BMC Bioinformatics

DATABASE

Open Access

YamOmics: a comprehensive data resource on yam multi-omics

Yi Zhao[1†], Xuteng Ye[1†], Jun Cheng[1†], Li Yin[1], Danyu Shen[2], Daolong Dou[1,2] and Jinding Liu[1*]

[†]Yi Zhao, Xuteng Ye and Jun Cheng contributed equally to this work.

*Correspondence:
Jinding Liu
liujd@njau.edu.cn
[1]Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China
[2]Department of Plant Pathology, Nanjing Agricultural University, Nanjing 210095, China

Abstract

Yams (*Dioscorea* spp.) are a highly important class of horticultural crops, serving as a staple food for millions of people in Africa and contributing significantly to food security. They are also widely cultivated in East Asia as medicinal herbs, bringing substantial economic incomes. Diverse omics data play a pivotal role in advancing yam research and breeding. However, these data are often scattered, lacking in systematic organization and analysis, which underscores the need for centralized and comprehensive data management. In view of this, we gathered extensive omics data and developed the Yam Omics Database (YamOmics; https://biotec.njau.edu.cn/yamdb). The database currently offers a vast and diverse range of omics data, covering genomic, transcriptomic and plastomic data from 41 distinct yam species, along with detailed records of genomic variants from 935 germplasms, and gene expression profiles from 191 samples. Additionally, the database features thorough annotations, encompassing aspects like genome synteny, ortholog groups, signaling pathways, gene families and protein interactions. To support yam basic biology and breeding research, it is also equipped with a suite of user-friendly online tools, including PCR primer design, CRISPR design, expression analysis, enrichment analysis, and phylogenetic inference among Dioscorea accessions tools.

Introduction

Yams (*Dioscorea* spp.) are important horticulture crops that belong to monocotyledonous plants, but have some characteristics of dicotyledonous plants, making them important materials for fundamental botany research. Among more than 600 species of the genus *Dioscorea*, some have been domesticated for starch production, such as *D. alata*, *D. dumetorum*, *D. rotundata*, and more [1]. In some African countries, yams serve staple food, feeding millions of people, and they are also important sources of dietary diversification and are gradually increasing their market share in the global and local market (Fig. 1). On the other hand, since diosgenin saponins were first isolated from the rhizomes of *D. tokoro* in the 1930s [2], some *Dioscorea* species are also used as a favored source of medicinal plants, such as *D. nipponica*, *D. polystachya*, and *D. zingiberensis* [1, 3]. Over the past twenty years, worldwide production of yams has seen a twofold increase. However, this growth has been primarily driven by the expansion of cultivation areas, as

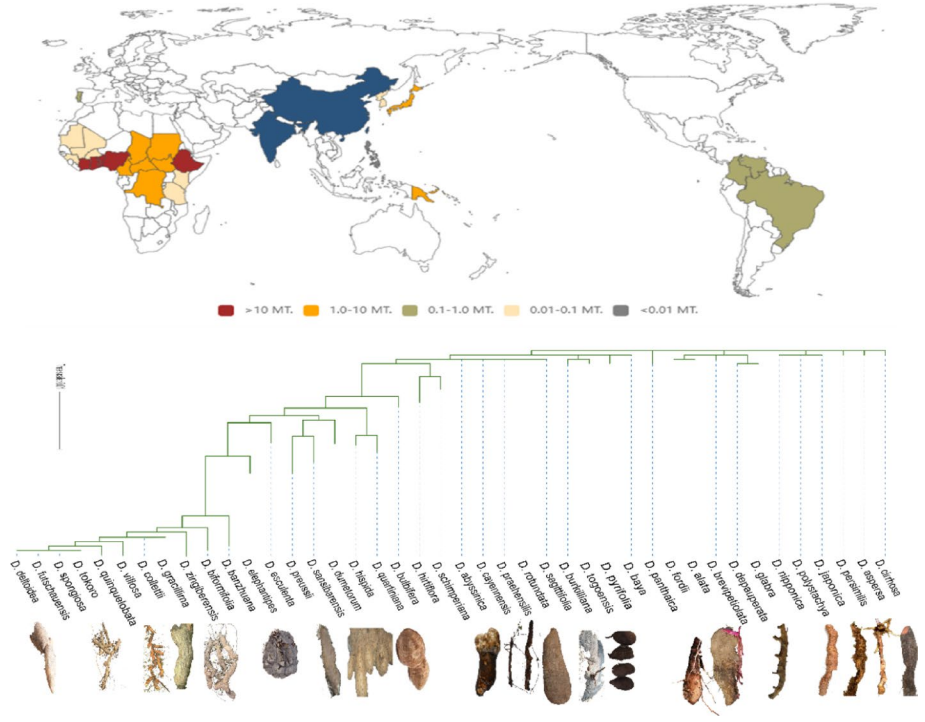

Fig. 1 The global distribution of yam producing region, phylogeny relationship among forty-one species and some tuber phenotype

opposed to enhancements in productivity (FAOSTAT 2020). Notably, the application of omics tools has laid a theoretical foundation for breaking through the bottleneck of low productivity, such as by deciphering the genetic basis of agronomically important traits and optimizing breeding strategies [4].

In 2017, to facilitate genetic research and molecular breeding in yams, the first yam genome, *D. rotundata* (white Guinea yam), was sequenced to elucidate the mechanism of sex determination, to find a genomic region associated with female heterogametic sex determination [5]. Based on this finding, a molecular marker was developed for sex identification of Guinea yam plants at the seedling stage. Subsequently, in 2020, 336 accessions of white Guinea yam were resequenced. The research resolved its origin, revealing that it was a hybrid derived from crosses between the wild rainforest species *D. praehensilis* and the savannah adapted species *D. abyssinica* [6]. In 2022, the genome of *D. alata* was generated combined with a dense genetic map from African breeding populations. Using these genomic tools, users located the quantitative trait loci for resistance to anthracnose and several tuber quality traits, providing important references for yam breeding. In terms of medicinal Dioscorea species, the genome of *D. zingiberensis* was also sequenced to elucidate the biosynthesis mechanism of diosgenin saponins [7]. The comparative genomic analysis indicates that tandem duplication plus whole-genome duplication event, creating an important evolutionary resource for the diosgenin saponin biosynthetic pathway. In addition, the genomes of *D. tokoro* [8] and *D. dumetorum* [9] were also assembled and annotated to offer a large genetic data resource for their biology research. Apart from the nuclear genome, a number of chloroplast genome sequences were produced to identify the phylogenetic placement among multiple Dioscorea species [10]. Although there are no available reference genomes for some

important species, such as *D. nipponica* and *D. polystachya*, a significant amount of transcriptome sequencing datasets have been generated to explore gene function in various biological processes. In *D. nipponica*, RNA from different organs and methyl jasmonate treated rhizomes were sequenced to investigate dioscin biosynthesis [11]. To learn the molecular regulatory mechanism during microtuber formation of *D. polystachya*, RNAs from different development stages of microtuber were sequenced to reveal key pathways and hormone activities [12].

Despite the substantial significance of omics data, a critical gap has been identified in this field: yam omics data covering genomes, transcriptomes, plastomes and variomes are widely scattered across numerous databases and individual research publications. A centralized integration platform for these heterogeneous datasets has not yet been established, and a tool capable of systematically annotating, visualizing and analyzing data within a yam-specific context is also lacking.

To address this urgent need, here we present YamOmics (Fig. 1), a web-based retrieval and analysis platform, covering genomic, transcriptomic and plastomic data from 41 distinct yam species, together with a large number of genomic variants from 935 yam accessions, and gene expression profiles from 191 yam samples. The database boasts three key features. First, it encompasses a comprehensive gene annotation, covering aspects like domain, family, pathway, and collinearity. To ensure responsible data interpretation, we have implemented a rigorous evidence hierarchy for downstream analyses. Specifically, computational outputs such as BLAST search and Expression analysis are assigned confidence levels based on strict filtering parameters (e.g., E-value) to help users prioritize high-confidence candidates. Second, it features an array of sophisticated visualization tools that enable users to intuitively grasp the retrieval results. Finally, it is equipped with a suite of online tools designed to enhance the user experience and facilitate efficient database utilization, including BLAST search, PCR primer design, CRISPR RNA design, expression analysis, enrichment analysis, and phylogenetic inference among Dioscorea accessions.

Material and method

Genome synteny

We collected the genome assemblies and annotation data for five yam species including *D. alata* (GCA_020875875.1), *D. rotundata* (GCF_009730915.1), *D. tokoro* (GCA_902716375.1), *D. dumetorum* (GCA_902712375.1), and *D. zingiberensis* (GCA_026586065.1)—from the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/datasets/genome). The other four genome assemblies, except for *D. dumetorum*, were assembled to the chromosome level. Because chromosome-level assemblies are essential for identifying interchromosomal syntenic blocks and comparing homologous segment arrangements across species, the four assemblies were used for genomic collinearity analysis of pairwise combinations. The BLASTP program was used for identification of homologous genes between two species with an E-value threshold of 1e − 5, followed by filtering for ≥ 30% sequence identity and ≥ 50% alignment coverage. The output of BLASTP was fed as evidence for MCScanX-2020-0-23 [13] (with default parameters) to identify two genomic collinearity blocks.

Plastome phylogeny

A total of 41 complete chloroplast genome assemblies of Dioscorea were retrieved from the CGIR database (https://ngdc.cncb.ac.cn/cgir). To construct the reference plastome phylogeny used in YamOmics, all plastome sequences were globally aligned using MAFFT v7.429 with the—auto strategy and default gap penalties [14]. The resulting multiple sequence alignment was used for maximum-likelihood (ML) phylogenetic inference with PHYLIP v3.696, and branch support was evaluated with 1000 bootstrap replicates [15]. To enable online phylogenetic placement in YamOmics, we implemented a rapid workflow that integrates MAFFT v7.429 and FastTree v2.1. This module allows users to upload either whole-chloroplast-genome sequences or plastid marker-gene sequences, align them against the reference plastome dataset, and quickly infer their placement within the Dioscorea plastome phylogeny [16].

Transcriptome assembly and annotation

To obtain more genetic resources, we used Trinity-2.15.1 [17] to *de novo* assemble transcriptomes for 14 Dioscorea species that have no genome assembly available (Table S1). TransDecoder-2.1.0 was used to identify ORFs in the assembled transcripts with the assistance of Pfam database. The transcript with longest ORF was selected as representative of an assembled gene locus. BUSCO/5.3.0 [18] was used to estimate transcriptome assembly completeness. Orthologous groups were inferred using OrthoFinder (v2.3.3) on the predicted proteomes of all Dioscorea species, with DIAMOND (v2.1.8) employed for the all-vs-all protein similarity searches and all other parameters kept at their default settings.

Variant calling

As for the variome, we collected resequencing datasets from *D. rotundata* and *D. alata* to call single nucleotide polymorphisms (SNPs) and short insertions and deletions (INDELs) (Table S2). Fastp v0.23.0 [19] was used to conduct quality control for the raw WGS datasets. The clean reads were aligned to the reference genomes using BWA-MEM v0.7.17 [20], followed by duplicate marking with MarkDuplicates and base quality score recalibration (BQSR) following the GATK Best Practices pipeline. Variants were called using HaplotypeCaller in GVCF mode, and joint genotyping across all accessions for each species was performed using GenotypeGVCFs in GATK v4.3.0 [21]. Multi-allelic variants were removed, retaining only biallelic SNPs and INDELs for downstream filtering. Hard filtering thresholds recommended by GATK were then applied to remove low-confidence variants. For SNPs, the criteria were QD < 2, QUAL < 30, FS > 60, MQ < 40, MQRankSum < − 12.5, and ReadPosRankSum < − 8. For INDELs, the thresholds included QD < 2, QUAL < 50, FS > 100, and ReadPosRankSum < − 20. For *D. rotundata*, which has sufficient sequencing depth and accession number, we retained biallelic variants with MAF ≥ 0.05 and genotype missing rate ≤ 0.3 for database construction. Because *D. alata* has fewer accessions and lower sequencing depth, a more lenient criterion (mean depth ≥ 5 and MAF ≥ 0.05) was adopted to maximize the number of reliable variants. We also performed the Hardy–Weinberg equilibrium (HWE) test to remove SNPs with HWE P < 1e−5. The final high-confidence variants were annotated using snpEff [22].

PPI prediction

We employed an interolog-based approach to predict protein–protein interactions, following the workflow described in our previous study [23]. Briefly, we first retrieved all experimentally validated interacting protein pairs from the STRING database [24]. We then aligned the proteomes of each plant species against the STRING proteins using BLASTP, with an E-value cutoff of 1e − 20, a minimum query coverage of 80%, and a minimum sequence identity of 20%. If two proteins within a given proteome mapped to two different proteins that are known to interact in STRING, we inferred that these two proteins form a putative interaction (interolog). Finally, we calculated an interaction score for each predicted pair using the following formula:

$$\text{Score} = \frac{1}{2} \times \left(\frac{Scov_1 + Qcov_1}{2} \times id_1 + \frac{Scov_2 + Qcov_2}{2} \times id_2 \right)$$

where Scov_1 denotes the percentage of the alignment length relative to the first protein in the known interacting pair, and Qcov_1 represents the percentage of the alignment length relative to the query protein. id_1 is the sequence identity of the first alignment. Similarly, Scov_2, Qcov_2, and id_2 describe the coverage and identity of the second alignment. A higher score indicates a higher likelihood of interaction, whereas a lower score suggests a lower probability of interaction.

Gene annotation

We conducted various data mining and annotations to improve the usability of the database. First, primer-3-2.6.1 [25] was used to design PCR primers for each gene CDS sequence. Second, InterProScan-5.2.1 [26] was used to identify functional domains in gene protein sequences. Third, TFBSTools [27] was used to detect transcription factor binding sites using the PFMs from the JASPAR [28] database. Finally, BLASTP was used to align all gene proteins to the Arabidopsis proteome (Araport11), enabling the retrieval of their corresponding gene names and family classifications.

Expression analysis

Differential expression (DE) analysis in YamOmics was implemented based on DESeq2 (v1.42.0) [29] and edgeR (v4.0.2) [30]. For each RNA-seq sample, raw gene-level read counts were precomputed using a unified pipeline and stored in the database. In the web interface, users first select a dataset/tissue and define two comparison groups by choosing samples based on metadata annotations. The selected raw count matrix is then used as input to DESeq2 or edgeR to identify differentially expressed genes, and statistical significance is reported as *P*-values with multiple-testing correction. Users can customize the DEG calling criteria by adjusting thresholds for log2 fold change and *P*-value (or adjusted *P*-value/FDR), and the output includes a ranked DEG table with log2 fold change, significance values, and gene annotations. For co-expression analysis, TPM expression matrices were quantile-normalized across samples using between-sample quantile (QNT) normalization. Pairwise gene co-expression was estimated using Pearson correlation coefficients, and users can filter gene–gene associations by specifying correlation and *P*-value cutoffs to construct co-expression networks. Functional enrichment analysis of gene sets was performed using clusterProfiler (v3.16.1) [31], and enriched terms were reported with corrected significance values.

CRISPR design

A custom Perl script was used to enumerate candidate SpCas9 gRNAs (20-nt protospacers with NGG PAM) within the user-specified region. Genome-wide off-target loci were then identified using BatMis v3.00 (≤ 3 mismatches, NGG PAM on the reference genome), and the resulting on- and off-target sequences were scored with the crisproff pipeline (v1.1.1) to obtain off-target scores and an overall CRISPRspec specificity score for each gRNA [32].

Web implementation

YamOmics is a web application, and all data are stored and managed using MySQL v5.1.71, a free relational database management system. On the server side, PHP v5.4.16, Perl, and R v4.0.3 are used for data processing. For the front-end, HTML, CSS, JavaScript, and JQuery are used to design the user interface. JavaScript plug-ins, including DataTables and ECharts, are incorporated to display tabular data and interactive visualizations. JBrowse v1.12.1, a fast and scalable genome browser, is embedded to enable users to explore genes within their genomic context.

Database content and usage

Database content

The database encompasses a comprehensive range of omics data from over 40 *Dioscorea* species, including genomic, transcriptomic, plastomic and variomic data (Fig. 2). Utilizing the 4 genome assemblies from *D. alata*, *D. rotundata*, *D. zingiberensis and D. tokoro*, we successfully identified a total of 6303 collinear blocks through comparative genomic analysis between different species. Furthermore, we achieved *de novo* transcriptome assembly for 14 species, such as *D. polystachya*, *D. abyssinica*, *D. cirrhosa*, *D. cochleariapiculata*, *D. communis*, *D. composita*, *D. japonica*, *D. nipponica*, *D. polygonoides*, *D. praehensilis*, *D.* soso, *D. sylvatica*, *D.* trifida, and *D. villosa*, resulting in the identification of 1,062,791 genes. These genes, along with 149,061 genes annotated from the genome, have been divided into 80,629 orthologous groups (Table S3). We also compiled 191 RNA-seq datasets, enabling detection of approximately 17.9 million expression levels for both *de novo* assembled and genome-annotated genes within these RNA-seq samples. Our prediction efforts led to the identification of around 10.4 million interolog-based protein-protein interactions. From 935 germplasm accessions of *D. alata* and *D.*

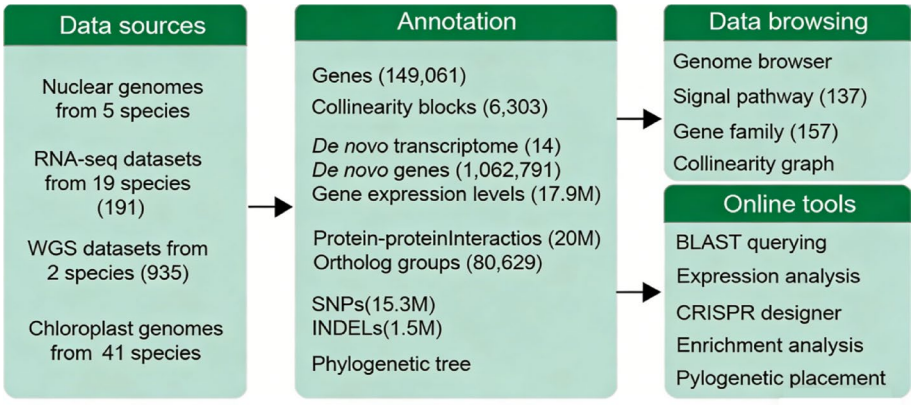

Fig. 2 Summary of YamOmics contents and functions

rotundata, we called a total of 15.3 million SNPs and 1.5 million INDELs. Lastly, we have gathered the complete chloroplast genome assemblies of 41 Dioscorea species, which served as a basis for constructing a phylogenetic tree.

Database construction

YamOmics, our comprehensive database, is readily accessible at https://biotec.njau.edu.cn/yamdb. It features a user-friendly navigation menu at the top of every page. The database's functionalities are organized into three main categories: (i) 'Omics browsers' offering an immersive experience in browsing genome, transcriptome, plastome and variome data. (ii) 'Gene exploration' supporting exploration of genes in the signaling pathways, gene family, genome collinearity and BLAST search capabilities. (iii) 'Analysis tools' equipped with a suite of online analysis tools, such as gene differential expression/coexpression analysis, CRISPR gRNA design, gene enrichment analysis, and phylogenetic placement identification within the Dioscorea species.

Omics browsers

Through the navigation menu or function button in the home page, users can conveniently open omics page to browse related information. At the top of the genome page, there is a drop-down list box for users switching to other species. The first section displays summary information of a genome assembly, such as assembly accession number, genome size, N50, and more. Then, there a genetic linkage map showing gene density in chromosomes (Fig. 3A) and JBrowse for users to explore genes in genomic context (Fig. 3B). Next, there is a table displaying statistical information for genes in each chromosome (Fig. 3C). Clicking on the view button, user can load genes of interests into an interactive table.

The transcriptome page has similar layout to the genome page. First, the initial section in the transcriptome page display summary information of a transcriptome assembly. Then, a pie chart (Fig. 3D) exhibits the integrity of the transcriptome assembly, and a bar chart (Fig. 3E) shows the length distribution of gene protein sequences. For *de novo* assembled transcriptomes, genes can't be summarized by chromosomes; thus, they are categorized by orthologous groups and listed in an interactive table. There, a bar chart shows the size of an orthologous group across different species (Fig. 3F). The red bar

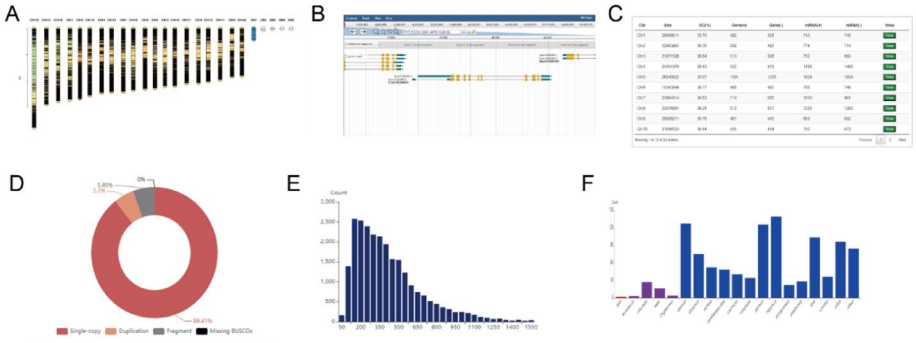

Fig. 3 Genome and Transcriptome Visualization. **A** Gene distribution density in chromosomes. **B** Exploration of genes in genomic context. **C** Gene table with several features. **D** Completeness estimation for *de novo* assembled transcriptome. **E** Protein sequence distribution of *de novo* assembled transcriptome. **F** Orthologous group size in genomic and transcriptomic proteomes

represents the current species, the purple bars indicate species with genome available, and the blue bars indicate species with transcriptome available.

The chloroplast genome page also has a drop-down list box for users to select one of 41 species of interest. The first section is a circular genome structure diagram showing gene position (Fig. 4A). Then a circular tree diagram displays the position of a species of interest in the Dioscorea phylogenetic tree (Fig. 4B). Finally, there is an interactive table showing genes and SSRs in the chloroplast genome.

In the variome page, there are 3 pie charts showing the proportion of SNP and INDEL variants (Fig. 4C), the proportion of variants in different gene feature regions (Fig. 4D), the proportion of exonic variant effects (Fig. 4E). The following is a heatmap showing variant density across all chromosomes (Fig. 4F). Variants can be queried by genome region and the retrieved variants are listed in an interactive table. By clicking on a variant ID, users can open a page showing its details, including allele frequency (Fig. 4G), genotype frequency (Fig. 4H) and genotyping missing frequency (Fig. 4I). At the page bottom, there is a table showing the genotype of the current variant site in all germplasm accessions. In addition, the table also provides hyperlinks for users to access information of accessions and study projects.

Gene exploration

To facilitate gene exploring, the collected genes can be accessed by signaling pathways, gene family, genome collinearity and BLAST search. On the signaling pathways page, there is a bar diagram showing gene distribution in 5 types of pathways including cellular processes (red), metabolism (purple), environmental information processing (blue), genetic information processing (green) and organismal systems (orange), and 18

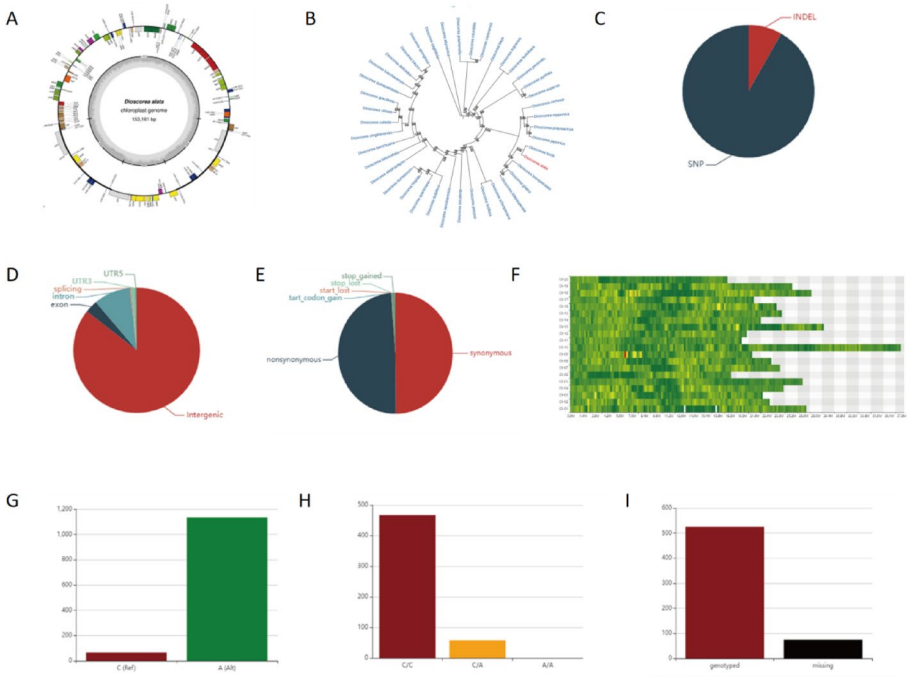

Fig. 4 Chloroplast genome and variome visualization. **A** Gene distribution in chloroplast genome. **B** Phylogenetic tree of 41 *Dioscorea* species. **C** Proportion of SNP and INDEL variants. **D** Proportion of variants in different gene feature regions. **E** Proportion of exonic variant effect. **F** Variant distribution in chromosomes. **G** Allele frequency of a variant site. **H** Genotype frequency of a variant site. **I** Missing frequency in genotyping of a variant site

subcategories of pathways (Fig. 5A). Through the interactive bar chart, users can quickly open a page to show a pathway of interest. In the pathway, the nodes with mapping genes are marked in red (Fig. 5B). Hovering the mouse over a node displays a message box showing gene IDs. By clicking on gene ID, users can open a page to display gene detail.

On the gene family page, there is a table exhibiting the current species possessed gene families with family size and subfamily number (Fig. 5C). If users select a family, a bar chart appears below, showing its all subfamilies (Fig. 5D). By clicking on a subfamily bar, its gene members will be listed in a following table with hyperlinks to gene page.

In addition to pathway and gene family, genes can be explored by genome collinearity. In the genome collinearity page, it allows users to select any two genomes to list genes in their collinear blocks. In the first section of the page, an interactive bar chart is used to show the proportion of genes in collinear blocks (Fig. 5E). Then, a circle mapping graph of collinear blocks between Dioscorea genomes is present to help users to filter out genes in a specified collinear block (Fig. 5F). Clicking on a connection line, the genes in the corresponding block will be listed in a table below. Furthermore, the database is also equipped with an online BLAST program (Fig. 5G), which allows users to search homologous genes by sequence similarity.

Gene page

Whether starting from the omics interfaces or the gene interfaces, users can open a page to show gene details (Fig. 6). At the top of the gene page, there is a table exhibiting

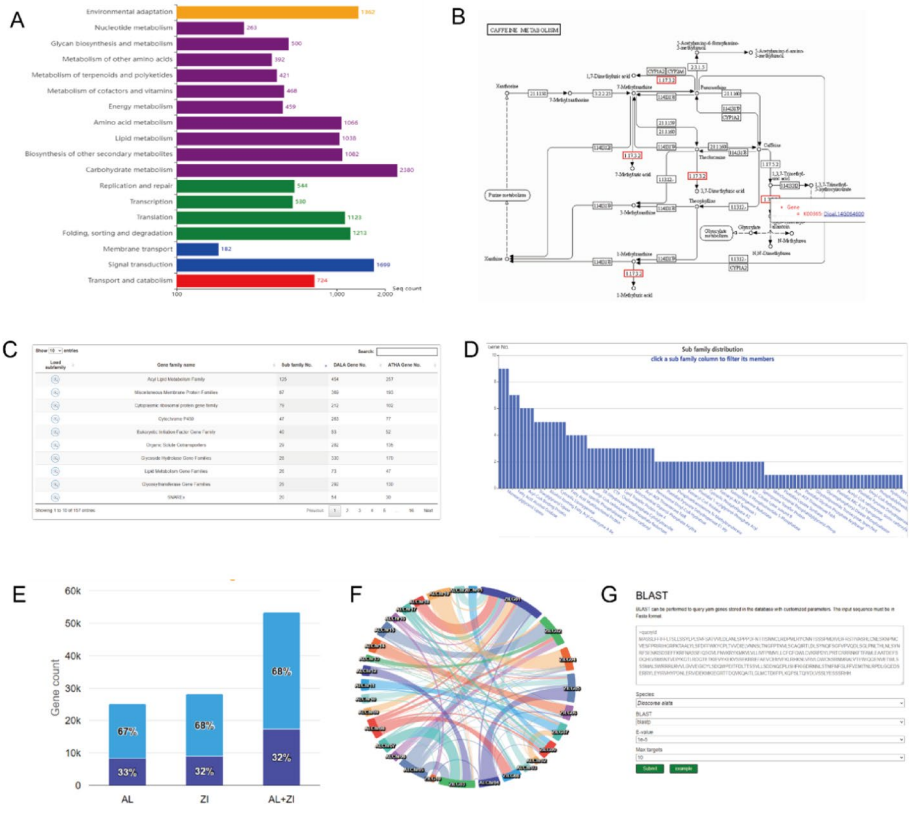

Fig. 5 Exploring genes by multiple approaches. **A** Gene distribution in different types of signaling pathways. **B** Visual chart of caffeine metabolism pathway. **C** Gene family list with family size and subfamily numbers. **D** Subfamily distribution graph of a specified family. **E** Collinear gene proportions between Dioscorea genomes. **F** Mapping graph of collinear blocks between Dioscorea genomes. **G** BLAST search interface

gene details, including gene ID, species, genome position, annotation information, and hyperlinks to download cDNA, CDS, and protein sequences (Fig. 6A). The following section include five panels named gene model, TFBS, primer, domain, expression, and interaction, respectively. The genome model panel displays the gene's intron-exon structure (Fig. 6B). The TFBS panel offers a table to list the predicted transcription factor binding sites (Fig. 6C). The primer panel provides a table showing the recommended qPCR primers with related parameters such as temperature, stability and more (Fig. 6D). The domain panel contains a graph showing domain topology in the protein sequence and a table displaying domain details (Fig. 6E). The expression panel contains a boxplot exhibiting gene expression levels in the collected RNA-seq samples (Fig. 6F). Finally, the interaction panel also contains a bar diagram and a table exhibiting the predicted protein-protein interactions (Fig. 6G).

Analysis tools

YamOmics is equipped with four online analysis tools to support yam research. In the expression analysis page, users can freely select samples and specify parameters to conduct differential gene expression and coexpression analysis (Fig. 7A). Differentially expressed genes are listed in an interactive table with log2FC, *P*-value, and more (Fig. 7B). By clicking on the explore button, users can further find genes that have a coexpression profile with the current gene. In addition to the coexpression genes listed in a table below (Fig. 7C), the expression profiles are visualized in a curve diagram (Fig. 7D).

To help users investigate gene function, the database contains an extremely user-friendly CRISPR/Cas9 sgRNA design tool. By simply specifying a genomic region,

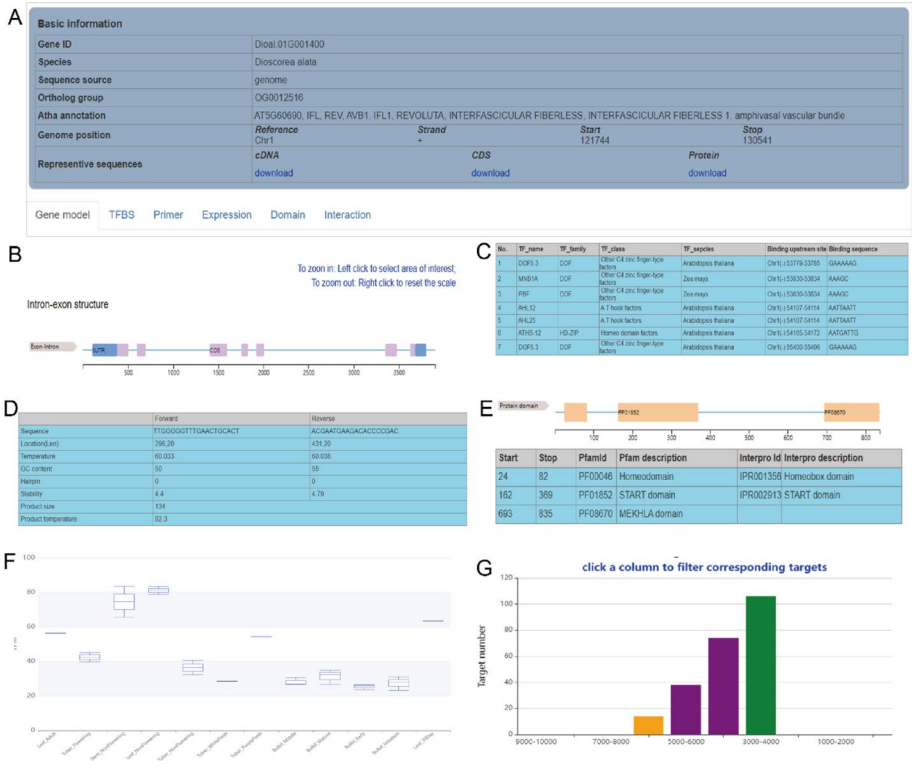

Fig. 6 Display of gene details. **A** Summary of gene information. **B** Gene structure including UTR, CDS, and intron. **C** Transcription factor binding site table. **D** qPCR primer table. **E** Domain topology of protein sequence. **F** Boxplot showing gene expression levels. **G** Bar chart showing the distribution of predicted interaction scores

Fig. 7 Online analysis tools. **A** Expression analysis interfaces for users to specify parameters and select samples. **B** Differentially expressed gene list. **C** Co-expression gene table. **D** Gene expression profiles in the collected RNA-seq samples. **E** CRISPR/Cas9 sgRNA table with specificity scores. **F** Table estimating off-target scores of a sgRNA. **G** Enrichment analysis results table. **H** A scatter plot to visualize enrichment analysis results. **I** Phylogenetic placement identification interfaces for users to paste DNA marker sequences or upload a chloroplast genome sequence file. **J** Phylogenetic tree showing the position of a querying sequence among 41 Dioscorea species

including the chromosome number and interval coordinates, users can acquire sgRNAs tailored for editing that specific area (Fig. 7E). Additionally, the tool offers valuable insights into the editing efficiency and potential off-target effects (Fig. 7F).

Enrichment analysis is used to study the overall function of a gene set, which is commonly based on gene ontology terms or pathways. If users upload a set of gene IDs, the tool can return a table showing enrichment analysis results with significance levels and gene ratios (Fig. 7G). In addition, the scatter plot below is used to visualize the analysis results (Fig. 7H).

The phylogenetic placement identification tool has been meticulously crafted to assist users in discerning the phylogenetic relationships among Dioscorea accessions. To achieve this goal, users can paste marker gene DNA sequences from the chloroplast genome including rbcL, matK, psbA, and more, or upload a complete chloroplast

genome sequence file (Fig. 7I). The phylogenetic placement tool can construct a phylogenetic tree online, in which the queried sequence/genome is marked in red (Fig. 7J).

Conclusion and future directions

In summary, we have constructed the YamOmics database, which serves as a central portal for Dioscorea species and provides comprehensive genome, transcriptome, plastome, and variome resources. The rich data resource and the diverse retrieval, visualization, and analysis functions make it a one-stop resource for yam researchers and breeders. We explicitly distinguish between primary genomic researchers and secondary users, such as breeders, Primary users can leverage cor e tools such as genome synteny analysis, ortholog inference, and plastome phylogenetics to explore fundamental biology; secondary users can utilize practical tools including CRISPR guide design, variant exploration, and expression analysis for breeding-related research. Compared with existing plant multi-omics databases (such as PLAZA and Ensembl Plants), YamOmics is specifically tailored for Dioscorea and offers several distinctive features, including plastome-aware, yam-specific variant exploration and an integrated CRISPR design toolkit customized for yam genomes. While the platform integrates heterogeneous datasets, we emphasize that computationally predicted outputs, such as CRISPR gRNA candidates and functional annotations based on de novo assemblies, should be viewed as exploratory hypotheses that require experimental validation, such as Sanger sequencing, before practical application. It directly accelerates basic research, such as the mining of stress-resistance genes, and supports the development of high-quality yam varieties, thereby addressing key challenges in yam improvement and industrial development. As for future directions, the database will be continuously updated as more data, including phenotypic and epigenetic data, become available. The platform's long-term utility is secured through versioned snapshots deposited in persistent repositories to ensure research reproducibility.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-026-06393-4.

Supplementary Material 1

Acknowledgments
Not applicable.

Author contributions
J.L. designed and managed the project. Y.Z., X.Y., J.C. and L.Y. collected and conducted data analysis. Y.Z., X.Y., J.C. and D.S. constructed the database. L.Y., J.C., Y.Z. and D.D. participated in discussions. J.L., and D.D. wrote the manuscript.

Funding
This study was supported from the China Agriculture Research System (CARS-21), the National Natural Science Foundation of China (32270208, and 32230089) and the Fundamental Research Funds for the Central Universities (KYCXJC2025005 and KYCXJC2024002).

Data availability
All data are available at: https://biotec.njau.edu.cn/yamdb. The specific accession numbers and sources of public datasets have been provided in supplementart-table.xlsx, including Table S1 for RNA-seq data, Table S2 for WGS data and Table S3 for Orthologous group distribution. These public datasets are derived from the Sequence Read Archive (SRA). The database will be regularly updated and released as versioned snapshots with a persistent Zenodo DOI: https://doi.org/10.5281/zenodo.17796473.

Declarations

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 11 September 2025 / Accepted: 29 January 2026
Published online: 08 February 2026

References

1. Epping J, Laibach N. An underutilized orphan tuber crop-Chinese yam: a review. Planta. 2020;252:58.
2. Marker RE, Tsukamoto T. Diosgenin1. J Am Chem Soc. 1940;62:2525–32.
3. Zhai C, Lu Q, Chen X, Peng Y, Chen L, Du S. Molecularly imprinted layer-coated silica nanoparticles toward highly selective separation of active Diosgenin from Dioscorea Nipponica Makino. J Chromatogr A. 2009;1216(12):2254–62.
4. Chen Y, Tariq H, Shen D, Liu J, Dou D. Omics technologies accelerating research progress in Yams. Vegetable Res. 2024;4:e014.
5. Tamiru M, Natsume S, Takagi H, White B, Yaegashi H, Shimizu M, Yoshida K, Uemura A, Oikawa K, Abe A et al. Genome sequencing of the staple food crop white Guinea Yam enables the development of a molecular marker for sex determination. Bmc Biol. 2017;15:86.
6. Sugihara Y, Darkwa K, Yaegashi H, Natsume S, Shimizu M, Abe A, Hirabuchi A, Ito K, Oikawa K, Tamiru-Oli M, et al. Genome analyses reveal the hybrid origin of the staple crop white Guinea Yam. P Natl Acad Sci USA. 2020;117(50):31987–92.
7. Li Y, Tan C, Li Z, Guo J, Li S, Chen X, Wang C, Dai X, Yang H, Song W, et al. The genome of Dioscorea zingiberensis sheds light on the biosynthesis, origin and evolution of the medicinally important Diosgenin saponins. Hortic Res. 2022;9:uhac165.
8. Natsume S, Yaegashi H, Sugihara Y, Abe A, Shimizu M, Oikawa K, White B, Kudoh A, Terauchi R. Whole genome sequencing of a wild yam species Dioscorea tokoro reveals a genomic region associated with sex. *bioRxiv* 2022.
9. Siadjeu C, Pucker B, Viehover P, Albach DC, Weisshaar B. High contiguity de Novo genome sequence assembly of trifoliate Yam (*Dioscorea dumetorum*) using long read sequencing. Genes. 2020;11(3):274.
10. Zhao Z, Wang X, Yu Y, Yuan S, Jiang D, Zhang Y, Zhang T, Zhong W, Yuan Q, Huang L. Complete chloroplast genome sequences of dioscorea: characterization, genomic resources, and phylogenetic analyses. PeerJ. 2018;6:e6032.
11. Sun W, Wang B, Yang J, Wang W, Liu A, Leng L, Xiang L, Song C, Chen S. Weighted gene co-expression network analysis of the Dioscin rich medicinal plant Dioscorea Nipponica. Front Plant Sci. 2017;8:789.
12. Li J, Zhao X, Dong Y, Li S, Yuan J, Li C, Zhang X, Li M. Transcriptome analysis reveals key pathways and hormone activities involved in early microtuber formation of Dioscorea opposita. Biomed Res Int. 2020;2020:8057929.
13. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49.
14. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res. 2019;47(W1):W5–10.
15. Shimada MK, Nishida T. A modification of the PHYLIP program: A solution for the redundant cluster problem, and an implementation of an automatic bootstrapping on trees inferred from original data. Mol Phylogenet Evol. 2017;109:409–14.
16. Chu G, Warnow T. SCAMPP+FastTree: improving scalability for likelihood-based phylogenetic placement. Bioinform Adv. 2023;3(1):vbad008.
17. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–U130.
18. Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–54.
19. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
21. De Summa S, Malerba G, Pinto R, Mori A, Mijatovic V, Tommasi S. GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics. 2017;18(Suppl 5):119.
22. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.
23. Yin ZY, Liu JD, Dou DL. RLKdb: A comprehensively curated database of plant receptor-like kinase families. Mol Plant. 2024;17(4):513–5.
24. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638–46.
25. Koressaar T, Lepamets M, Kaplinski L, Raime K, Andreson R, Remm M. Primer3_masker: integrating masking of template sequence with primer design software. Bioinformatics. 2018;34(11):1937–8.
26. Jones P, Binns D, Chang HY, Fraser M, Li WZ, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.
27. Tan G, Lenhard B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics. 2016;32(10):1555–6.
28. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K, Kumar V, Lemma RB, Lucas J, Cheneby J, Baranasic D, et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2024;52(D1):D174–82.

29. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
30. Robinson MD, McCarthy DJ, Smyth GK. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
31. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. ClusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innov. 2021;2(3):100141.
32. Alkan F, Wenzel A, Anthon C, Havgaard JH, Gorodkin J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. Genome Biol. 2018;19(1):177.

Publisher's note