















SapBase: A central portal for functional and comparative genomics of Sapindaceae species

Jiawei Li^{1,2,3†} , Chengjie Chen^{1,2,3†} , Zaohai Zeng^{1,2,3} , Fengqi Wu^{1,2,3} , Junting Feng^{1,2,3} , Bo Liu^{1,2,3} , Yingxiao Mai^{1,2,3} , Xinyi Chu¹ , Wanchun Wei^{1,2,3} , Xin Li³ , Yanyang Liang^{1,2,3} , YuanLong Liu^{1,2,3} , Jing Xu^{1,2,3}  and Rui Xia^{1,2,3*} 

1. State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Horticulture, South China Agricultural University, Guangzhou 510640, China

2. Guangdong Laboratory for Lingnan Modern Agriculture, South China Agricultural University, Guangzhou 510640, China

3. Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in South China at the Ministry of Agriculture and Rural Affairs, South China Agricultural University, College of Horticulture, Guangzhou 510640, China

[†]Jiawei Li and Chengjie Chen contributed equally to this work.

*Correspondence: Rui Xia (rxia@scau.edu.cn)



Jiawei Li



Rui Xia

genomic data center capable of storing, sharing, and analyzing these data. Here, we introduced SapBase, that is, the Sapindaceae Genome Database. SapBase houses seven published plant genomes alongside their corresponding gene structure and functional annotations, small RNA annotations, gene expression profiles, gene pathways, and synteny block information. It offers user-friendly features for gene information mining, co-expression analysis, and inter-species comparative genomic analysis. Furthermore, we showcased SapBase's extensive capacities through a detailed bioinformatic analysis of a *MYB* gene in litchi. Thus, SapBase could serve as an integrative genomic resource and analysis platform for the scientific exploration of Sapindaceae species and their comparative studies with other plants.

Keywords: comparative genomics, database, SapBase, Sapindaceae

Li, J., Chen, C., Zeng, Z., Wu, F., Feng, J., Liu, B., Mai, Y., Chu, X., Wei, W., Li, X., et al. (2024). SapBase: A central portal for functional and comparative genomics of Sapindaceae species. *J. Integr. Plant Biol.* **00**: 1–10.

ABSTRACT

The Sapindaceae family, encompassing a wide range of plant forms such as herbs, vines, shrubs, and trees, is widely distributed across tropical and subtropical regions. This family includes economically important crops like litchi, longan, rambutan, and ackee. With the wide application of genomic technologies in recent years, several Sapindaceae plant genomes have been decoded, leading to an accumulation of substantial omics data in this field. This surge in data highlights the pressing need for a unified

INTRODUCTION

The Sapindaceae family, also known as the soapberry family, comprises 141 genera and approximately 1,900 species (Acevedo-Rodríguez et al., 2010). These species are predominantly found in tropical and subtropical regions and manifest as trees, shrubs, as well as woody or herbaceous vines. The family exhibits diverse reproductive features, with some species being dioecious and others

monoecious. The economic importance of many Sapindaceae species is notable, with some providing delicious fruits such as litchi (*Litchi chinensis*), longan (*Dimocarpus longan*), rambutan (*Nephelium lappaceum*), and ackee (*Blighia sapida*)—the national fruit of Jamaica, and others valued for their abundant secondary metabolites, such as saponins from soapberry (*Sapindus mukorossi*) and seed oil from yellowhorn (*Xanthoceras sorbifolium*). Additionally, some species, including maple (*Acer spp.*) and buckeye

(*Aesculus glabra*) are valued for their timber while others like balloon-vine (*Cardiospermum halicacabum*) are recognized for their medicinal properties.

The last decade has seen a surge in next-generation sequencing (NGS) and genomic technologies, leading to the sequencing of full genomes for several Sapindaceae plants (Lin et al., 2017; Liang et al., 2019; Yang et al., 2019; Zhang et al., 2021; Hu et al., 2022; Xue et al., 2022). Notably, our recent publication on the litchi genome has also garnered significant attention (Edger, 2022; Hu et al., 2022; Lyu, 2022). These mark the advent of the post-genome era for the Sapindaceae family. Despite these advancements, there remains a lack of a public genomic database for any Sapindaceae species, let alone an integrative database for the entire family. Addressing this gap, we have combined our in-house NGS data with all available public data for seven Sapindaceae plants to create the **Sapindaceae Genome DataBase** (i.e., SapBase) (www.sapindaceae.com). This platform aims to furnish comprehensive genomics resources and offer a powerful online analytic tool for the scientific exploration of Sapindaceae species and their comparative studies with other plants.

RESULTS

Overview of SapBase

Currently, SapBase hosts genomic resources for seven Sapindaceae species, including litchi, longan, rambutan, soapberry, balloon-vine, acer, and yellowhorn (Figure 1B). The database encompasses 16 genomes with complete sequences, 411 resequencing genomic data sets totaling 4.82 TB, and 919 RNA-seq data sets (10.3 TB) from 49 projects, and 501 small RNA (sRNA) loci sourced from the sRNAanno database (Chen et al., 2021). Altogether, SapBase contains data on 514,422 genes comprising 893,747 transcripts, with functional annotations provided for 501,479 of these genes. Furthermore, 4,577 functional domains have been annotated for 392,123 genes. SapBase

has also mapped 79,862,416 interaction relations among 145,248 proteins and identified 89,025 synteny blocks across all Sapindaceae species, covering 134,016 genes. Additionally, 486 gene co-expression modules have been identified through integrative analyses of these omics data. Access to all these resources is facilitated through SapBase's four primary function categories (Figure 1C), that is, **BROWSE** for data and result browsing, **SEARCH** for comprehensive and efficient information retrieval, **ANALYSES** for various data processing, analysis and visualization, and **DOWNLOAD** for data deposit and download. With SapBase, users are empowered to conduct bio-sequence analyses, explore gene expression atlases, and engage in comparative genomic studies.

Bio-sequence analysis with SapBase

SapBase, as a multifunctional resource hub, provides users with a variety of search strategies (Figure 1C). Initiating with a mere gene identifier, users can easily learn its genomic position, functional annotation, gene structure annotation, domain architecture, and sequences (Figure 2A). It has integrated Basic Local Alignment Search Tool (BLAST), the most commonly used sequence search engine, for the quick nucleic acid or protein sequence comparison among species of interest. Furthermore, SapBase introduces a practical gene ID Convert function, enabling the alignment of genes from Sapindaceae species with their potential homologs in other well-studied plant species (Figure 2B), including Arabidopsis, rice, and tomato. To extend the search capacities of SapBase, we designed a sophisticated "Meta-Search" module, which offers a "Google-like" search function (Figure 2C), permitting users to conduct searches within SapBase using a single text field. This feature accepts various types of input, such as gene identifiers, descriptions of gene functions, genomic intervals, or even DNA/protein sequences. SapBase is designed to automatically recognize the type of input, execute a comprehensive search, and return the best matching results for the search terms.

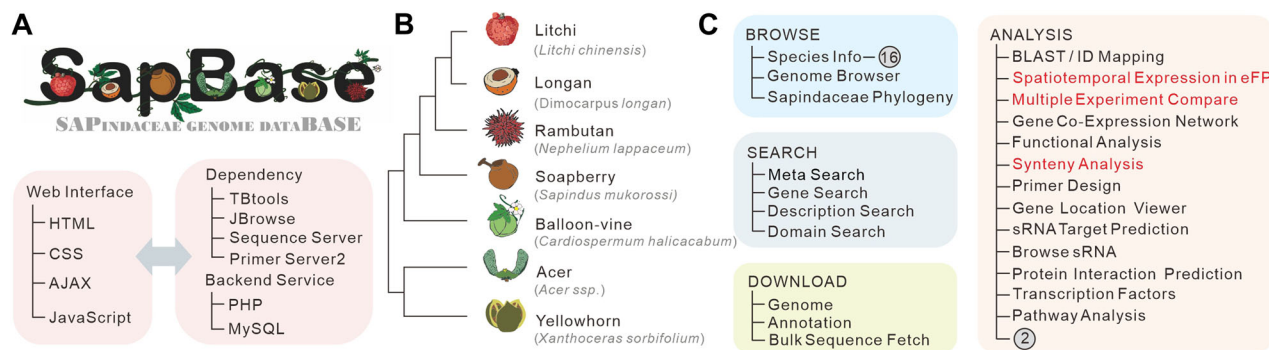


Figure 1. Overview of SapBase

(A) Internal structure of SapBase. (B) Phylogenetics relationship of all currently available species in SapBase. (C) Four primary functional categories of SapBase.

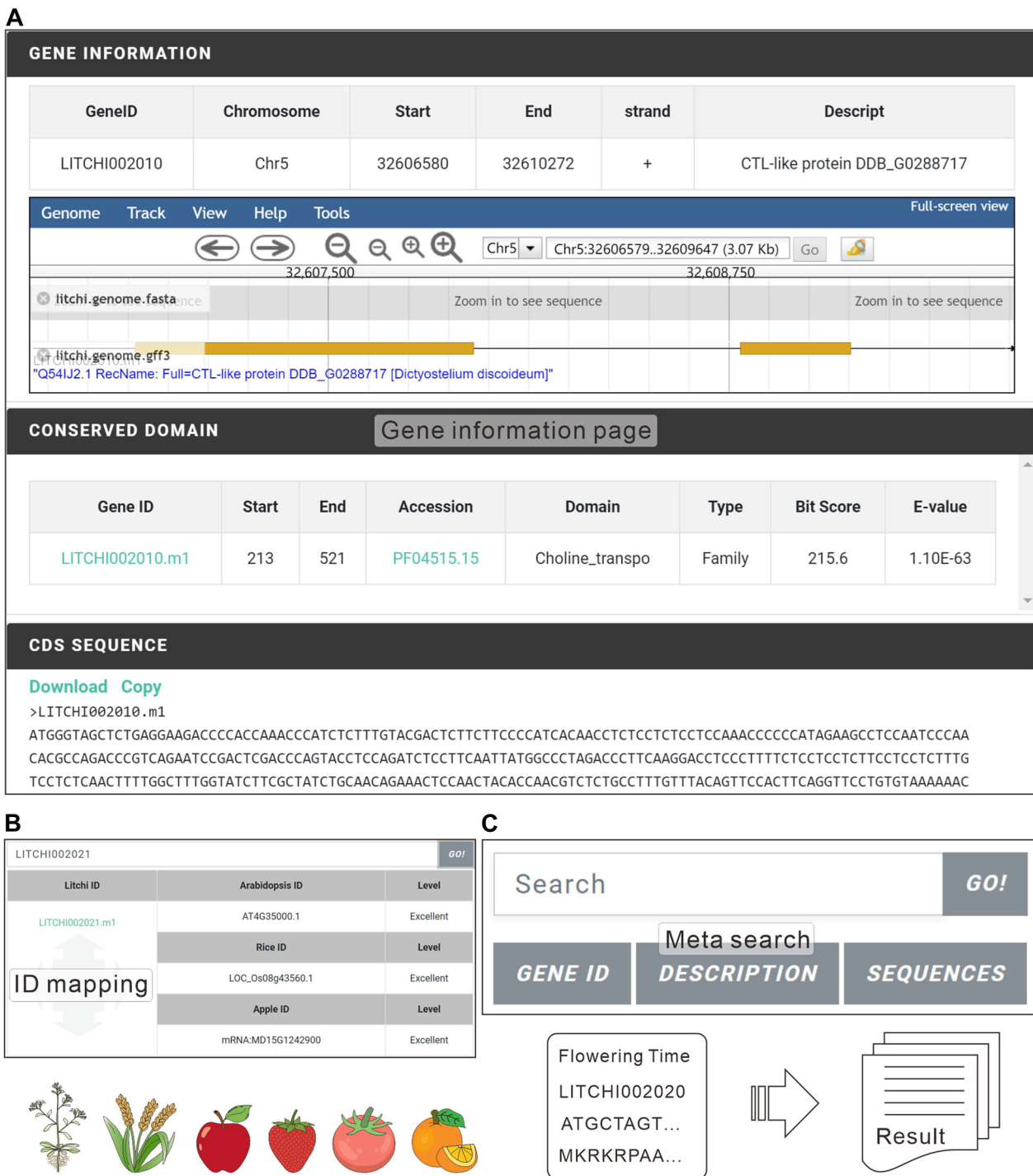


Figure 2. Sequence analysis function modules in SapBase

(A) A demo page of gene information description. (B) A sample outcome of the identifier conversion function within SapBase. (C) Illustration of the “Meta-Search” function, which permits diverse input content in a singular text field component for database exploration.

Expression data exploration

The functionality of a gene is intimately linked to its expression pattern. SapBase offers a diverse array of interfaces for the exploration of expression data (Figure 3). A notable feature, the Spatiotemporal Expression visualization using

electronic Fluorescent Pictograph (eFP), has been designed to enable intuitive visualization of gene expression in plants. eFP profiles have been meticulously constructed for all seven plants within the database (Figure 3A), allowing users to analyze the spatiotemporal distribution of gene expression by

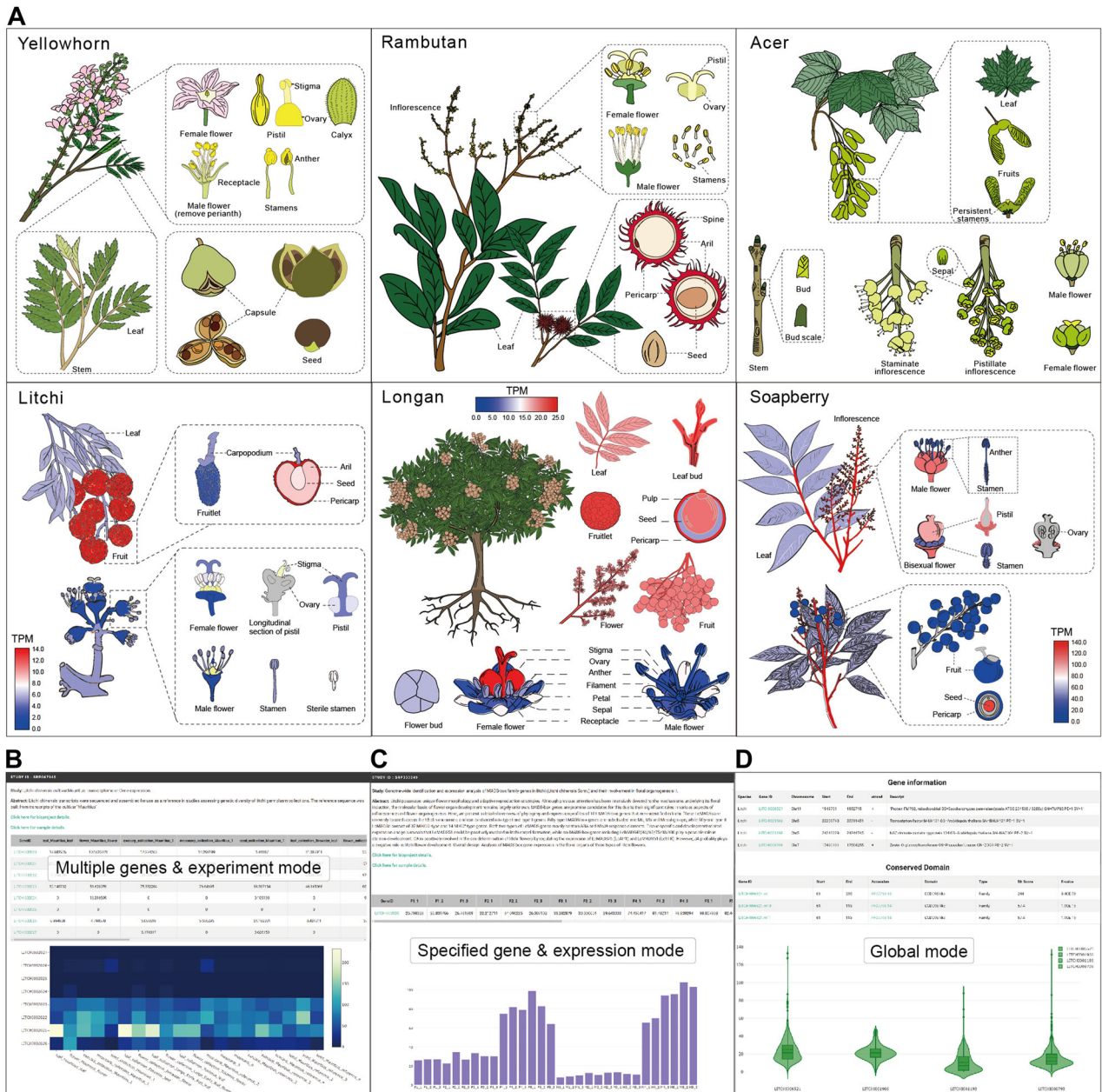


Figure 3. Modules for exploring expression data in SapBase

(A) Cartoon heatmaps available on SapBase for yellowhorn, litchi, rambutan, longan, acer and soapberry. (B) A demo result for an inquiry of expression profiles of multiple genes. (C) A demo result for an inquiry of expression profile of a single gene. (D) A demo result (violin plot) for an inquiry of comparative analysis of global gene expression in SapBase.

merely inputting a gene identifier. Additionally, SapBase has aggregated all publicly available RNA-seq datasets of Sapindaceae plants, facilitating quick inspection of gene expression patterns across various experiments or under different conditions. When users input multiple gene identifiers, a heatmap is generated to display the data (Figure 3B), while the input of a single gene identifier results in a barplot (Figure 3C). Moreover, users can also compare the global expression patterns of multiple genes using violin plots (Figure 3D). Furthermore, SapBase incorporates a gene

co-expression function. This function leverages expression profiles to construct comprehensive co-expression gene networks, enabling users to easily identify genes that are co-expressed and interconnected with a gene of interest simply by entering its identifier.

Comparative genomics functions

SapBase is designed to be a comprehensive data hub for all Sapindaceae species. As the genomic resources for more species become available, SapBase enables users to engage

in a wide array of comparative genomic analyses directly within the platform. For instance, by providing a list of gene identifiers, users can visualize and compare the location of genes across different species (Figure 4A). In addition, a synteny analysis module is available for rapid investigation into the evolution and diversification of large syntenic gene blocks (Figure 4B), which is invaluable for gene gain-or-loss analysis. SapBase also offers a function for inferring homologs, which allows users to efficiently identify the best homologous gene set for their genes of interest. This feature is more robust than a simple sequence BLAST search and could be used for gene clade contraction-or-expansion analysis (Figure 4C). Moreover, to facilitate comparative genomics analysis at the gene family level, SapBase incorporates OrthoVenn2 (Xu et al., 2019) to provide an interactive interface. This interface grants users access to homologous groups with species-specific gene presence or absence events (Figure 4D, E).

Extended modules

SapBase extends its functionality with a comprehensive suite of integrative data analysis pipelines and tools, including Gene Set Function Enrichment, Gene Pathway Analysis, and Protein Interaction Network modules, and sRNA Target Prediction and polymerase chain reaction Primer Design tools. Furthermore, SapBase offers access for downloading a wealth of Sapindaceae genomic data and resources, comprising raw sequencing data, complete genome sequences, gene annotations, and expression profiles. For those needing to obtain sequences for a large number of genes or specific chromosome regions in bulk, the “Bulk Data Fetch” feature efficiently facilitates this process. Moreover, SapBase includes a **RESOURCE** page, which is a well-curated compilation of state-of-the-art software, relevant genomic databases, and web servers, serving as a valuable resource for researchers.

Case study: Characterization of a MYB transcription factor related to anthocyanin metabolism

SapBase aims to establish a comprehensive one-stop genomic resource and analysis platform for Sapindaceae plants. To showcase the excellent capacities of SapBase, we delve into a case study focusing on the bioinformatic analysis of a single gene in litchi.

The distinct radiant red skin of litchi is one of its important economic traits (Ghosh, 2000). The red color of the skin is mainly attributed to its anthocyanin content (Rivera-López et al., 1999). The MYB family of transcription factors has been reported to play a crucial role in anthocyanin synthesis. In our previous study on the litchi genome, we identified a MYB gene, *LITCHI008189*, which is potentially involved in anthocyanin synthesis (Allan et al., 2008; Xu et al., 2015).

Utilizing SapBase, we start by querying the gene identifier, *LITCHI008189*, revealing its chromosome location to be 2115398 to 2119483 on Chr7 (Figure 5A). It is annotated as “MYB113,” a gene previously identified to directly regulate

anthocyanin synthesis in *Arabidopsis thaliana* (Gonzalez et al., 2008). The genome browser module reveals that the gene has three exons with moderate lengths of both 5' and 3' untranslated regions, indicating a comprehensive annotation of this gene (Figure 5A), which is further corroborated by RNA-seq data (Figure 5F). The presence of two Myb_DNA-binding domains confirms it as an *R2R3-MYB* gene (Figure 5B), a common yet crucial MYB type in plants known for regulating anthocyanin synthesis (Stracke et al., 2007; Naing and Kim, 2018).

Next, a pivotal question would be the expression profile (i.e., where and extent) of *LITCHI008189*. Through SapBase, spatiotemporal gene expression can be visually plotted as heatmaps (i.e., eFP Graph). Notably, *LITCHI008189* is preferentially expressed in mature litchi fruit skin (Figure 5C), aligning with high anthocyanin content regions. Transcriptome analysis of litchi skin at different developmental stages (SRP047115) indicates a positive correlation between *LITCHI008189* expression and color transition from green to red (Figure 5D). The gene co-expression module in SapBase further identifies genes related to *LITCHI008189*, facilitating in the unraveling of downstream gene regulation pathways.

Additionally, SapBase also supports evolutionary gene analysis. Through the Evolution Path function, users can compare syntenic gene blocks across related species. Within this module, *LITCHI008189* is found to reside within a MYB gene cluster comprising four members (Figure 5E). Of note, different species have varying member counts in this MYB gene cluster across species. For example, rambutan and litchi both have four MYB members, while longan has only two, and yellowhorn has only one. Interestingly, the skin of rambutan and litchi is red, while the skin of longan and yellowhorn is yellow-white, hinting a potential link between gene cluster composition and fruit skin color, which is worth further exploration.

Further examination in the genome browser of SapBase revealed intriguing details within this gene cluster (Figure 5F). Specifically, *LITCHI008189* exhibits significant RNA-seq read coverage within the pericarp transcriptome, while *LITCHI008190* shows a significant RNA-seq read coverage within the stamen transcriptome. It is generally believed that the existence of the same gene family members in one cluster is caused by tandem repeat events. This pattern suggests that while members of the same family are grouped together due to tandem duplication events, the different expression patterns of *LITCHI008189* and *LITCHI008190* imply their involvement in distinct biological processes, despite their similar molecular functions. Thus, these new findings provide rationales for designing experiments to further validate the findings.

CONCLUSION

Through the aggregation of publicly available genomes and omics data for seven Sapindaceae species, we have

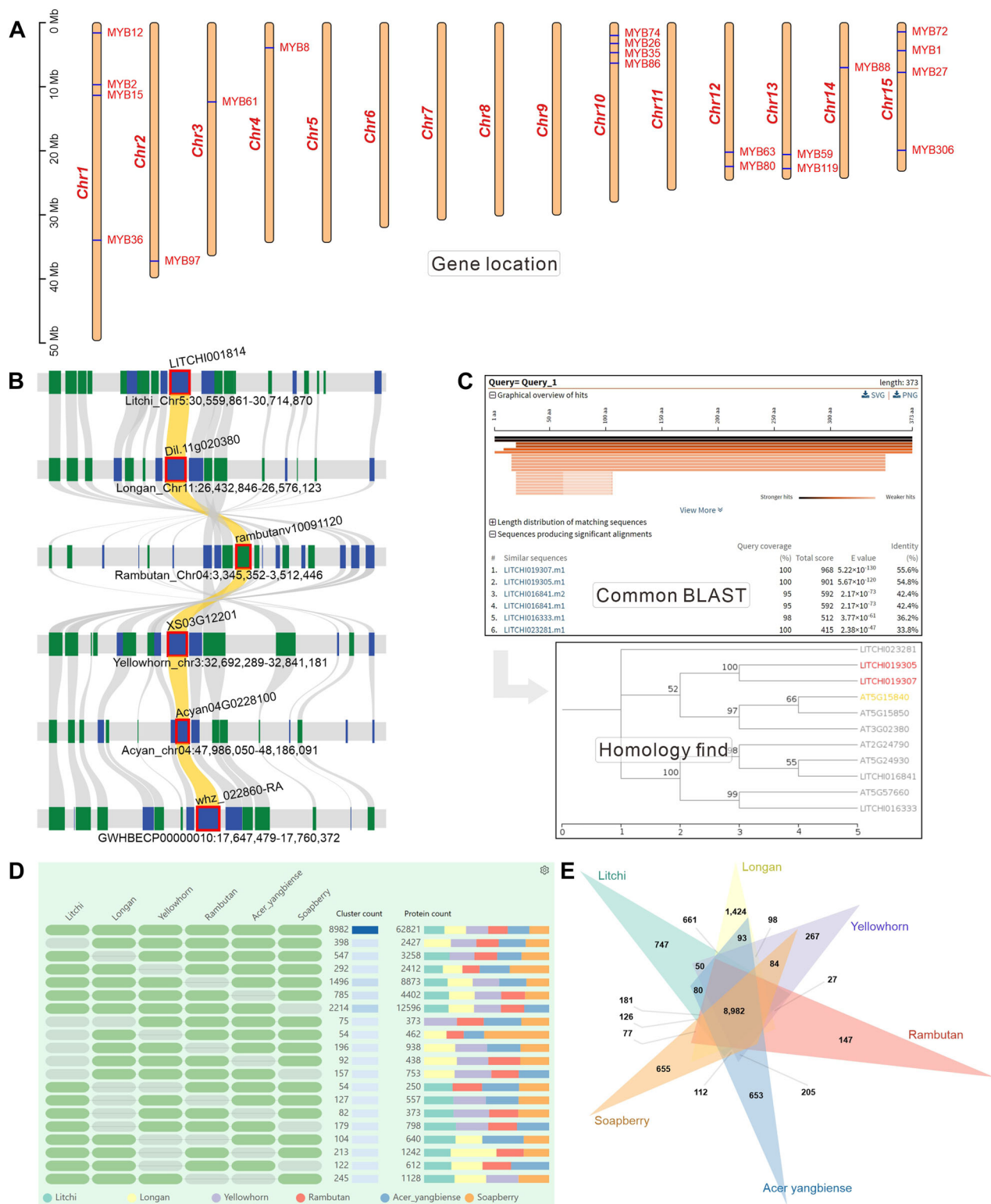


Figure 4. Comparative genomics modules in SapBase

(A) Visualization of position of multiple genes across different chromosomes. (B) A sample output for evolutionary conserved block analysis in SapBase. The target gene is highlighted with red boundary and the correlation relationships are represented as brazier curves. (C) Comparison of result generated from Basic Local Alignment Search Tool (BLAST) and the Homology Find function. (D, E) Overview of conservation status of all gene families in Sapindaceae species, analyzed and visualized by OrthoVenn2. All components are interactive.



Figure 5. An illustrative analysis outcome of a MYB gene within SapBase

(A) The basic sequence information of *LITCHI008189*. (B) The conserved domain annotation of *LITCHI008189* with red symbolizing high expression levels and blue signifying lower expression levels. (C) A cartoon heatmap of *LITCHI008189* across different RNA-seq experiments and its co-expression network. (D) Expression profiles of *LITCHI008189* across different RNA-seq experiments and its co-expression network. (E) The evolutionary conserved block encompassing *LITCHI008189* leading to the identification of a MYB cluster. (F) A genome browser snapshot of *LITCHI008189*.

established SapBase, a comprehensive one-stop resource and analysis platform for the Sapindaceae family. The database ensures convenient and effective access to a wide array of genomic resources, facilitating daily endeavors of researchers in the field. Envisioned as a dynamic, long-term project, SapBase is committed to ongoing maintenance and updates, solidifying its role as a pivotal data hub and analytic platform for the research community focused on Sapindaceae or related fields.

DATABASE IMPLEMENTATION

Genomic resource collection

We downloaded reference genome sequences and associated annotation information for each Sapindaceae species following links in previous studies (Lin et al., 2017; Liang et al., 2019; Yang et al., 2019; Zhang et al., 2021; Hu et al., 2022; Xue et al., 2022). TBtools software (Chen, Wu, et al., 2023) was employed to extract CDS, protein sequences and promoter sequences. sRNA annotation information was directly synchronized from the sRNAanno database, which was previously established by us (Chen et al., 2021). All conserved domains in protein sequences were predicted using a stand-alone pfam_scan pipeline (Mistry et al., 2021).

Acquire and analyze RNA-seq, sRNA-seq and resequencing data

We downloaded and aggregated the raw resequencing data, RNA-seq data and sRNA-seq data of Sapindaceae species from the Sequence Read Archive and sRNAanno databases (Chen et al., 2021) by implementing a robust workflow for data processing. This included mapping all reads to Sapindaceae reference genomes using Burrows–Wheeler Aligner (Li and Durbin, 2009), and sorting the mapped reads according to genomic coordinates using Samtools (Li and Handsaker, Wysoker, et al., 2009). Reads from different Illumina lanes were merged using “samtools merge.” Subsequently, duplicates were removed with Picard (<https://github.com/broadinstitute/picard>), and call of individual-specific gvcf was performed with Genome Analysis Toolkit (GATK) (Poplin et al., 2017). Finally, joint calling of single nucleotide polymorphisms (SNPs) was performed with GenotypeGVCFs. Following the quality control of SNPs with bcftools (Danecek et al., 2021), the SNPs underwent stringent filtering using GATK VariantFiltration with the following settings: DP < 300 || DP > 3000 || QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0, finally obtaining high-quality SNPs in Variant Call Format (VCF) files. For RNA-seq data, following initial quality control with FastQC (Andrews, 2010) and adapter removal by Fastp (Chen, 2023), all reads were mapped to the corresponding reference genome using Spliced Transcripts Alignment to a Reference (STAR) (Dobin et al., 2013). Finally, the expression level of transcripts was normalized to transcripts per million by StringTie (Pertea et al., 2015). For sRNA-seq data, we used sRNAmminer (Li et al., 2023) to remove adapters. Clean reads of at least 15 nt in

length were retained for further analysis. Potential small RNA and phased secondary (PHAS) loci were then identified using sRNAmminer with default parameters.

Synteny analysis, co-expression analysis and protein interaction prediction

We employed MCScanX for synteny analysis across Sapindaceae species, and the results were further optimized using an in-house script (Wang et al., 2012). For each experimental dataset comprising at least eight samples, co-expressed gene modules were identified using the weight gene co-expression network analysis (WGCNA) method (Langfelder and Horvath, 2008). Based on a reciprocal BLAST approach, we established precise gene mapping relationships for Sapindaceae species and a range of extensively studied plants. Leveraging these mapping relationships between Sapindaceae plants and *A. thaliana*, we transferred all *A. thaliana* protein interaction information downloaded from the String database (Szklarczyk et al., 2023) to each Sapindaceae plant.

Overall function implementation

SapBase uses Nginx as the back-end support, with MySQL serving as the robust storage solution. The platform's front-end graphics interface is crafted using programming languages such as PHP, HTML, and JavaScript. The overall framework of the website was constructed using the common JavaScript library, bootstrap.js. To ensure an optimal BLAST search experience, we have integrated SequenceServer (Priyam et al., 2019) in our platform. PrimerServer2 (<https://github.com/billzt/PrimerServer2>) is incorporated as a Primer Design module. The rest of the analysis and visualization modules in SapBase (such as cartoon heatmap, synteny block visualization, gene set enrichment analyses and sRNA targeting analyses, etc.) were implemented using the TBtools software developed by us (Chen et al., 2023).

ACKNOWLEDGEMENTS

This work is supported by the Key Area Research and Development Program of Guangdong Province (2022B0202070003 and 2021B0707010004). This work is also supported by the National Science Foundation of China (#32072547 and #32102320) and the open competition program of top 10 critical priorities of Agricultural Science and Technology Innovation for the 14th Five-Year Plan of Guangdong Province (2022SDZG05). We thank all members of XIALAB and the National Litchi and Longan Industrial Technology Consortium of China for their suggestions and testing of SapBase. We are also grateful for the efforts of other researchers who have been devoted to the genomic research of Sapindaceae plants. We express our gratitude to all the SapBase users for their support. We also thank Guiyang Watch Biotechnology for their technical advice on the development of SapBase “meta-search” module.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

R.X. conceived the project; R.X., J.L. and C.C. designed the database. J.L. performed all the coding of the website. Z.Z., F.W., J.F., B.L., Y.M., X.C., W.W., X.L., Y.L. (Liu), Y.L. (Liang) and J.X. tested the functions. R.X., J.L. and C.C. prepared the figures and wrote the manuscript. All authors read and approved the final manuscript.

Edited by: Long Mao, Institute of Crop Sciences, CAAS, China

Received Feb. 11, 2024; **Accepted** Apr. 23, 2024

REFERENCES

- Acevedo-Rodríguez, P., Van Welzen, P., Adema, F., Van Der Ham, R.** (2010). *Flowering plants. Eudicots: Sapindales, cucurbitales, myrtaceae* (Springer), pp. 357–407.
- Allan, A.C., Hellens, R.P., and Laing, W.A.** (2008). MYB transcription factors that colour our fruit. *Trends Plant Sci.* **13**: 99–102.
- Andrews, S.** (2010). FastQC: A quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Chen, C., Li, J., Feng, J., Liu, B., Feng, L., Yu, X., Li, G., Zhai, J., Meyers, B.C., and Xia, R.** (2021). sRNAanno—A database repository of uniformly annotated small RNAs in plants. *Hortic. Res.* **8**: 45.
- Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., Liu, Y., Feng, J., Chen, H., He, Y., et al.** (2023). TBtools-II: A “One for All, All for One” bioinformatics platform for biological big-data mining. *Mol. Plant* **16**: 1733–1742.
- Chen, S.** (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**: e107.
- Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., and Davies, R.M.** (2021). Twelve years of SAMtools and BCFtools. *Gigascience* **10**: giab008.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R.** (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.
- Edger, P.P.** (2022). The power of chromosome-scale, haplotype-resolved genomes. *Mol. Plant* **15**: 393–395.
- Ghosh, S.** (2000). World trade in litchi: Past, present and future. *Acta Horticulturae* **558**: 23–30.
- Gonzalez, A., Zhao, M., Leavitt, J.M., and Lloyd, A.M.** (2008). Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. *Plant J.* **53**: 814–827.
- Hu, G., Feng, J., Xiang, X., Wang, J., Salojärvi, J., Liu, C., Wu, Z., Zhang, J., Liang, X., Jiang, Z., et al.** (2022). Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. *Nat. Genet.* **54**: 73–83.
- Langfelder, P., and Horvath, S.** (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **9**: 1–13.
- Li, G., Chen, C., Chen, P., Meyers, B.C., and Xia, R.** (2023). sRNAMiner: A multifunctional toolkit for next-generation sequencing small RNA data mining in plants. *Sci. Bulletin.* **69**: 784–791.
- Li, H., and Durbin, R.** (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P.** (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- Liang, Q., Li, H., Li, S., Yuan, F., Sun, J., Duan, Q., Li, Q., Zhang, R., Sang, Y.L., and Wang, N.** (2019). The genome assembly and annotation of yellowhorn (*Xanthoceras sorbifolium* Bunge). *GigaScience* **8**: giz071.
- Lin, Y., Min, J., Lai, R., Wu, Z., Chen, Y., Yu, L., Cheng, C., Jin, Y., Tian, Q., and Liu, Q.** (2017). Genome-wide sequencing of longan (*Dimocarpus longan* Lour.) provides insights into molecular basis of its polyphenol-rich characteristics. *GigaScience* **6**: gix023.
- Lyu, J.** (2022). Two domestication routes intersect. *Nat. Plants* **8**: 96.
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L., Tosatto, S.C., Paladin, L., Raj, S., and Richardson, L.J.** (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**: D412–D419.
- Naing, A.H., and Kim, C.K.** (2018). Roles of R2R3-MYB transcription factors in transcriptional regulation of anthocyanin biosynthesis in horticultural plants. *Plant Mol. Biol.* **98**: 1–18.
- Perteua, M., Perteua, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L.** (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotech.* **33**: 290–295.
- Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., and Roazen, D.** (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* **2017**: 201178.
- Priyam, A., Woodcroft, B.J., Rai, V., Moghul, I., Munagala, A., Ter, F., Chowdhary, H., Pieniak, I., Maynard, L.J., and Gibbins, M.A.** (2019). Sequenceserver: A modern graphical user interface for custom BLAST databases. *Mol. Biol. Evol.* **36**: 2922–2924.
- Rivera-López, J., Ordorica-Falomir, C., and Wesche-Ebeling, P.** (1999). Changes in anthocyanin concentration in Lychee (*Litchi chinensis* Sonn.) pericarp during maturation. *Food Chem.* **65**: 195–200.
- Stracke, R., Ishihara, H., Huep, G., Barsch, A., Mehrtens, F., Niehaus, K., and Weisshaar, B.** (2007). Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *Plant J.* **50**: 660–677.
- Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., and Pyysalo, S.** (2023). The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**: D638–D646.
- Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.-h., Jin, H., Marler, B., and Guo, H.** (2012). MScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**: e49.
- Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., Zhang, G., Gu, Y.Q., Coleman-Derr, D., and Xia, Q.** (2019). OrthoVenn2: A web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **47**: W52–W58.
- Xu, W., Dubos, C., and Lepiniec, L.** (2015). Transcriptional control of flavonoid biosynthesis by MYB–bHLH–WDR complexes. *Trends Plant Sci.* **20**: 176–185.
- Xue, T., Chen, D., Zhang, T., Chen, Y., Fan, H., Huang, Y., Zhong, Q., and Li, B.** (2022). Chromosome-scale assembly and population diversity analyses provide insights into the evolution of Sapindus mukorossi. *Hortic. Res.* **9**: uhac012.

Yang, J., Wariss, H.M., Tao, L., Zhang, R., Yun, Q., Hollingsworth, P., Dao, Z., Luo, G., Guo, H., and Ma, Y. (2019). De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan Province, China. *GigaScience* **8**: giz085.

Zhang, W., Lin, J., Li, J., Zheng, S., Zhang, X., Chen, S., Ma, X., Dong, F., Jia, H., and Xu, X. (2021). Rambutan genome revealed gene networks for spine formation and aril development. *Plant J.* **108**: 1037–1052.



Scan using WeChat with your smartphone to view JIPB online



Scan with iPhone or iPad to view JIPB on Twitter