

1 **SapBase (Sapindaceae Genomic DataBase): a** 2 **central portal for functional and comparative** 3 **genomics of Sapindaceae species**

4

5 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},
6 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
7 Xu^{1,2,3}, Rui Xia^{1,2,3*}

8

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

11 ²*Guangdong Laboratory for Lingnan Modern Agriculture, South China Agricultural University, Guangzhou,*
12 *Guangdong, 510640, China*

13 ³*Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in South China, Ministry of*
14 *Agriculture and Rural Affairs, South China Agricultural University, Guangzhou, Guangdong 510640, China*

15

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

17

18 Sapindaceae is a family of flowering plants, also known as the soapberry family, comprising 141
19 genera and about 1900 species (Pedro et al., 2010). Most of them are distributed in tropical and
20 subtropical regions, including trees, shrubs, also woody or herbaceous vines. Some are dioecious,
21 while others are monoecious. Many Sapindaceae species possess great economic value; some furnish
22 delicious fruits, like lychee (*Litchi chinensis*), longan (*Dimocarpus longan*), rambutan (*Nephelium*
23 *lappaceum*); and ackee (*Blighia sapida*) - the national fruit of Jamaica; some produce abundance
24 secondary metabolites, like saponin from soapberry (*Sapindus mukorossi*), and seed oil from
25 yellowhorn (*Xanthoceras sorbifolium*); some yield valuable timber including maple (*Acer spp.*) and
26 buckeye (*Aesculus glabra*); and some are of great herbal medicinal value, like balloon-vine

27 (*Cardiospermum halicacabum*).

28

29 In the last decade, with the rocketing of next generation sequencing (NGS) and genomic technologies,
30 the full genome sequences of several Sapindaceae plants have been resolved (Lin et al., 2017; Liang et
31 al., 2019; Yang et al., 2019; Zhang et al., 2021; Hu et al., 2022; Xue et al., 2022). Among them, our
32 recent publication of the lychee genome attracted broad attention (Edger, 2022; Hu et al., 2022; Lyu,
33 2022). Now the post-genome era arrives for Sapindaceae, however, there is no public genomic
34 database available for any Sapindaceae species, let alone an integrative database for the whole
35 Sapindaceae family. A unified data platform is in urgent need to collect, manage and share relevant
36 data resources. Therefore, we integrated our home-brew NGS data with all publicly available data for
37 seven Sapindaceae plants and constructed the **Sapinaceae Genomic DataBase**, named SapBase
38 (www.sapindaceae.com), in order to provide genomic resources and an online powerful analytic
39 platform for scientific research on Sapinaceae species and comparative studies with other plants.

40

41 **Data Source**

42 Currently, SapBase hosts genomic resources for seven Sapindaceae species (Fig. 1A), including 16
43 full genome sequences, 411 sets of resequencing genomic data (411 sets, 4.82 TB), 919 RNAseq data
44 (10.3 TB) from 49 projects, and 501 sRNA loci from the sRNAanno database (Chen et al., 2021). In
45 total, there are 514,422 genes (893,747 transcripts), with 501,479 of them having functional
46 annotations. 4,577 functional domains are annotated from 392,123 genes. SapBase also predicts
47 79,862,416 interaction relations between 145,248 proteins. 89,025 synteny blocks between every two
48 Sapindace species were identified covering 134,016 genes. Besides, 486 gene co-expression modules
49 were singled out by the integrative analyses of these omics data. All these data are accessible at
50 SapBase via the four major function categories (Fig. 1B): **BROWSE** for data and result browsing;
51 **SEARCH** for comprehensive and efficient information retrieval; **ANALYSES** providing various data
52 processing, analysis and visualization functions, and **DOWNLOAD** responsible for data deposit and
53 download.

54 **Search Strategy**

55 As a multifunctional resource hub, SapBase provides a set of search strategies. Starting from a simple
56 gene identifier, users can obtain its functional annotation, gene structure annotation, domain
57 annotation, and sequences. BLAST, the most commonly used sequence search engine, has been
58 integrated into SapBase for quick nucleic acid or protein sequence comparison among species of
59 interest. A practical ID Convert function is provided for mapping genes from Sapindaceae species to
60 their best homologs in most-researched plant species, such as *Arabidopsis*, rice, tomato, etc. To
61 maximize the search capacity of SapBase, we designed a sophisticated “Meta Search” module, which
62 provides a “Google-like” search function. Users are allowed to search using any related information,
63 such as gene identifiers, gene function descriptions and even DNA/protein sequences, and SapBase
64 will automatically identify the input content, carry out data search, and return the best matching
65 results.

66 **Data Analysis**

67 Aside from a datahub, we also aim to develop SapBase as a powerful analytic platform. Currently, a
68 great variety of analysis functions are available in SapBase (Fig.1B). Spatiotemporal Expression
69 shown in eFP (electronic Fluorescent Pictograph) is a feature function designed to intuitively visualize
70 gene expression in customized pictographic representations of plants. We have constructed eFP
71 profiles for all the seven plants (Fig. 1C). There are two other entries in the Gene Expression module:
72 Multiple Experiment Comparison and Co-expression. The former accommodates transcriptional
73 expression profiles generated from all publicly available RNAseq datasets of Sapindaceae plants (Fig.
74 1D), which can be used for quick assessment of the expression changes of ideal genes under different
75 experimental conditions. The Co-expression function, based on the expression profiles, is
76 implemented to integrate all co-expression gene networks, where users can easily obtain the
77 co-expressed and interconnected counterparts by simply entering the identifier of a gene of interest.

78

79 SapBase also provides functions for comparative genomic analyses. Synteny Analysis can be used to

80 quickly explore the evolution and diversification of large syntenic gene blocks (Fig. 1E), and
81 Homology Find function allows users to quickly obtain the optimal homologous gene set for the genes
82 of interest. In addition, a pack of other functions is incorporated in the SapBase, from various
83 integrative data analysis pipelines like Gene Function Enrichment, Gene Pathway Analysis, and
84 Protein Interaction Network, to out-of-the-box tools, including Gene Location Viewer, small RNA
85 Target Prediction, PCR Primer Design, etc.

86

87 **Others**

88 SapBase provides entries for downloading all Sapindaceae genomic data and resources, including raw
89 sequencing data, genome sequences, gene annotations, and expression matrices. Users can use the
90 “Bulk Data Fetch” to grab sequences for a large number of genes or chromosome regions in a batch
91 mode. A well-curated list of state-of-the-art software, related genomic databases, or web servers are
92 also recommended on the RESOURCE page.

93

94 **Conclusion**

95 By collating publicly released genomes and omics data for seven Sapindaceae species, we have
96 developed SapBase, which provides a one-stop-shop for all Sapindaceae genomic resources, ensuring
97 a convenient and efficient access and usage of all these resources for daily research. As a long-term
98 development project, SapBase will be continuously maintained and updated as a central datahub and
99 analytic platform for researchers working on Sapindaceae or related areas.

100 **Acknowledgment**

101 We thank all members of XIALAB and the National Litchi and Longan Industrial Technology
102 Consortium of China for their suggestions and testing of SapBase. We are also grateful for the effort
103 other researchers have devoted to the genomic research of Sapindaceae plants.

104

105 **References**

106

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.

Edger, P. P. (2022). The power of chromosome-scale, haplotype-resolved genomes. *Mol. Plant* **15**:393–395.

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.

Liang, Q., Li, H., Li, S., Yuan, F., Sun, J., Duan, Q., Li, Q., Zhang, R., Sang, Y. L., Wang, N., et al. (2019). The genome assembly and annotation of yellowhorn (*Xanthoceras sorbifolium* Bunge). *GigaScience* **8**:6.

Lin, Y., Min, J., Lai, R., Wu, Z., Chen, Y., Yu, L., Cheng, C., Jin, Y., Tian, Q., Liu, Q., et al. (2017). Genome-wide sequencing of longan (*Dimocarpus longan* Lour.) provides insights into molecular basis of its polyphenol-rich characteristics. *GigaScience* **6**:5.

Lyu, J. (2022). Two domestication routes intersect. *Nat. Plants* **8**:96.

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**.

Yang, J., Wariss, H. M., Tao, L., Zhang, R., Yun, Q., Hollingsworth, P., Dao, Z., Luo, G., Guo, H., Ma, Y., et al. (2019). De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan Province, China. *GigaScience* **8**:7.

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

107

108

109 Figure legend:

110 **Figure 1. Features of Sapinaceae Genomic DataBase (SapBase).** (A) List of current species with
111 genomic data hosted in SapBase. (B) Structure of main functions in SapBase. Numbers in circles show
112 denotes the species number or the number of other functions that cannot be shown here due to space
113 limit. Functions highlighted in red are the three selected functions used for demonstration. (C) eFP
114 (electronic Fluorescent Pictograph) used for gene expression presentation of six Sapinaceae species:
115 Yellowhorn, Rambutan, Acer, Litchi, Longan and Soapberry. The former three are in their original
116 form without displaying expression data while the latter three are in their heatmap form to show the
117 expression of certain genes. Red color denotes a high expression level of gene expression while blue
118 color corresponds to a low expression level. All eFP graphs in SapBase are interactive to view spatial
119 expression of genes. (D) A representative screenshot from the Multiple Experiment Compare function.
120 Expression levels of several genes are listed in a table and displayed in an interactive heatmap. (E) An
121 exemplary result generated from the Synteny Analysis function, using the gene (“LITCHI001814”)
122 as the input.

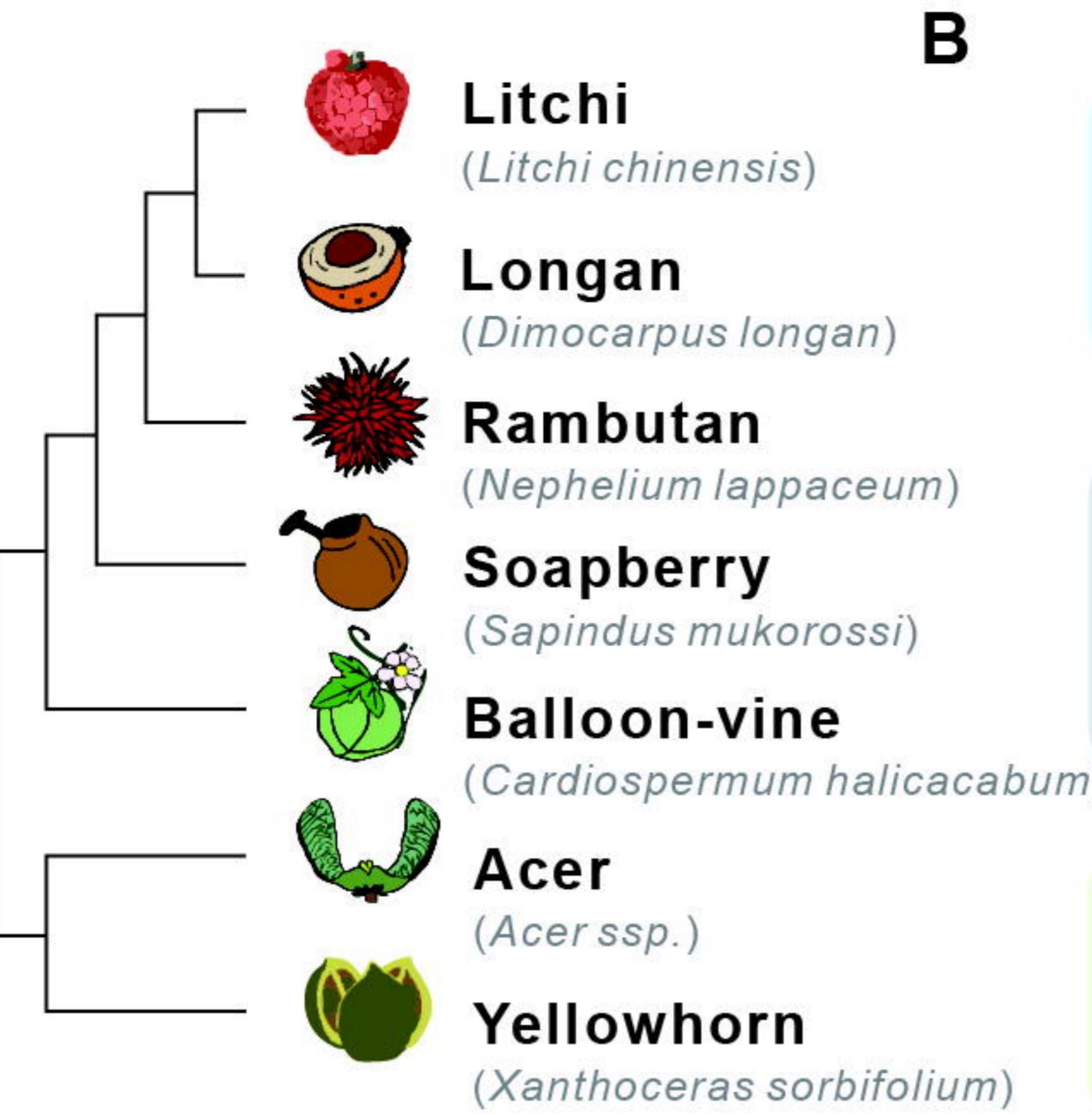
123

124

A

SapBase

SAPINDACEAE GENOME DATA BASE



B

BROWSE

- Species Info — 16
- Genome Browser
- Sapindaceae Phylogeny

SEARCH

- Meta Search
- Gene Search
- Description Search
- Domain Search

DOWNLOAD

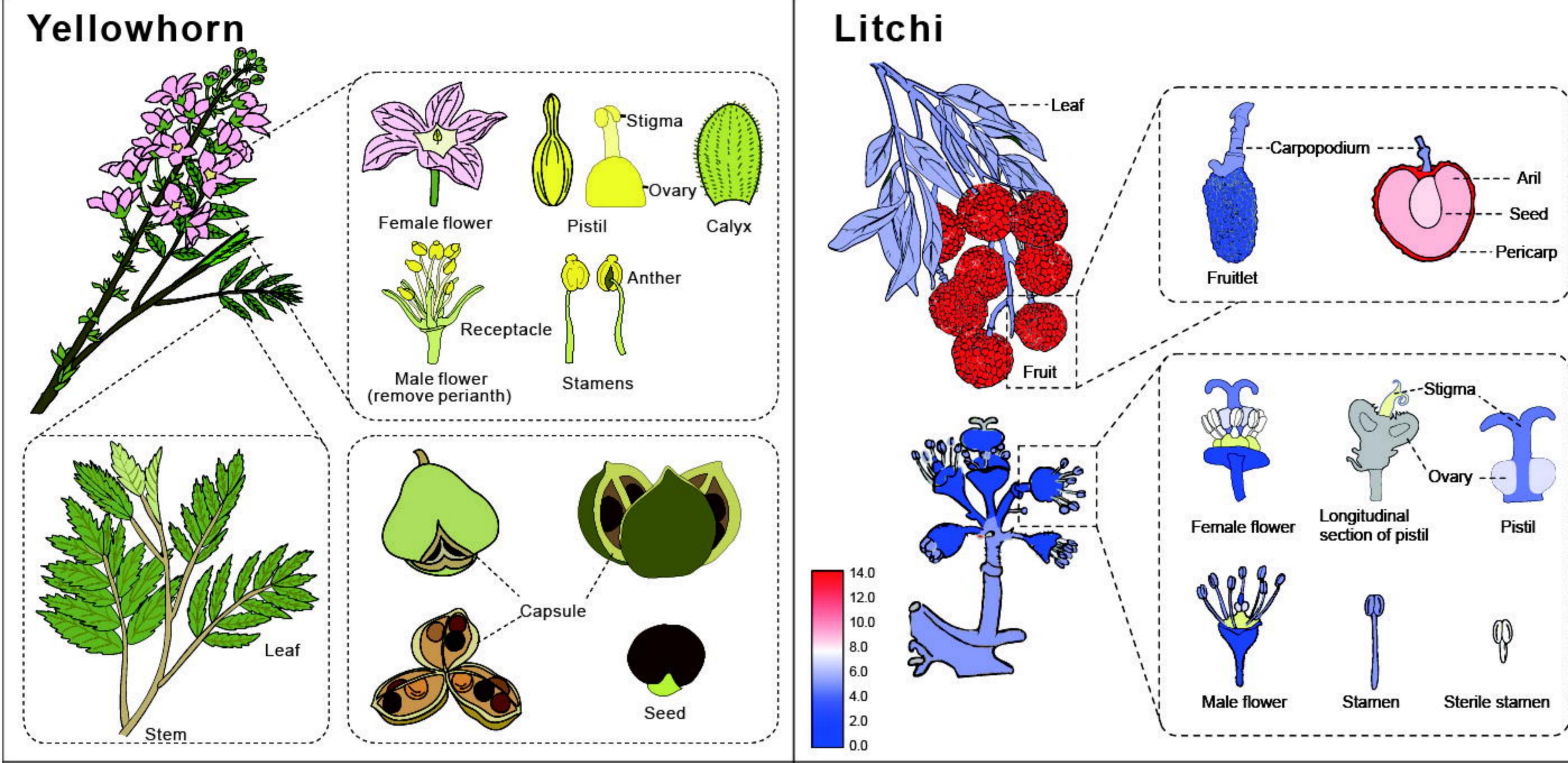
- Genome
- Annotation
- Bulk Sequence Fetch

ANALYSIS

- BLAST / ID Convert
- Spatiotemporal Expression in eFP
- Multiple Experiment Compare
- Gene Co-Expression Network
- Functional Analysis
- Synteny Analysis
- Primer Design
- Gene Location Viewer
- sRNA Target Prediction
- Browse sRNA
- Protein Interaction Prediction
- Transcription Factors
- Pathway Analysis

4

C



D

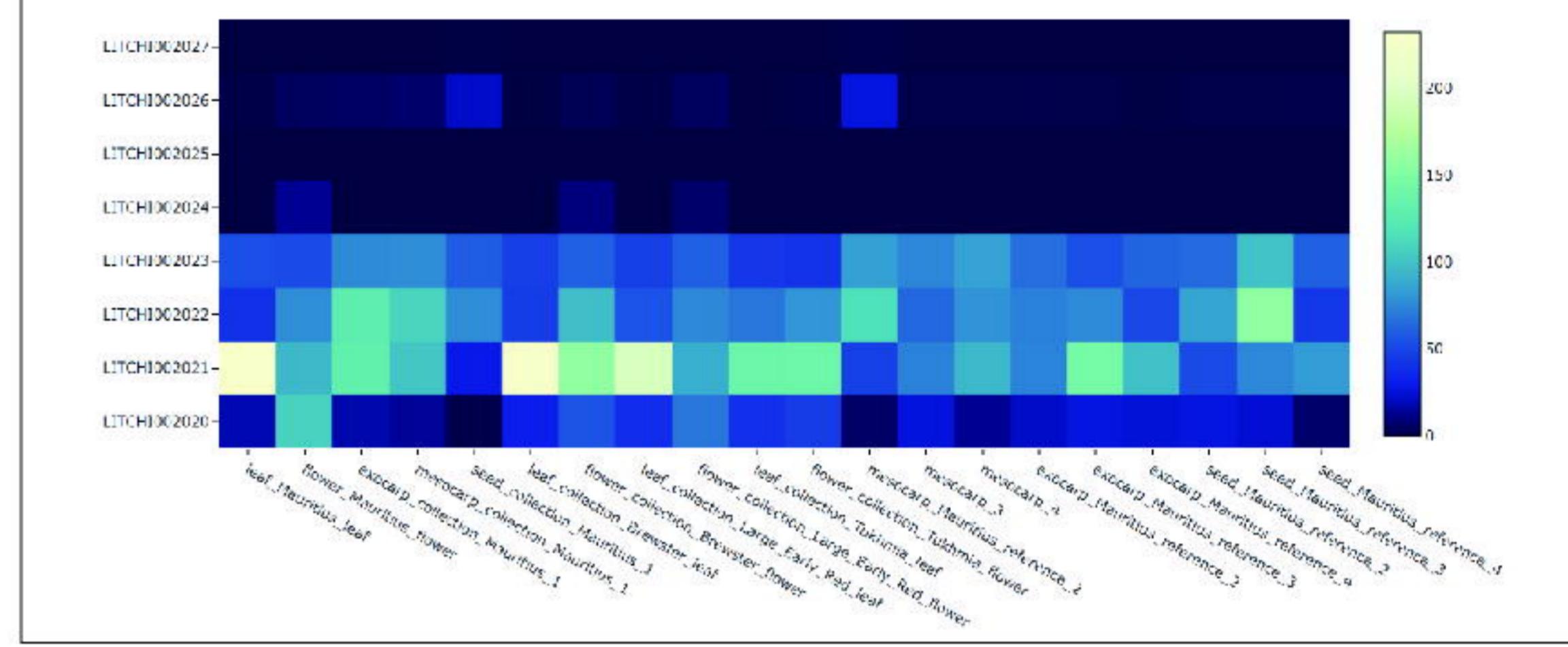
STUDY ID : SRP067945

Study: *Litchi chinensis* ssp. *Mauritius* transcriptome or Gene expressionAbstract: *Litchi chinensis* transcripts were sequenced and assembled for use as a reference in studies assessing genetic diversity of litchi germplasm collections. The reference sequence was built from transcripts of the cultivar 'Mauritius'.

Click here for bioproject details.

Click here for sample details.

GeneID	loc_Mauritius.fcf	flower_Mauritius.flower	exocarp_collection_Mauritius_I	mesocarp_collection_Mauritius_I	seed_collection_Mauritius_I	leaf_collection_Brewster_leaf	flower_collected
UT04002320	17.632976	07.623673	17.274255	14.230159	1.44387	31.87911	55
UT04002321	232.29782	95.000311	31.204416	11.176246	25.705095	23.099007	151
UT04002322	39.704836	77.251828	126.705769	109.161709	76.881186	45.669796	97
UT04002323	53.10732	5.628059	75.532234	76.64695	93.3.7181	16.165096	60
UT04002324	C	13.310559	0	0	0.0913	0	9
UT04002325	C	C	0	0	0	0	0
UT04002326	0.964308	4.744078	3.065318	5.598203	21.182331	0.424311	3.
UT04002327	C	C	0.170317	0	0.626153	0	1



E

