The bayberry database: a multiomic database for *Myrica rubra*, an important fruit tree with medicinal value

Haiying Ren[1*] , Yuanhao He[1,2], Xingjiang Qi[1], Xiliang Zheng[1], Shuwen Zhang[1], Zheping Yu[1] and Fengrong Hu[1]

Abstract

Background: Chinese bayberry (*Myrica rubra* Sieb. & Zucc.) is an important fruit tree in China, and has high medicinal value. At present, the genome, transcriptome and germplasm resources of bayberry have been reported. In order to make more convenient use of these data, the Bayberry Database was established.

Results: The Bayberry Database is a comprehensive and intuitive data platform for examining the diverse annotated genome and germplasm resources of this species. This database contains nine central functional domains to interact with multiomic data: home, genome, germplasm, markers, tools, map, expression, reference, and contact. All domains provide pathways to a variety of data types composed of a reference genome sequence, transcriptomic data, gene patterns, phenotypic data, fruit images of *Myrica rubra* varieties, gSSR data, gene maps with annotation and evolutionary analyses. The tools module includes BLAST search, keyword search, sequence fetch and enrichment analysis functions.

Conclusions: The web address of the database is as follows http://www.bayberrybase.cn/. The *Myrica rubra* database is an intelligent, interactive, and user-friendly system that enables researchers, breeders and horticultural personnel to browse, search and retrieve relevant and useful information and thus facilitate genomic research and breeding efforts concerning *Myrica rubra*. This database will be of great help to bayberry research and breeding in the future.

Keywords: Chinese bayberry, Database, Genome, Transcriptome, Tools, Germplasm resource

Background

The Myricaceae family consists of three genera, including approximately 50 species, and it is mainly distributed in warm and humid areas of Asia (e.g., India, Japan, and China), South America (e.g., Brazil), North America (e.g., the USA), Africa (e.g., Kenya), and Europe (e.g., Spain, Switzerland, Norway, and France) [1–4]. Chinese bayberry (*Myrica rubra* Sieb. & Zucc.) is a fruit tree with high economic value, with a cultivation area in China of approximately 334,000 ha and an annual yield

of approximately 950,000 t. In addition, many kinds of organs of Chinese bayberry are applied in traditional Chinese medicine. Chinese bayberry extracts contain antioxidants that may serve to alleviate health issues such as diarrhea, inflammation, and cancer [1, 5].

With rapid advances in genome technology, obtaining high-quality genome assemblies of most plant species is no longer difficult [6, 7]. Through genome sequencing, a 289.92 Mb genome for fine mapping was assembled with 26,325 predicted protein-coding genes [8, 9]. To elucidate the evolution and antioxidant activity of Chinese bayberry, we performed in-depth transcriptomic analysis to determine the genes that regulate antioxidant and pharmacological activities [8, 9]. The paper provides some vital data for the genetic improvement of bayberry

*Correspondence: renhy@zaas.ac.cn; 479046536@qq.com
[1] Institute of Horticulture, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China
Full list of author information is available at the end of the article

Ren *et al. BMC Plant Biol*   (2021) 21:452

Page 2 of 11

in the future and is conducive to the understanding of its genetic evolution.

Fruit quality is also a significant agricultural trait of bayberry and is difficult to dissect. Fruits with a distinctive flavor and color are more popular on the market. The color of bayberry fruit is primarily determined by the level of anthocyanin accumulation, which exists stably in plant cells, making the fruit purple or red [10, 11]. In addition to fruit color, total sugar content is the primary index of fruit flavor quality [12]. Sugar affects the flavor of the fruit and is the primary source of energy. Mature bayberry fruit primarily contains sucrose, fructose and glucose; of these sugars, sucrose accounts for the highest proportion [13]. The key period of color development of red bayberry is the whiting stage, after which the sugar components accumulate rapidly, and the total sugar content reaches the maximum at the harvest stage, when the fruit flavor quality is the best [12, 14]. Moreover, research findings suggest that xenia affects fruit yield and quality, and various pollen types have a marked impact on the quality of bayberry fruit [15, 16]. Research on this phenomenon has practical and immediate significance for the fruit tree industry.

With the rapid development of genomic technology and bioinformatic data, determining how multiomics data should be stored, interlinked and presented in one interface requires a sustained effort. From a researcher's point of view, some user-friendly tools, such as genome browsers, cultivar banks, genetic maps and design tools to assist selection, are necessary. Currently, there are several websites for plant research that are publicly available, such as the Cucurbitaceae database (GourdBase, http://www.gourdbase.cn/) [17] and Malvaceae database (MaGenDB, http://magen.whu.edu.cn) [18], but there is no publicly available database for bayberry. Collecting essential molecular breeding data, such as genome, transcriptome, genetic mapping and phenotypic information, of bayberry remains difficult. Therefore, it is urgently important to establish an integrated, professional and intelligent bayberry database to enable breeders, researchers and horticulturalists to retrieve and access many kinds of necessary data.

In this paper, a special database named the Bayberry Database fills the gap in the body of knowledge for significant plant genera, integrates large-scale multiomics data, implements integrative data visualization methods, and establishes a new functional comparison system to facilitate the research on molecular breeding and biological traits of bayberry. The primary purpose of this multifunctional database is to provide a reference genome resource, as well as germplasm (cultivars, fruit shape, and disease resistance), molecular marker and transcriptomic data, to enable marker-assisted breeding of bayberry. In addition, the Bayberry Database also features hundreds of fruit shape images, which is of considerable value to the study of bayberry fruit shape.

Utility and discussion
Bayberry database content
The Bayberry Database contains nine central functional domains: home, genome, germplasm, markers, tools, map, expression, reference, and contact (Fig. 1). At present, the Bayberry Database contains the following information: 26,325 genes with tissue expression patterns, fruit shape phenotypes for 102 diverse accessions, basic characteristics of 59 gSSR markers, transcriptomic data of 13,706 genes, and 2 genome maps of bayberry. The Bayberry Database also features a reference module in which 137 papers providing the above information can be freely browsed and downloaded with the authors' permission. Based on the above information, this database and its related tools will enable users to quickly retrieve large-scale functional information for biological research.
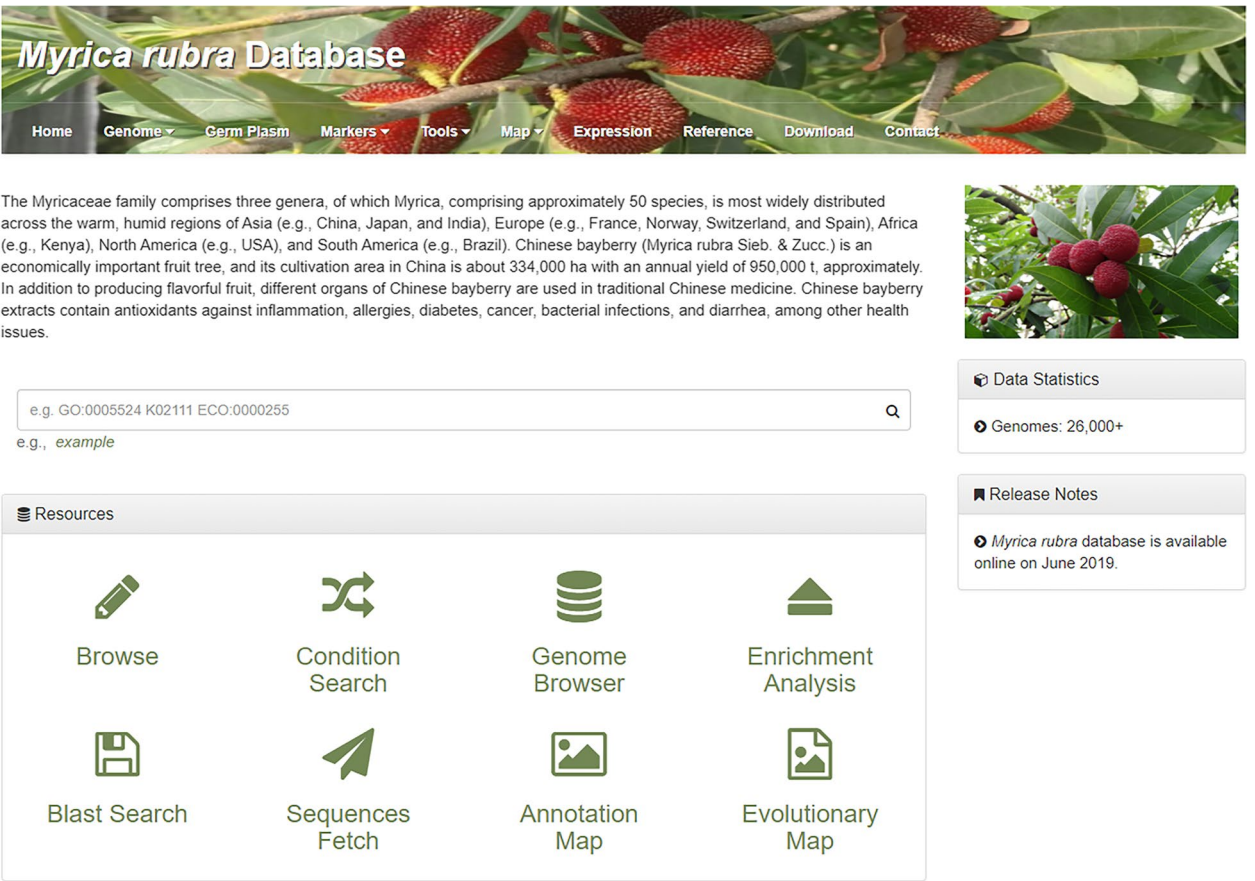
Application of the bayberry database
Genome search
The bayberry genome database offers the whole genome information of bayberry recently compiled by the authors' laboratory. The size of the whole genome of bayberry is approximately 289.92 Mb [8], which is considerably smaller than the genomes of apple (742 Mb) [19] and pear (527 Mb) [20]. When the user clicks on the 'Genome' heading, the column header label displays the 'Gene Browser', 'Condition Search', and 'Browse' subheadings (Fig. 2A). The user can access the sublinks and retrieve the required data by clicking on any one of these labels. For example, if the user clicks on 'Conditional Search', an additional layer will appear with two options that read 'Gene ID' and 'Region'. After clicking on the 'Region' option, 'Chromosome', 'Start' and 'End' are displayed (Fig. 2B). After entering the query information in the box, the user can then click on the 'Search' button, and a large quantity of data will display quickly according to the requirements (Fig. 2C).

Genome browser
The genome browser of the Bayberry Database comprises tracks describing the gene sequence, gene structure, mRNAs and other gene-related features and gives a visual display of explanatory notes on the bayberry genome (Fig. 3). Users can read the gene model on scaffolds from the genome browser. For example, if the user selects the genomic region of 4,130,542 bp to 4,411,741 bp on Scaffolds.01 for browsing, all of the genes in this zone will display as requested (Fig. 3A). If the user clicks on 'GENE_019554', the next level appears with specific gene

Fig. 1 Bayberry Database homepage. Nine main modules are displayed at the top of the interface and include the following: the genome, germplasm (phenome), markers, tools, maps, expression, reference, download and contact modules

data, such as the CDS, exons, mRNAs, and other features (Fig. 3B).

Germplasm resource data

At present, the germplasm module of the Bayberry Database contains phenotypic data of bayberry along with images of the fruits. The bayberry fruit shapes changed little, but there were differences in tree vigor, leaf type, fruit color, and single fruit weight, and the stages of early bloom and full bloom were slightly variable among the different cultivars. Most of the fruit shape types of bayberry appear in the germplasm module, which can be browsed and downloaded by users (Fig. 4A). The image data mainly exhibit distinctions in the shape, size and color of bayberry fruit. For example, the fruit color of GP00101 'Zaojia' is dark purple, and the fruit shape is intimate circle; the fruit color of GP00064 'Dongkui' is red, and the fruit shape is round. The module also provides a series of information on approximately 102 bayberry cultivars, which includes information such as the name, plant sex, place of origin, and level of disease
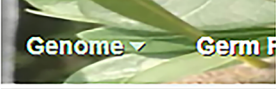
resistance (Fig. 4B). Users can enter the ID or name of cultivars in the search bar to browse and fetch information regarding a variety of bayberry cultivars.

Marker data

In the molecular marker module, 59 pairs of gSSR molecular markers were collected. Based on the genomic sequence information of bayberry [8], in total, 43,842 SSR loci have been identified, and 59 pairs of gSSRs have been cluster analyzed; the polymorphism information content (PIC), Nei's gene diversity (h*), effective number of alleles (Ne), and number of alleles (Na) can be found in the database [21].

Tool search

The tools module supplies users with diverse methods for genome searching. When the user clicks on 'Tools', the column title tags 'Enrichment Analysis', 'Sequence Fetch', 'Keyword Search', and 'Blast Search' are displayed. When the user clicks on 'Blast search', an extra layer appears with two options, 'Blastp' and 'Blastn'.

Fig. 2 Detailed functional view of the genome module. **A** 'Genome' includes 'Browse', 'Condition Search', and 'Genome Browser'. **B** Demonstration of the 'Search By Region' dialog box. **C** An example of the search result

Protein sequences can be compared with sequences in the protein database to identify the sequence from a distant source with Blastp, and given nucleic acid sequences can be compared with sequences in the nucleic acid database with Blastn. Users can enter a FASTA seq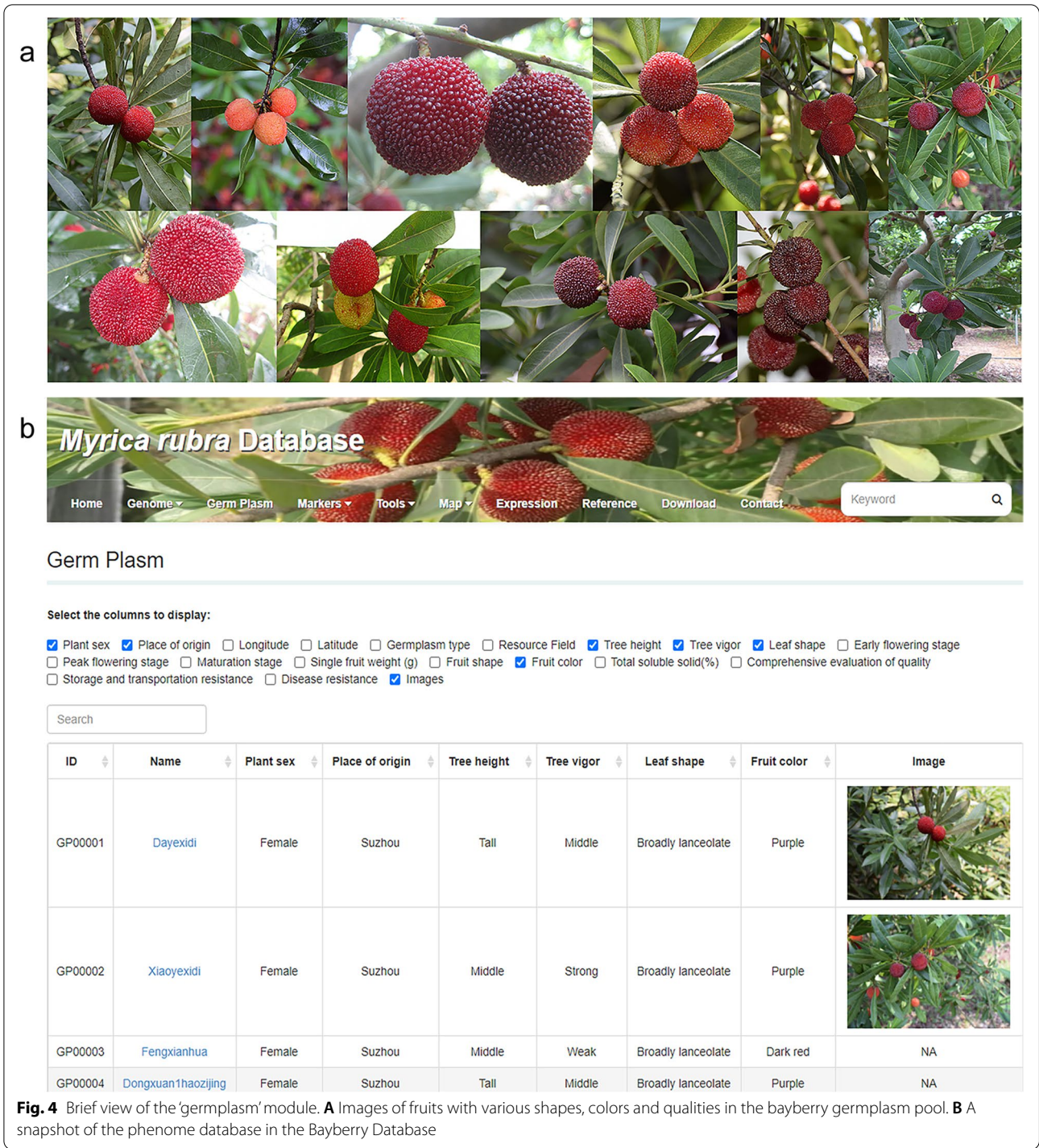uence in the box, select the database type (Fig. 5A), and click on the 'Search' button to obtain the nucleotide sequence comparison results; the sequence comparison results of Blastn or Blastp can be downloaded from the database (Fig. 5B). 'Keyword Search' is a more practical function in the tools module (Fig. 6). By inputting the desired keywords, the user can find

Fig. 3 Regional view of the genome using the Bayberry Database genome browser. **A** A graphic view of the region 4,130,542 bp to 4,411,741 bp on Scaffold1. **B** The interface after clicking on 'GENE_019554'
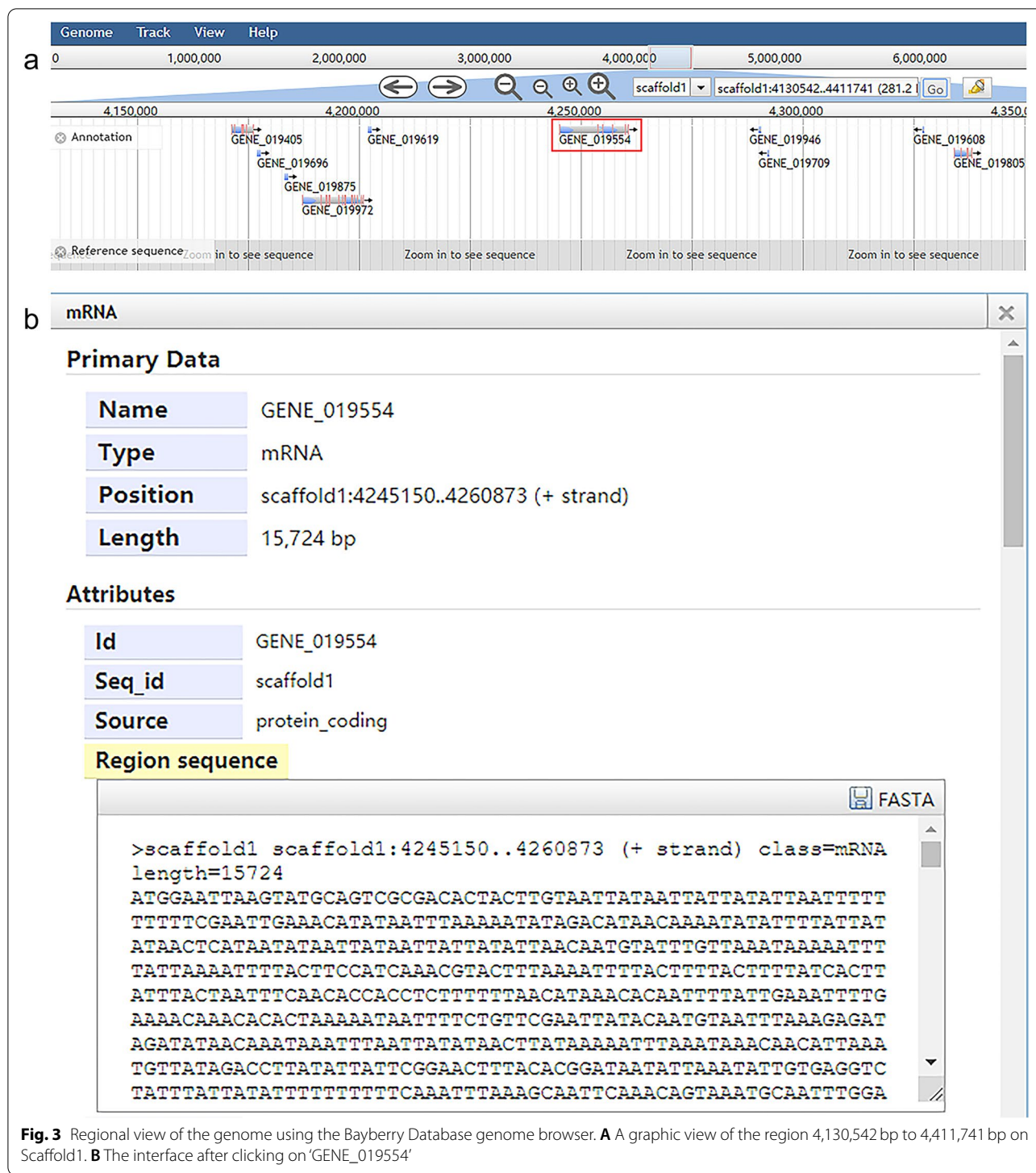
genes with that keyword content in the database. For example, if the user enters 'disease resistance' in the box and clicks on the 'Search' button (Fig. 6A), then the genes related to disease resistance will be displayed, such as 'GENE_000288' and 'GENE_000325'. (Fig. 6B). Detailed information on the genes can be obtained after clicking on their name (Fig. 6C). In the 'Sequence Extraction' function, users can input the starting fragment of the scaffold to obtain the DNA sequence corresponding to the genomic coordinates.

Ren *et al. BMC Plant Biol*   *(2021) 21:452*

Page 6 of 11

Fig. 4 Brief view of the 'germplasm' module. **A** Images of fruits with various shapes, colors and qualities in the bayberry germplasm pool. **B** A snapshot of the phenome database in the Bayberry Database

Gene functional enrichment analysis is a method to identify types of genes or proteins that are overrepresented in a large group of genes or proteins and may be related to disease phenotypes. The method employs statistical approaches to identify significantly enriched or depleted groups of genes. Transcriptomic technologies and proteomics results are often used to identify thousands of genes that are used for the analysis. There are two enrichment analysis methods in this database: 'GO' and 'KEGG'. Users can select the corresponding analysis methods according to their own needs.

Fig. 5 Screenshot of 'Blast Search' in the tools module. **A** Demonstration of the 'Blast Search' dialog box. **B** Query results after inputting the FASTA sequence and selecting the database type
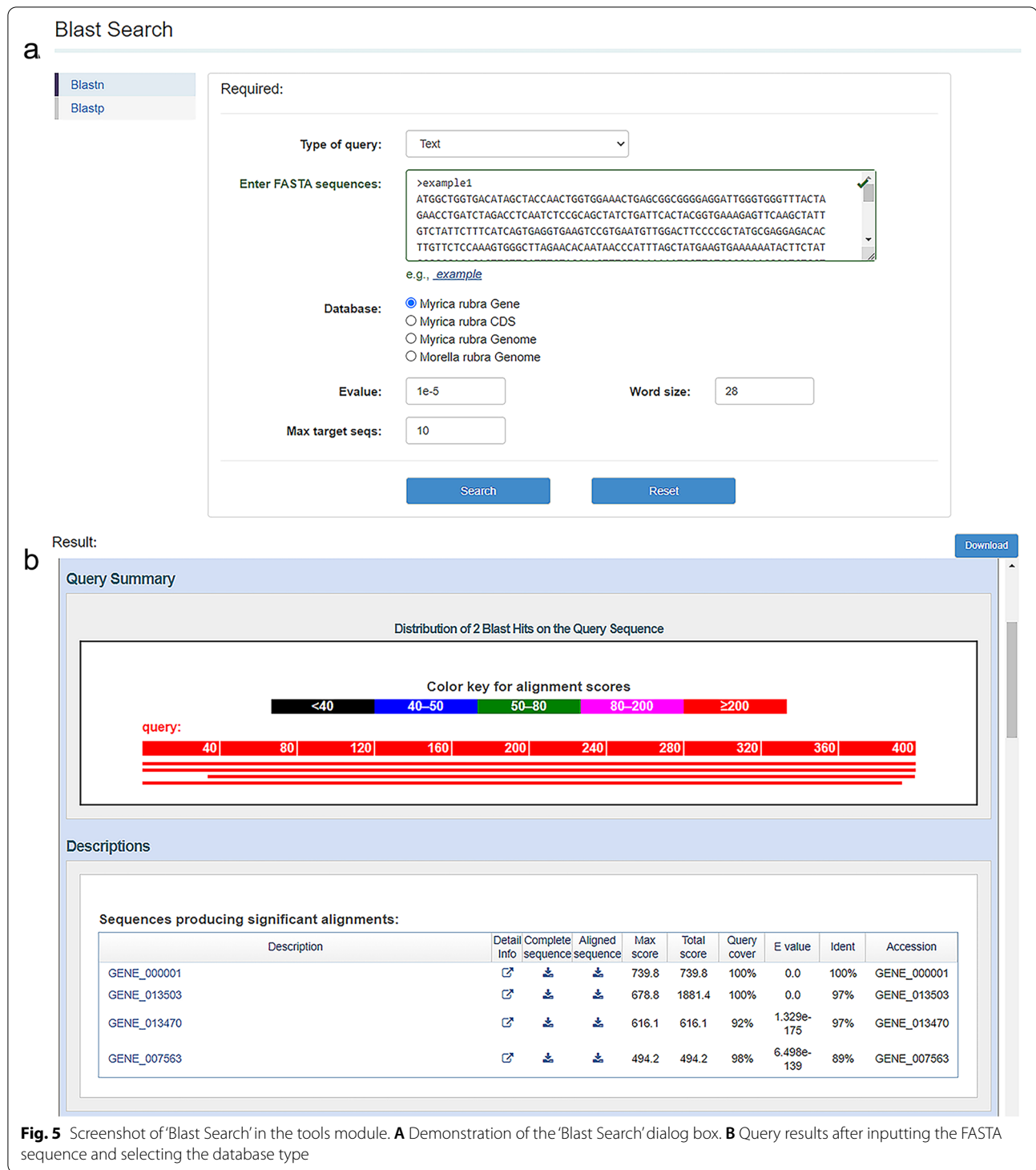
Genome map

The map module contains two kinds of bayberry gene maps. One is the red bayberry genetic haploblock (HB) map, which was constructed using RAD tag sequencing technology. The HB markers for the l m×ll segregation type are shown in red, those for nn×np are shown in green, and the ab × cd segregation type is shown in blue. The HBs in the centromere are indicated in bold. The other map is the alignment of female (a) and male (b) bayberry containing assembled scaffolds with the SNP marker linkage genetic map [22]. The map module contains the evolutionary genetic map of bayberry, a total of
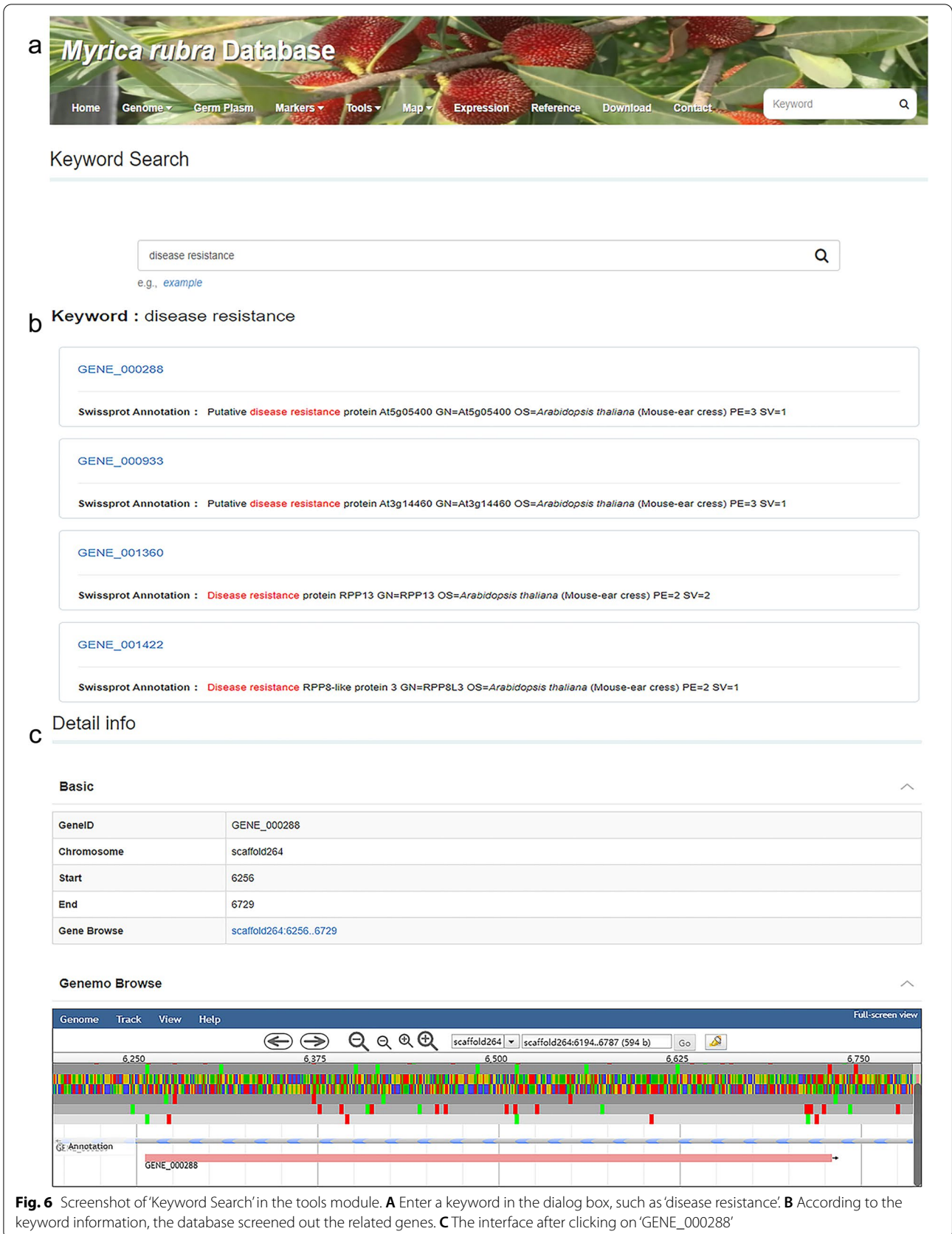
Fig. 6 Screenshot of 'Keyword Search' in the tools module. **A** Enter a keyword in the dialog box, such as 'disease resistance'. **B** According to the keyword information, the database screened out the related genes. **C** The interface after clicking on 'GENE_000288'

Ren *et al. BMC Plant Biol*   (2021) 21:452

Page 9 of 11

4270 gene families were shared by these four species, and 891 gene families were unique to bayberry. Based on 560 single-copy protein sequences, a phylogenetic tree was constructed for nine species (*M. rubra, Prunus mume, Citrullus lanatus, Morus notabilis, Fragaria × ananassa, Citrus sinensis, Carica papaya, Arabidopsis thaliana*, and *Solanum lycopersicum*) by using PhyML 4.0 [23]. Three species, *M. rubra, P. mume*, and *M. notabilis,* were clustered in a subclade that was most likely derived from a common ancestor approximately 93.95 Mya, whereas *P. mume* and *M. notabilis* diverged 82.41 Mya, and *C. lanatus* diverged from an ancestor shared with *M. notabilis, P. mume*, and *M. rubra* approximately 97.67 Mya [8]. In addition, comparative genomic analysis of bayberry and other fruits was also performed in the map module. A total of 13,216 unique gene families were produced by gene similarity clustering of the 26,325 predicted genes for *M. rubra* with those of *M. notabilis, F. × ananassa, Vitis vinifera*, and *S. lycopersicum*. We found that the five species shared 4074 gene families, and 712 gene families were unique to *M. rubra*. A total of 12,266 unique gene families were yielded by gene similarity clustering of the 26,325 predicted genes for *M. rubra* with those of *Oryza sativa, A. thaliana*, and *Cicer arietinum* [8].

Gene expression data

The gene expression module of the Bayberry Database contains transcriptome data of the fruit development process of two Chinese bayberry cultivars 'Zaojia 'and 'Dingao Bayberry' post pollination 10, 20, 30, 40, 50 days. Users can click on the sample name to obtain detailed information regarding the gene, such as gene expression, gene function, and cDNA sequence. The database offers a platform for visualization and download of specific gene data of bayberry, which can assist researchers and breeders in performing in-depth experiments.

Reference data

The reference module of Bayberry Database currently contains 126 publications supporting the aforementioned information from 1985 to 2020, which can be browsed and downloaded freely with the authors' permission. Among these publications, the completion and publication of bayberry genome sequencing results is helpful to understand the structure and function of the bayberry genome and provide an important reference for exploring the origin and evolution of species, the development and application of molecular markers, and the location and cloning of functional genes [8]. The process of molecular breeding may be accelerated by genome data. Based on the genomic sequence information of bayberry [8, 9], the 43,842 SSR loci in 623 scaffold sites were identified [21]. The genetic relationship between main cultivars and male plants of bayberry were analyzed by RAPD and ISSR markers, which provides theoretical guidance for the selection of male parents in bayberry cross breeding [24].

The Bayberry Database was made available online in June 2019 and is currently in the initial version. With the increasing number of innovative data in the Bayberry Database, various types of molecular markers (e.g., SSRs, indels, and QTLs), genome resequencing and new publications, the database will be continuously optimized with time. The data classification, integration and update will be emphasized in the Bayberry Database, and a large quantity of comprehensive data will be conveniently provided in the near future.

Some new resequenced genome, transcriptome and molecular marker data will be added to improve the functions of some modules of the Bayberry Database. The Synteny Viewer module will be added to the Bayberry Database because there is no publicly available database of the bayberry genome. This module has been used in the Genome Database for Rosaceae (https://www.rosaceae.org) [25] and the Citrus Genome Database (https://www.citrusgenomedb.org) [26]. The primary function of the module is to view genome synteny and homologous gene pairs between different species, and the gene synteny clusters between two genomes can be dynamically presented with Genome Synteny Viewer as a circular plot. The new map sets, phenome data, and novel search and download options will also be added. The long-term goals of the Bayberry Database include constructing an integrated, intelligent and comprehensive information system that integrates more genomic, transcriptomic and metabolomic data. With more bayberry cultivars available in omics data, the database will benefit both researchers and fruit-tree breeders.

Conclusions

The web address of the database is as follows http://www.bayberrybase.cn/. The Bayberry Database contains nine central functional domains: home, genome, germplasm, markers, tools, map, expression, reference, and contact. The *Myrica rubra* database is an intelligent, interactive, and user-friendly system that enables researchers, breeders and horticulturalists to browse, search and retrieve relevant and useful information and facilitates genomic research and breeding efforts concerning *Myrica rubra*. This database will be of great help to bayberry research and breeding in the future.

Ren *et al. BMC Plant Biol*   (2021) 21:452

Page 10 of 11

Methods

The bayberry database was deployed in Ubuntu 16.04 operation system and developed by AKKA 2.12 (web server), MySQL 5.7.26 (database server), Scala 2.12.2 and SBT 0.13.18. All data in database were managed and stored by MySQL Database Management System. The query function was enforced based on Slick 3.3.0 middleware tier. To visualize the genome, we used the Jbrowser 1.16.6. The website interface components were designed and implemented by the Bootstrap 3.3.0 and Play Framework 2.6.25. The website has been tested in several popular web browsers, including Firefox, Google Chrome and Internet Explorer. Source code is available in github https://github.com/Westsyan/Myrica_rybra_database.

Authors' contributions
HR conceived the paper. SZ collected and analyzed the data of germplasm resources and papers. HR and YH wrote the paper. All authors including HR, YH, XQ, XZ, ZY, and FH have read and agreed to contribute.

Availability of data and materials
Source code is available in github https://github.com/Westsyan/Myrica_rybra_database. The database and web interface can be accessed at http://www.baybe rrybase.cn/.

Declarations

Ethics approval and consent to participate
This paper was completed within the laws of the People's Republic of China. No specific permits were required for our field research. The study species is not included in the 'List of Protected Plants in China'.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Institute of Horticulture, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China. [2]College of Landscape Architecture, Nanjing Forestry University, Nanjing 210037, China.

References
1. Sun CD, Huang HZ, Xu CJ, Li X, Chen KS. Biological activities of extracts from Chinese bayberry (*Myrica rubra* Sieb. Et Zucc.): a review. Plant Food Hum Nutr. 2013;68(2):97–106.
2. Mathiesen L, Malterud KE, Sund RB. Antioxidant activity of fruit exudate and C-methylated dihydrochalcones from *Myrica gale*. Planta Med. 1995;61(6):515–8.
3. Erickson DL, Hamrick JL. Genetic and clonal diversity for *Myrica cerifera* along a spatiotemporal island chronosequence. Heredity. 2003;90(1):25–32.
4. Njung'e K, Muriuki G, Mwangi JW, Kuria KA. Analgesic and antipyretic effects of *Myrica salicifolia* (Myricaceae). Phytother Res. 2002;16(1):73–4.
5. Fang ZX, Zhang M, Sun YF, Sun JC. How to improve bayberry (*Myrica rubra* Sieb. Et Zucc.) juice color quality: effect of juice processing on bayberry anthocyanins and polyphenolics. J Agric Food Chem. 2006;54(1):99–106.
6. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science. 2008;320(5872):106–9.
7. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 2010;38(15):159.
8. Ren HY, Yu HB, Zhang SA, Liang SA, Zheng XA, Zhang SC, et al. Genome sequencing provides insights into the evolution and antioxidant activity of Chinese bayberry. BMC Genomics. 2019;20(1).
9. Qi XJ, Ren HY, Liang SM, Zheng XL, Wu YC. Preliminary report on genome sequencing of *Myrica rubra*. Zhejiang Agri Sci. 2015;56:564–1566 (in Chinsese).
10. Niu SS, Xu CJ, Zhang WS, Zhang B, Li X, Wang KL, et al. Coordinated regulation of anthocyanin biosynthesis in Chinese bayberry (*Myrica rubra*) fruit by a R2R3 MYB transcription factor. Planta. 2010;231(4):887–99.
11. Zhang WS, Li X, Zheng JT, Wang GY, Sun CD, Ferguson LB, et al. Bioactive components and antioxidant capacity of Chinese bayberry (*Myrica rubra* Sieb.And Zucc.) fruit in relation to fruit maturity and postharvest storage. Eur Food Res Technol. 2008;227(4):1091–7.
12. Liang SM, Xu YH, Wang W, Zheng XL, Ren HY, Peng J, et al. Study on the formation discipline of fruit appearance and main nutrient quality in the fruit development of Chinese bayberry. J Nucl Agri Sci. 2016;1142(6):1135–40 (in Chinsese).
13. Xie LJ, Ye XQ, Liu DH, Ying YB. Quantification of glucose, fructose and sucrose in bayberry juice by NIR and PLS. Food Chem. 2009;114(3):1135–40.
14. Zhu TT, Liang SM, Zhang SW, Zheng XL, Ren HY, Qi XJ. Study on fruit development and quality formation of (*Myrica rubra*) cv. Zaoxian Acta Agri Zhejiangensis. 2018;30(5):764–70 (in Chinsese).
15. Qin LZ, Li BG, Qi GH. The research advances of metaxenia. Hebei J Forestry Orchard Res. 2002;4:371–5 (in Chinsese).
16. Qi XJ, Zheng XL, Ren HY, Liang SM, Yan HP, WU CY. Effect of xenia on fruit quality and sucrose metabolism enzyme activity in red bayberry. J Fruit Sci. 2017;34(7):861–7 (in Chinsese).
17. Wang Y, Xu P, Wu XH, Wu XY, Wang BG, Huang YP, et al. GourdBase: a genome-centered multi-omics database for the bottle gourd (*Lagenaria siceraria*), an economically important cucurbit crop. Sci Rep. 2018;8.
18. Wang D, Fan W, Guo X, Wu K, Zhou S, Chen Z, et al. MaGenDB: a functional genomics hub for Malvaceae plants. Nucleic Acids Res. 2020;48(D1):D1076–84.
19. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, et al. The genome of the domesticated apple (*Malus × domestica* Borkh.). Nature Genet. 2010;42(10):833–9.
20. Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, et al. The genome of the pear (*Pyrus bretschneideri Rehd.*). Genome Res. 2012;23:396–408.
21. Zhang SW, Liang SM, Zheng XL, Ren HY, Zhu TT, Qi XJ. Development of genomic SSR and application in Chinese bayberry. Acta Horti Sinica. 2019;46(1):149–56 (in Chinsese).

22. Jia HM, Jia HJ, Cai QL, Wang Y, Zhao HB, Yang WF, et al. The red bayberry genome and genetic basis of sex determination. Plant Biotechnol J. 2019;17:397–409.
23. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52:696–704.
24. Xie XB, Qiu YY, Qi XJ, Zheng XL, Qiu LJ, Zhang YJ. Analysis of genetic relationship between male and female plants in *Myrica rubra* by RAPD and ISSR. J Fruit Sci. 2008;25:198–202 (in Chinsese).
25. Jung S, Lee T, Cheng CH, Buble K, Zheng P, Yu J, et al. 15 years of GDR: new data and functionality in the genome database for Rosaceae. Nucleic Acids Res. 2019;47(1):D1137–45.
26. Buble K, Jung S, Humann JL, Yu J, Cheng CH, Lee T, et al. Tripal MapViewer: A tool for interactive visualization and comparison of genetic maps. Database (Oxford). 2019:baz100. https://doi.org/10.1093/database/baz100.

Publisher's Note