly visualizations. Since its first release in 2021, the knowledgebase has been a steward of community-generated reference genomes and has expanded its pan-genome across subsequent releases (Fig. 2b). Specifically, homology-based gene trees aid in the characterization of a pan-gene set, as well as phylogenetic dating of genes and the identification of orthologs and paralogs across and within species. Due to the increased use cases of other sorghum genotypes by the community, notably RTx430 for CRISPR genome editing, Plant Variety Protection (ex-PVP) lines for proximity to breeding materials (Mascher et al. 2024), and Sorghum Association Panel accessions for a variety of agronomic traits (Casa et al. 2008), we have focused on either hosting these existing reference assemblies or generating our own for inclusion in SorghumBase and our pan-genome annotation pipeline. Expanding and strengthening the characterization of gene trees and their respective haplotypes hold value for numerous tasks. Such examples include inferring gene function, trait-based selection for germplasm inclusion into breeding programs, and improving genetic diversity within a mapping population.

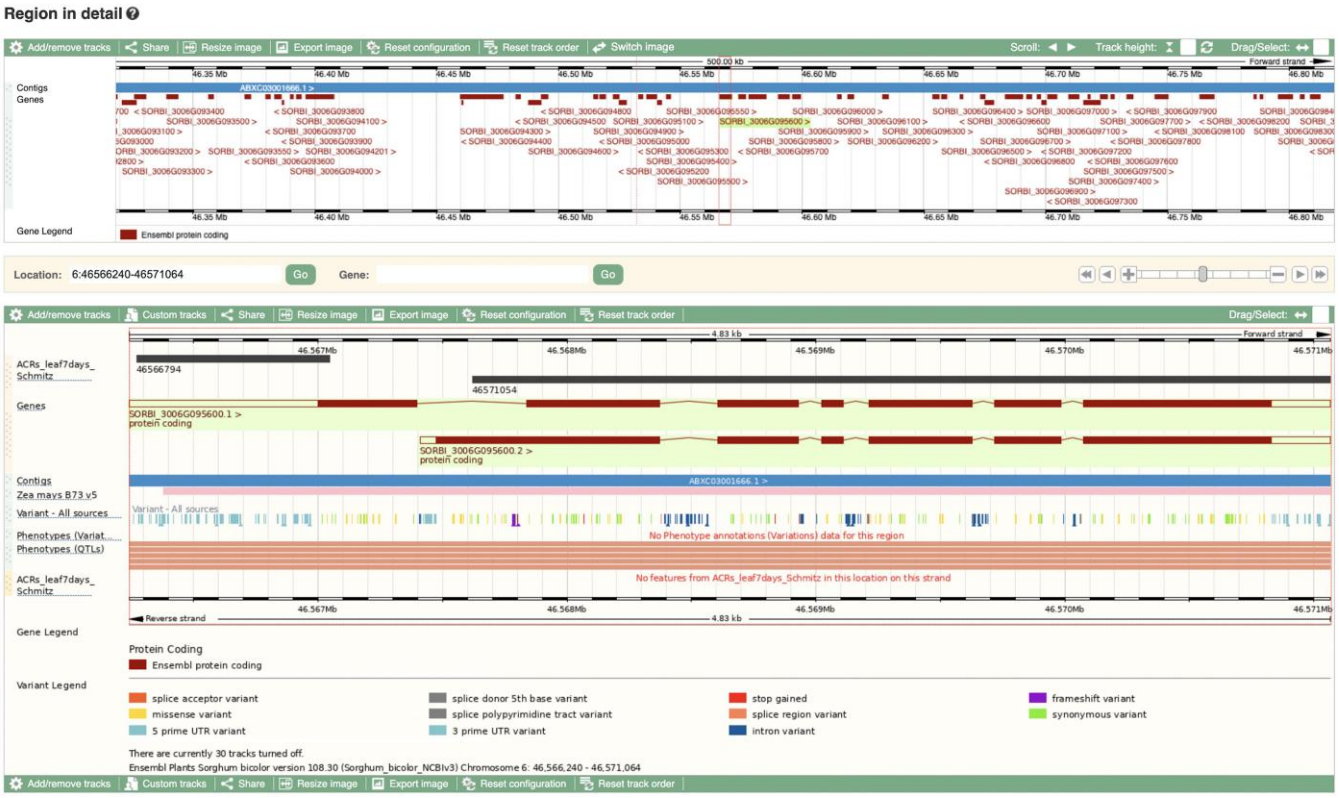

Fig. 3. SorghumBase genome browser view. SorghumBase contains a genome browser view with customizable tracks as well as the ability to add custom, private data for viewing. The highlighted in view below the Region in Detail view shows the *MSD2* gene model (SORBI_3006G095600) flanked by genomic tracks for active chromatin regions, synteny to Zea mays, all variant information, and selected phenomics tracks from QTL studies. A popover will appear when selecting tracks, such as with the QTL track for panicle width. Tracks can be modified based on visual style and extent of the experimental content hosted on the site. The available repetitive regions track and other ChIP-seq data for assorted histone methylation and acetylation studies and expression are not shown.

This transformation has resulted in novel functionality, data type integration, visualizations, and interoperability with other databases. As of Release 9, there are 88 sorghum genomes hosted on the site in addition to 7 outgroup genomes. This pan-genome approach sets a groundwork for constructing robust protein-based gene trees. Initially, 7 outgroups [*Arabidopsis thaliana* (TAIR10) (Berardini et al. 2015), *Chlamydomonas reinhardtii* (*Chlamydomonas_reinhardtii*_v5.5) (Merchant et al. 2007), *Drosophila melanogaster* (BDGP6) (dos Santos et al. 2015), *Oryza sativa* (IRGSP-1.0) (Kawahara et al. 2013), *Selaginella moellendorffii* (v1.0) (Banks et al. 2011), *Vitis vinifera* (IGGP_12x) (Jaillon et al. 2007), and B73 *Zea mays* (AGPv4) (Jiao et al. 2017)] were used in conjunction with 5 sorghum reference assemblies as inputs for the Ensembl Protein Comparative phylogenetic analysis. This has been increased in Release 9 to include 48 hosted sorghum genomes for gene tree calling (Ensembl Compara pipeline version-87 with database scheme e108) (Herrero et al. 2016). As of Release 9, 48 genomes were used to build 44,841 protein-coding gene family trees. These are constructed via the peptide encoded by the canonical transcript (Olson and Ware 2021) (i.e. a representative transcript for a given gene) of each 1,806,704 individual genes based on 1,856,883 input proteins. The gene tree analysis classifies evolutionary events (e.g. speciation and duplication) and detects annotation artifacts such as split genes (12,258 across all genomes in Release 9).

At present, full de novo structural annotation is performed exclusively on genomes that are either sequenced in-house or provided by active collaborators. For all other assemblies, we intake existing community-derived gene models as provided by the original sources (e.g. published experimental data). However, to ensure consistency in downstream analyses, we uniformly apply functional annotation across all genomes using InterProScan (v5.38-76.0) (Jones et al. 2014) and repeat annotation using EDTA (v2.1.0) (Ou et al. 2019). Upon completion and validation of our pan-gene index, we intend to implement a standardized structural and functional annotation pipeline across the entire genome collection and distribute these standardized annotations as browser-accessible track files. We construct our pan-gene index using a pan-genomic framework that integrates representative pan-gene models, which are selected based on a comparative analysis of gene family phylogenies generated via the Ensembl Compara pipeline (Supplementary Fig. 1). To propagate these pan-gene representatives onto the genome assemblies of other unannotated accessions, we employ Liftoff (v1.6.3) (Shumate and Salzberg 2021) and subsequently enhance the gene structures using available transcriptome evidence through PASA (v2.5.3) (Haas et al. 2003) and score the final protein coding models via the Annotation Edit Distance (Campbell et al. 2014). This approach was initially applied to 10 CP-NAM genomes (Voelker et al. 2022) from the SorghumBase genome collections, resulting in the annotation of approximately 41,000 gene models, 6,000 to 7,000 more than those in the BTx623 reference genome. Most of these additional genes are single-exon, with about 30% supported by transcriptomic or homology-based evidence. All models and genomic content are viewable in the genome browser with customizable tracks with display options for accessible chromatin

Fig. 4. SorghumBase gene tree views. a) The default view in the Homology tab for a gene search result is the Alignment Overview, which displays a cladogram of the selected gene model and its orthologs across the species in SorghumBase. The cladogram is customizable and each clade can be collapsed or expanded. Colored sections of the aligned gene model display the annotated functional domains in the protein structure. Users can flag gene models for manual curation by selecting the Curate button at the bottom of the Homology tab page (not shown). b) The Neighborhood Conservation view is another display option in the Homology tab that shows the 10 genes flanking the gene of interest (along the red centerline). Color coding informs gene model conservation and gene orientation along the chromosome (as well as strandedness) is also indicated along with any annotated noncoding elements. This view allows a rapid identification of major structural changes in the area between sorghum accessions and across more distant species.

regions and repetitive element annotations along with other syntenic, phenotypic, and variant data (Fig. 3). Recent site releases have focused on the inclusion of ex-PVP germplasm, which complements the original inbred assembly references like BTx623 that, while useful, are sometimes decades old. The ex-PVP germplasms represent closer approximations of what sorghum hybrid creation programs than older reference inbreds, and as such are useful in molecular and phenotypic discovery as well as trait integration.

The Homology tab in the gene search result pages presents these homology-based views that highlight functional protein domain alignments between different accessions in an overview (Fig. 4a), a detailed amino acid alignment view, and gene neighborhood view that displays flanking gene and noncoding elements (Fig. 4b). Using the *MSD2* Lipoxygenase gene SORBI_3006G095600 as an example, the gene trees and position information for surrounding genes are used to generate gene neighborhood views. MSD2 is involved in floral development and, when disrupted, results in complete fertility in both sessile and pedicellate spikelets, which increases seed setting in the panicle (Gladman et al. 2019). Figure 4 highlights the power of SorghumBase gene trees by allowing the customization of gene tree views in the cladogram while showing structural differences (Fig. 4a) across accessions or larger chromosomal changes relative to the queried gene model. In an attempt to engage the community in constant gene model improvement, a Curate button has been added under the homology views to allow users to flag gene models with potentially incorrect structures for future manual curation under the Apollo browser (Dunn et al. 2019) with RNA-seq evidence. Since this feature was introduced, 949 out of 9,585 BTx623 v3 curated gene models have been flagged for possible structural annotation issues.

Fig. 5. Gene expression views. a) Heatmap of normalized gene expression across different tissues for the gene of interest (*MSD2*) and its paralogs. The heatmap and the adjacent morphological diagram are interactable and highlighting the gene and tissue of interest generates a popover with additional expression information including the transcripts per million value (TPM). All expression information is generated through the EMBL EBI Expression Atlas. b) The eFP Browser tab shows electronic fluorescent pictographs (eFPs) generated to highlight gene expression across different plant structures that are derived from different publicly available studies. The eFP views were created in collaboration with the Bio-Analytic Resource for Plant Biology (BAR) at the University of Toronto.

Metabolic and regulatory pathways

Plants are an incredible chemical resource for natural product production. To facilitate the understanding of these complex natural pathways, the Pathway tab in the SorghumBase gene search page provides the visualization of pathways that cover metabolic and transport pathways, hormone signaling, genetic regulations of developmental processes, and transcriptional networks (Supplementary Fig. 2). These pathway views come from the Plant Reactome project, an open-access pathway database that manually curates various biological pathways and signaling cascades (Gupta et al. 2024). There are 271 orthology-based pathways for sorghum which were projected from curated Nipponbare rice pathways (Gupta et al. 2024). Users can interact with the Pathway tab view to customize the image space, navigate surrounding pathways and sub-cellular localization views, search and filter for pathway content, and export image views.

Gene expression

The published sorghum bulk RNA-seq expression data from the National Center For Biotechnology Information Sequence Read Archive (NCBI SRA) and the European Molecular Biology Lab European Nucleotide Archive (EMBL ENA) are prioritized based on the sorghum community needs on abiotic stress, growth and development studies. Currently, 11 published bulk RNA-seq datasets of sorghum have been selected for SorghumBase. This dataset is extensively manually curated, validated, and configured by SorghumBase in collaboration with the EMBL EBI Expression Atlas (www.ebi.ac.uk/gxa) prior to ingestion into the Atlas (Papatheodorou et al. 2020; Moreno et al. 2022; George et al. 2024). The expression dataset is indexed and integrated into the gene search results feature and then accessible via the SorghumBase search browser.

The Expression tab in the gene results of the SorghumBase search interface allows users to visualize gene expression levels across tissue and different environmental conditions via a heatmap (Fig. 5a). The Expression subtab All Studies allows users to explore the baseline expression of their gene of interest across different tissues and developmental stages for all the studies hosted on the Expression Atlas. Currently, the baseline gene expression data from eight sorghum BTx623 datasets have been curated and processed by the EMBL EBI Expression Atlas team in collaboration with SorghumBase. The differential expression data of 3 studies is available directly from the EMBL EBI Expression Atlas. There is also a Paralog view that will show the baseline or differential expression profiles for the gene of interest as well as all of its predicted paralogs. Both of these views allow users to identify tissue-specific expression for a given gene and its closest related gene models, which can rapidly convey possible gene activity and redundancy for future molecular characterization.

The eFP Browser tab under the Expression tab displays electronic fluorescent pictograph (eFP) images from 10 studies of sorghum seedlings, roots, stem, leaves, panicles, seeds, and more across different developmental stages and under various abiotic stress conditions (Fig. 5b). The eFP Browser (Winter et al. 2007) was developed by the Bio-Analytic Resource (BAR) for Plant Biology of the University of Toronto (https://www.bar.utoronto.ca/) in collaboration with the SorghumBase team to facilitate visual exploration of gene expression. The eFP images of different studies show the different tissues and organs at various stages of development under different environmental conditions. Each image displays a different hue reflecting organ level expression for the gene being viewed. The color is generated based on a linear scale from zero (yellow) to the maximum expression (red) of the gene on the specific transcripts per million (TPM) value of the queried gene among the tissues corresponding to the expression level of that gene in that tissue making it easier to compare the gene expression across different tissues among different studies.

Variation and germplasm selection

The SorghumBase team has prioritized FAIR principles through hosting and curation of variation data across sorghum germplasm. This includes 78 million SNP and indel variants that have been mapped to the BTx623v3.1 reference assembly, including over 65 million natural variants from over 4,400 accessions

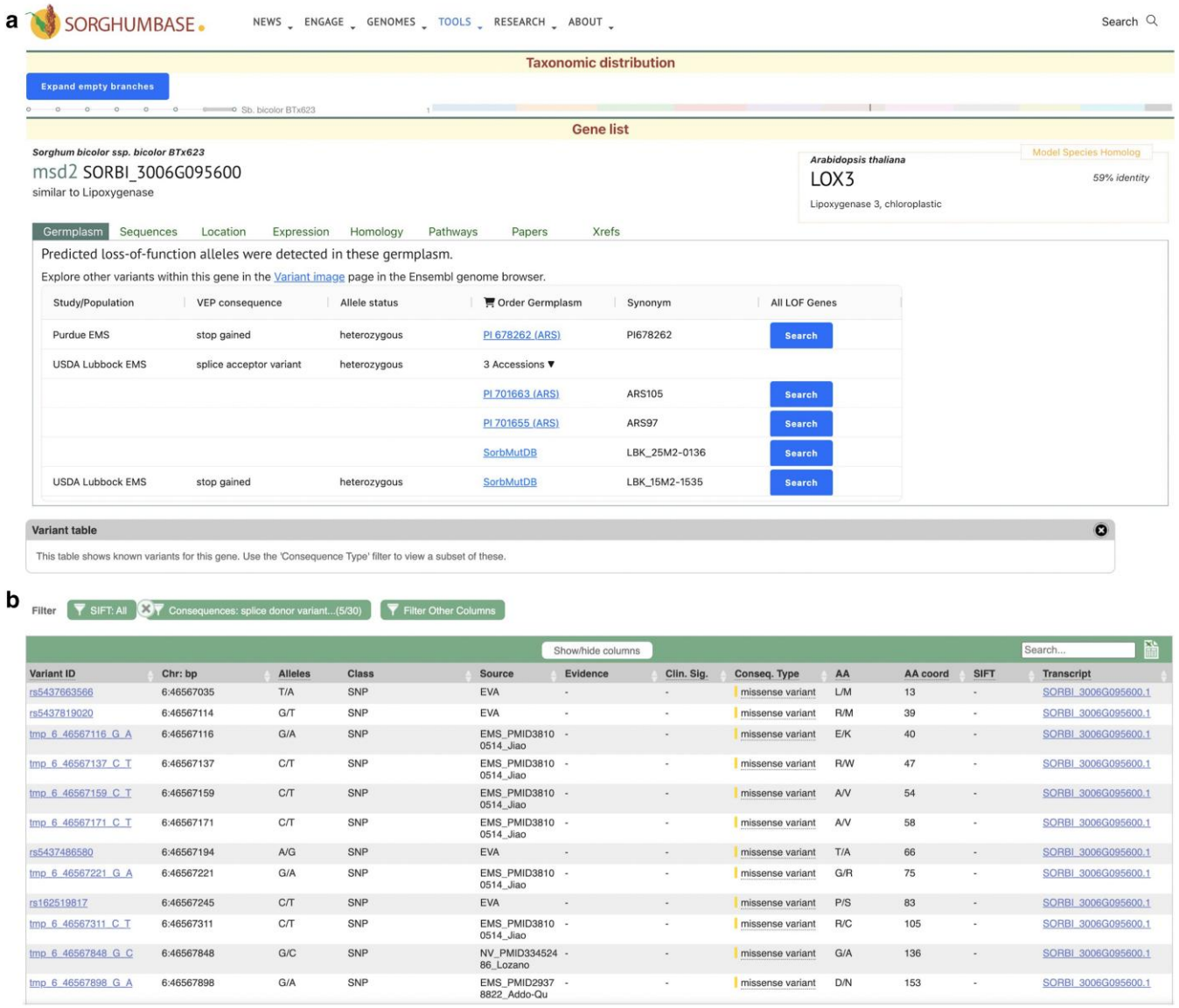

Fig. 6. Germplasm with impactful variants. a) The Germplasm tab in the gene search result page displays all likely loss-of-function variants that result from PTVs in the gene model. All germplasms with a PTV are shown in relation to the population (study) in which they were originally characterized. This figure shows Purdue and USDA Lubbock EMS-induced populations and multiple natural variant populations. Users can click the link in the "Order Germplasm" column to be sent to the germplasm repository page that curates and distributes the germplasm stocks to researchers (currently GRIN and SorbMutDB). This type of PTV search can be done reciprocally for a germplasm, where you can select the link in the "All LOF Genes" column and SorghumBase will display all gene models with PTVs in that specific germplasm. This can be a quick reference for researchers doing forward genetic screens to identify possible candidate genes for observed phenotypes in a given germplasm. b) The Variant Table for the *MSD2* gene (Lipoxygenase) displays all variants in and surrounding the gene model. This is a sortable, filterable table with additional feature columns for all variants associated with the gene model of interest. Selecting the Variant ID link will send the user to a more detailed variant information page that includes information regarding allele frequencies, germplasm association, and any existing phenotypic information.

(Lasky et al. 2015; Lozano et al. 2021; Boatwright et al. 2022; Kumar et al. 2024) nearly 13 million variants from ethyl methanesulfonate (EMS)-induced mutagenesis in BTx623v3.1 (Jiao et al. 2016, 2024; Addo-Quaye et al. 2018), 36.1 million variants from 940 accessions resequenced by the Sorghum Genome Toolbox project (https://www.globalsorghuminitiative.org/) mapped to BTx623v5.1, and over 32 million variants in the Tx2783 assembly (Zhou et al. 2024). The SorghumBase team has spearheaded the assignment of 41 million rsIDs from the BTx623v3.1 genome assembly to all 88 hosted sorghum accessions prioritized in coordination with the sorghum community. Assignment of an rsID is done by identifying an identical allele based on conservation of its flanking sequence; this permanent identifier permits interoperability across databases and a reliable association of phenotypes with genetic variants (Supplementary Fig. 3). All variants that localize within gene models are subjected to a Variant Effect Predictor analysis to assess the impact on protein structure and function (Hunt et al. 2022). SorghumBase indexes germplasm associated with gene model variants that have been classified as protein truncating variants (PTVs), which are likely to result in loss-of-function phenotypes, such as stop gained, splice site alteration, or frameshift. This Germplasm tab shows all predicted PTVs in wild or EMS-induced accessions (Fig. 6a), with their name or unique plant introduction (PI) numbers and a link to the germplasm

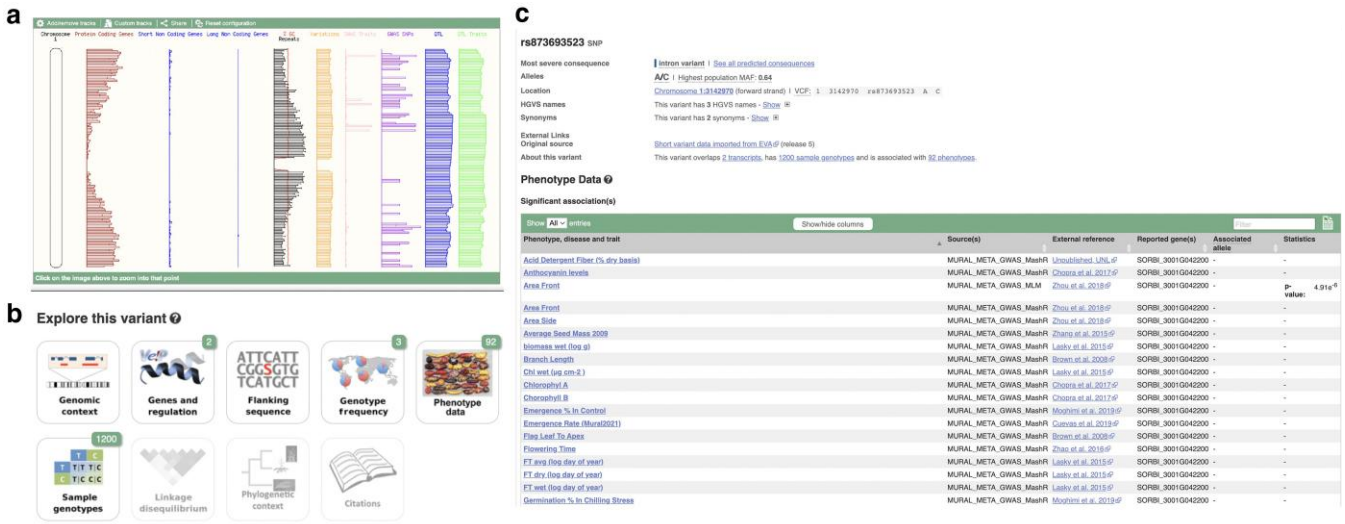

Fig. 7. Phenomic information in SorghumBase. a) A karyotype-like view of chromosome 1 that has interactive mappings of GWAS traits, GWAS SNPs, QTL traits, and QTL SNPs along with additional genomic metrics. Clicking on a portion of the karyotype will send the user to more detailed pages on the respective trait or variant. b) An example of a detailed variant information page that contains phenotype data along with standard data on genomic context, flanking sequences, genotype frequency, and sample genotypes (this identifies which germplasm in which populations contain major or minor alleles associated with the variant ID). c) The Phenotype Data page for an rsID variation displaying all the QTLs that have used this SNP. In this particular example, rs873693523 has been used in several studies, but only one QTL study "Area Front" under the "Phenotype, disease and trait" column shows a statistically significant hit.

repository where stocks can be ordered: SorghumMutDB (Jiao et al. 2024) or GRIN (Byrne et al. 2018) from the U.S. National Plant Germplasm System. This search feature is also reciprocal: any germplasm can be searched via its PI number and the results page will list all gene models that have PTVs for that given germplasm. This feature is useful for breeders to quickly identify any impactful variant within a gene model during a forward genetic screen. All variants can be viewed in the genome browser as a customizable track as well as 2 detailed views for every gene model in the form of a variant image and variant table. The variant image is an interactive genome browser page showing all variants within and surrounding a given gene model, and the variant table is a tabular list of the variants that is searchable and filterable based on variant characteristics like study source and variant impact (Fig. 6b). More detailed information for each variant (both with and without rsIDs) exists as an Ensembl page view and includes positional information, allele frequency within a given population, any available phenotypic information, and accession name or PI numbers and genotypes for all germplasm within a population that contains a specific variant.

Phenotypic data

SorghumBase continues to host QTLs from the Sorghum QTL Atlas (Mace et al. 2019) (https://aussorgm.org.au/sorghum-qtl-atlas/) and GWAS data from over 230 phenotypic traits sets across 25 studies (Mural et al. 2021). Additionally, we have begun to use consistent trait ontologies for the phenotypic descriptions to improve searchability and comparisons across studies (Fig. 7a, b). These data have also been incorporated into the genome browser view as customizable tracks (Fig. 3). Phenotypic information has been linked to their respective variants, and detailed information is contained within the phenotype section of variant pages, which includes links to the original study and standardized trait ontologies that summarize other variants associated with that trait (Fig. 7c).

Future directions

SorghumBase will continue to host data that is valuable to sorghum stakeholders while also engaging the community to better understand the needs and features that the SorghumBase team can provide. All improvements and advancements will be in line with FAIR principles through our collaborations with other partners to advance community standards while improving integration of genetic, phenotypic, and other data that is valued by sorghum stakeholders. However, this is complicated by not all standards existing or are widely adapted for phenomics data, so in order to work with sorghum and other crops, we must ensure standards are uniformly applied and informative. Our priority will still be the inclusion of reference genome assemblies to further bolster our pan-gene index: genomic variation, rsID assignment, and interoperability with germplasm repositories will remain a point of focus. But this priority will come with a shift in focus from individual reference assemblies to a pan-genome approach to benefit stakeholders, particularly breeders; ultimately benefiting cross-species and sorghum accession comparisons. While we have made headway into improving phenotypic metadata within the knowledgebase, our team will make further advances in standardizing trait ontologies, as well as engaging with other organizations like Planteome, AgBioData, and Breeding Insight to appropriately interoperate field- and trait-based data and metadata into the website, including being Breeding API (BRAPI)-compliant. This persistent integration of phenotypic data highlights the intention of SorghumBase to strengthen its role within the sorghum research community by effectively coupling genomic information with its phenomics complement; this is complicated by the diverse and non-uniform nature of phenotypic data and metadata. However, there are a subset of new phenotypic targets that will be incorporated into future releases of SorghumBase, including geospatial collection data for hundreds of sorghum accessions as well as grain size and color metrics. To further improve upon our open-source pipelines for genomic content generation, annotation, and

visualization, we will incorporate machine learning approaches to assist in (i) generating candidate gene lists and markers for breeding traits based on existing publications, (ii) curating lists of gene symbols and their associated ontologies and biological pathways, (iii) incorporation of existing protein-prediction models such as AlphaFold (https://alphafold.ebi.ac.uk/), and (iv) aiding in creation of research articles and other selected community outreach materials. On top of all our data hosting, members of the SorghumBase development team will continue to develop and operate workshops, seminars, and online training materials to inform and encourage stakeholders to use and improve the site. This reflects SorghumBase's ultimate goal of providing an essential resource to the community while adapting its features to current agricultural and biotechnical methods.

Data availability

All datasets described in this article have been previously published, are public, and can be found linked from https://www.sorghumbase.org. All web resources discussed are summarized here:

- https://breedinginsight.org/ Breeding Insight (USDA genotype-to-phenotype initiative)
- https://www.grin-global.org/ U.S. National Plant Germplasm System (germplasm repository and distribution)
- https://www.youtube.com/@sorghumbase9338 SorghumBase training videos (also located on SoghumBase.org)
- www.ebi.ac.uk/gxa EMBL EBI Expression Atlas (gene expression information)
- https://www.bar.utoronto.ca/ Bio-Analytic Resource (BAR) for Plant Biology (electronic fluorescent pictographs)
- https://www.globalsorghuminitiative.org/ Sorghum Genome Toolbox project
- https://aussorgm.org.au/sorghum-qtl-atlas/ Sorghum QTL Atlas (QTL database)

Supplemental material available at GENETICS online.

Acknowledgment

We thank the domestic and international sorghum community members for public release of data, ongoing discussions about community needs, and general site feedback. We acknowledge all members of our various working groups, particularly the SorghumBase User Working Group (Stephen Kresovich, Mitch Tuinstra, and Laura Mayor), the National Sorghum Producers (https://sorghumgrowers.com), and Sorghum Checkoff (https://www.sorghumcheckoff.com/) for their continued support and feedback. We thank our infrastructure partners, specifically their support on standards and making their sorghum data FAIR: AgBioData Consortium, SorghumOz database for their phenotype data, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) for their germplasm, Ensembl-EVA for their work on rsID assignment, and Joint Genome Institute (JGI) for their creation and access to numerous reference genome assemblies. We gratefully acknowledge Mercer University faculty member Dr. John Stanga and his undergraduate students for their collaboration in sorghum gene model curation.

Funding

SorghumBase is funded through the United States Department of Agriculture grant USDA-ARS 8062-21000-051-000D. This work was performed with assistance from the United States National Institutes of Health Grant S10OD028632-01. This research used resources provided by the SCINet project and/or the AI Center of Excellence of the USDA Agricultural Research Service, ARS project numbers 0201-88888-003-000D and 0201-88888-002-000D. Additional genome resources were supported through the Salk Institute for Biological Studies Harnessing Plants Initiative (HPI) with funding from the TED Audacious, Bezos Earth Fund, Sempra and Hess Corporation to TPM and NS. This research was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under DE-AR0000594, and the Bill & Melinda Gates Foundation under Award Number OPP1129063. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or of the Bill & Melinda Gates Foundation. NJP received funding from the National Sciences and Engineering Research Council of Canada (NSERC), which helped support the development of the Sorghum eFP Browser.

Conflicts of interest

None declared.

Literature cited

Addo-Quaye C, Tuinstra M, Carraro N, Weil C, Dilkes BP. 2018. Whole-genome sequence accuracy is improved by replication in a population of mutagenized sorghum. G3 (Bethesda). 8: 1079–1094. https://doi.org/10.1534/g3.117.300301.

Banks JA *et al.* 2011. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science. 332:960–963. https://doi.org/10.1126/science.1203810.

Berardini TZ *et al.* 2015. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. Genesis. 53:474–485. https://doi.org/10.1002/dvg.22877.

Boatwright JL *et al.* 2022. Sorghum association panel whole-genome sequencing establishes cornerstone resource for dissecting genomic diversity. Plant J. 111:888–904. https://doi.org/10.1111/tpj.15853.

Byrne PF *et al.* 2018. Sustaining the future of plant breeding: the critical role of the USDA-ARS national plant germplasm system. Crop Sci. 58:451–468. https://doi.org/10.2135/cropsci2017.05.0303.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinformatics. 48:4.11.1–4.11.39. https://doi.org/10.1002/0471250953.bi0411s48.

Cannon EKS *et al.* 2025. Guidelines for gene and genome assembly nomenclature. Genetics. 229:iyaf006. https://doi.org/10.1093/genetics/iyaf006.

Casa AM *et al.* 2008. Community resources and strategies for association mapping in sorghum. Crop Sci. 48:30–40. https://doi.org/10.2135/cropsci2007.02.0080.

Deng CH *et al.* 2023. Genotype and phenotype data standardization, utilization and integration in the big data era for agricultural sciences. Database (Oxford). 2023:baad088. https://doi.org/10.1093/database/baad088.

dos Santos G *et al.* 2015. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 43: D690–D697. https://doi.org/10.1093/nar/gku1099.

Dunn NA *et al.* 2019. Apollo: democratizing genome annotation. PLoS Comput Biol. 15:e1006790. https://doi.org/10.1371/journal.pcbi.1006790.

George N et al. 2024. Expression atlas update: insights from sequencing data at both bulk and single cell level. Nucleic Acids Res. 52: D107–D114. https://doi.org/10.1093/nar/gkad1021.

Gladman N et al. 2019. Fertility of pedicellate spikelets in sorghum is controlled by a jasmonic acid regulatory module. Int J Mol Sci. 20: 4951. https://doi.org/10.3390/ijms20194951.

Gladman N et al. 2022. SorghumBase: a web-based portal for sorghum genetic information and community advancement. Planta. 255:35. https://doi.org/10.1007/s00425-022-03821-6.

Gupta P et al. 2024. Plant reactome knowledgebase: empowering plant pathway exploration and OMICS data analysis. Nucleic Acids Res. 52:D1538–D1547. https://doi.org/10.1093/nar/gkad1052.

Haas BJ et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31:5654–5666. https://doi.org/10.1093/nar/gkg770.

Harper L et al. 2018. AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. Database. 2018:bay088. https://doi.org/10.1093/database/bay088.

Herrero J et al. 2016. Ensembl comparative genomics resources. Database (Oxford). 2016:bav096. https://doi.org/10.1093/database/bav096.

Hunt SE et al. 2022. Annotating and prioritizing genomic variants using the ensembl variant effect predictor-A tutorial. Hum Mutat. 43:986–997. https://doi.org/10.1002/humu.24298.

Jaillon O et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 449: 463–467. https://doi.org/10.1038/nature06148.

Jiao Y et al. 2016. A sorghum mutant resource as an efficient platform for gene discovery in grasses. Plant Cell. 28:1551–1562. https://doi.org/10.1105/tpc.16.00373.

Jiao Y et al. 2017. Improved maize reference genome with single-molecule technologies. Nature. 546:524–527. https://doi.org/10.1038/nature22971.

Jiao Y et al. 2024. A large sequenced mutant library—valuable reverse genetic resource that covers 98% of sorghum genes. Plant J. 117: 1543–1557. https://doi.org/10.1111/tpj.16582.

Jones P et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Kawahara Y et al. 2013. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice. 6:4. https://doi.org/10.1186/1939-8433-6-4.

Kersey PJ et al. 2018. Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res. 46: D802–D808. https://doi.org/10.1093/nar/gkx1011.

Kumar N, Boatwright JL, Boyles RE, Brenton ZW, Kresovich S. 2024. Identification of pleiotropic loci mediating structural and non-structural carbohydrate accumulation within the sorghum bioenergy association panel using high-throughput markers. Front Plant Sci. 15:1356619. https://doi.org/10.3389/fpls.2024.1356619.

Lasky JR et al. 2015. Genome-environment associations in sorghum landraces predict adaptive traits. Sci Adv. 1:e1400218. https://doi.org/10.1126/sciadv.1400218.

Lozano R et al. 2021. Comparative evolutionary genetics of deleterious load in sorghum and maize. Nat Plants. 7:17–24. https://doi.org/10.1038/s41477-020-00834-5.

Mace E et al. 2019. The Sorghum QTL Atlas: a powerful tool for trait dissection, comparative genomics and crop improvement. Theor Appl Genet. 132:751–766. https://doi.org/10.1007/s00122-018-3212-5.

Marrano A et al. 2025. A teaching and training framework to promote findable, accessible, interoperable, and reusable data generation in agriculture. Database (Oxford). 2025:baaf034. https://doi.org/10.1093/database/baaf034.

Mascher M, Jayakodi M, Shim H, Stein N. 2024. Promises and challenges of crop translational genomics. Nature. 636:585–593. https://doi.org/10.1038/s41586-024-07713-5.

Merchant SS et al. 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 318: 245–250. https://doi.org/10.1126/science.1143609.

Moreno P et al. 2022. Expression atlas update: gene and protein expression in multiple species. Nucleic Acids Res. 50:D129–D140. https://doi.org/10.1093/nar/gkab1030.

Mural RV et al. 2021. Meta-analysis identifies pleiotropic loci controlling phenotypic trade-offs in sorghum. Genetics. 218:iyab087. https://doi.org/10.1093/genetics/iyab087.

Olson AJ, Ware D. 2021. Ranked choice voting for representative transcripts with TRaCE. Bioinformatics. 38:261–264. https://doi.org/10.1093/bioinformatics/btab542.

Ou S et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20:275. https://doi.org/10.1186/s13059-019-1905-y.

Papatheodorou I et al. 2020. Expression atlas update: from tissues to single cells. Nucleic Acids Res. 48:D77–D83. https://doi.org/10.1093/nar/gkz947.

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. Bioinformatics. 37:1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016.

Tello-Ruiz MK et al. 2021. Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. Nucleic Acids Res. 49:D1452–D1463. https://doi.org/10.1093/nar/gkaa979.

Voelker WG et al. 2022. Ten new high-quality genome assemblies for diverse bioenergy sorghum genotypes. Front Plant Sci. 13: 1040909. https://doi.org/10.3389/fpls.2022.1040909.

Wilkinson MD et al. 2016. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 3:160018. https://doi.org/10.1038/sdata.2016.18.

Winter D et al. 2007. An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS One. 2:e718. https://doi.org/10.1371/journal.pone.0000718.

Zhou Y et al. 2024. A high-performance computational workflow to accelerate GATK SNP detection across a 25-genome dataset. BMC Biol. 22:13. https://doi.org/10.1186/s12915-024-01820-5.

Editor: T. Berardini