Computational and Structural Biotechnology Journal 18 (2020) 2766–2773

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

BnaSNPDB: An interactive web portal for the efficient retrieval and analysis of SNPs among 1,007 rapeseed accessions

Tao Yan, Qian Wang, Antony Maodzeka, Dezhi Wu, Lixi Jiang *,[1]

Provincial Key Laboratory of Crop Gene Resource, Zhejiang University, 866 Yu-Hang-Tang Road, Hangzhou 310058, PR China

ARTICLE INFO

Article history:
Received 11 July 2020
Received in revised form 18 September 2020
Accepted 18 September 2020
Available online 29 September 2020

Keywords:
SNP
Database
Rapeseed
Genomic variation

ABSTRACT

The rapid development of high-throughput sequencing technology and the decrease in sequencing costs provide valuable resources and great opportunities for researchers to investigate genomic variations across hundreds or even thousands of accessions in the post-genomic era. The management and exploration of these large-scale genomic variations heavily rely on programming and command-line environments, which are challenging and time-consuming for most experimental biologists and plant breeders. Here, we present BnaSNPDB, an interactive web portal with a user-friendly interface that provides multiple analysis modules for retrieving, analyzing, and visualizing single nucleotide polymorphisms among 1,007 accessions of worldwide rapeseed germplasm. It is compatible, extendable, and portable to be easily set up on different operating systems, and can be accessed at http://121.41.229.126:3838/bnasnpdb and http://rapeseed.zju.edu.cn:3838/bnasnpdb. Its whole dataset and code are available at https://github.com/YTLogos/BnaSNPDB. This database is essential for accelerating studies on the functional genomics and screening of the molecular markers of molecular-assisted breeding in rapeseed.

1. Introduction

With the rapid development of high-throughput sequencing (HTS) technology and the decreasing sequencing costs, biological data are growing exponentially, leading to the big data era of biology. Rapeseed (*Brassica napus* L.) is an economically important crop grown mainly as a source of edible oil and protein-rich livestock feed. The advent of high-throughput genomic technologies and the availability of the reference genome sequence of *B. napus* [1–3] provide new insights into the genomic localization of quantitative trait loci (QTL) associated with agronomic traits, the identification of genes underlying QTLs and variation in allelic sequences, the design of markers, the exploration of evolutionary history, and the practice of molecular breeding [4,5].

Single nucleotide polymorphisms (SNPs) are the most common type of genomic variation and used in various biological studies, such as domestication studies [6], genome selection [7], and genome-wide association studies (GWAS) [8]. The genomic variation data of numerous rapeseed accessions have been reported because of the abundant germplasm resources [4,5]. Large panels of raw sequencing data have been deposited to the sequence read archive (SRA) of the National Center of Biotechnology Information (NCBI), which is freely accessible to all researchers. However, raw data acquired from whole-genome re-sequencing should be aligned to the reference genomes for SNPs that can be stored in standard Variant Call Format (VCF) files [9]. This process, which requires a great number of command-line tools, bioinformatics software, analysis pipelines, and packages, is a challenging and time-consuming task for most experimental biologists and plant breeders. A number of databases such as SNP-Seek for rice [10], SorGSD for sorghum [11], ZEAMAP for maize [12], and CerealsDB for wheat [13], have been constructed to store and analyze SNPs of various organisms. Field researchers have been benefited from these databases in identifying many functional genes, such as the key genes related to nitrogen use efficiency in rice [14] and the genes for fertility restoration in maize [15]. Based on the identification of functional genes, molecular breeding and genetic engineering targeting these genes can be further carried out. However, existing databases have been constructed with multiple programming languages and database management systems, such as MySQL, which are not compatible, extendable, and portable enough to be migrated to other platforms. Several *Brassica* databases such as brassica.info (http://www.brassica.info) and CropSNPdb [16], have been built to store and analyze genomic variations. These databases are essential for functional genomic

* Corresponding author.
E-mail address: jianglx@zju.edu.cn (L. Jiang).
[1] ORCID: 0000-0002-8579-0763.

https://doi.org/10.1016/j.csbj.2020.09.031
2001-0370/

studies on rapeseed although only a simple query interface is provided or built based on the SNP array data produced by *Brassica* 60 K arrays (Supplementary Table S1).

In this study, using the R/Shiny framework [17,18], we developed a SNP database for *B. napus* (BnaSNPDB), an interactive web portal with a user-friendly interface that provides multiple analysis modules for visualizing and exploring SNPs among 1,007 rapeseed germplasm accessions based on the data reported in a previous research [4]. BnaSNPDB is compatible, extendable, portable, and easy to be established in different operating systems (Linux, macOS, and Microsoft Windows) with an installation guide at https://github.com/YTLogos/BnaSNPDB. With a fast and efficient access to large-scale genomic variation data, it is helpful for experimental biologists and plant breeders who investigate genomic variations without the need for programming skills. It is also useful for experimental biologists who are working in population genetics, functional genomics, and molecular-marker assisted rapeseed breeding.

2. Results

2.1. Construction of BnaSNPDB

BnaSNPDB was designed for deployment in different operating systems (Linux, macOS, and Microsoft Windows) by using the R/Shiny framework and R packages to retrieve and analyze genotypic variations. The genotypes of 1,007 rapeseed accessions across 2,404,340 SNP sites were stored as R data files in BnaSNPDB and could be efficiently retrieved without using the SQL language. BnaSNPDB was implemented using the R/Shiny framework, which possesses convenient functions for retrieving and visualizing genotypic data. BnaSNPDB provides an efficient way to rapidly access genotypic data in user-specific genomic regions, filter SNP sites and samples, and generate a genotype table as intermediate data. The intermediate genotypic data were stored in Random Access Memory (RAM) and then used for subsequent analyses and visualizations to enhance the capacity of genotypic data reuse for in-depth explorations. The whole dataset and code of BnaSNPDB are freely available at https://github.com/YTLogos/BnaSNPDB.

2.2. Interface and main functions of BnaSNPDB

BnaSNPDB supports the navigation of massive genomic variations in user-specified samples/accessions and genomic regions and performs lightweight analyses and visualizations through the R/Shiny framework. It offers uniform and flexible interfaces for manipulating query parameters and provides eight menus, namely, Home, LDheatmap, SNPdistribution, Phylogenetic, Diversity, Extraction, Documentation, and About (Fig. 1). The Home menu briefly introduces BnaSNPDB, and the Documentation menu displays the online uses in detail and the installation procedures of BnaSNPDB. The About menu presents information about technologies and packages used in BnaSNPDB. The remaining menus, including LDheatmap, SNPdistribution, Phylogenetic, Diversity, and Extraction, display versatile analysis and visualization functions. The graphics produced by these functions are available for download in various formats (e.g., png, pdf, tiff, jpeg, bmp, and svg), and the plain text file can be downloaded in different formats (e.g., txt, tsv, xlsx, and csv) for subsequent use.

Linkage disequilibrium (LD) heat map is widely used to display the pairwise LD measurements of SNPs in local GWAS signals. Here, the LDheatmap function provides the feasibility to calculate and visualize the LD between all pairs of SNP sites in a user-specified genomic region (Fig. 2A). For a user-specified genomic region, the LDs measured in r^2 between any pair of SNP sites are calculated and displayed as a heat map by using the R package LDheatmap [19]. This function supports a panel of parameters for filtering sites, such as the minor allele frequency (MAF), SNP effects, and rapeseed accessions, to restrict the SNPs used in calculation and visualization (Fig. 2B). Users can also display the heat map adjacently below a physical map showing the positions of SNPs, and the annotated gene models located in the user-specified genomic region (Fig. 2A). For example, this function can display the correlations between the SNPs in the *FLOWERING LOCUS C* (*FLC*) ortholog (*BnaA10g22080D*) and the *BnaC.SWEET4.a* (*BnaC07g24860D*) regions that indicate strong LDs (Fig. 2A; Supplementary Fig. S2) [4,20].

The SNPdistribution function is an intuitive way to visualize tabular genotype information as a heat map (Fig. 3A). A user needs to upload samples in one or more group lists. This function also

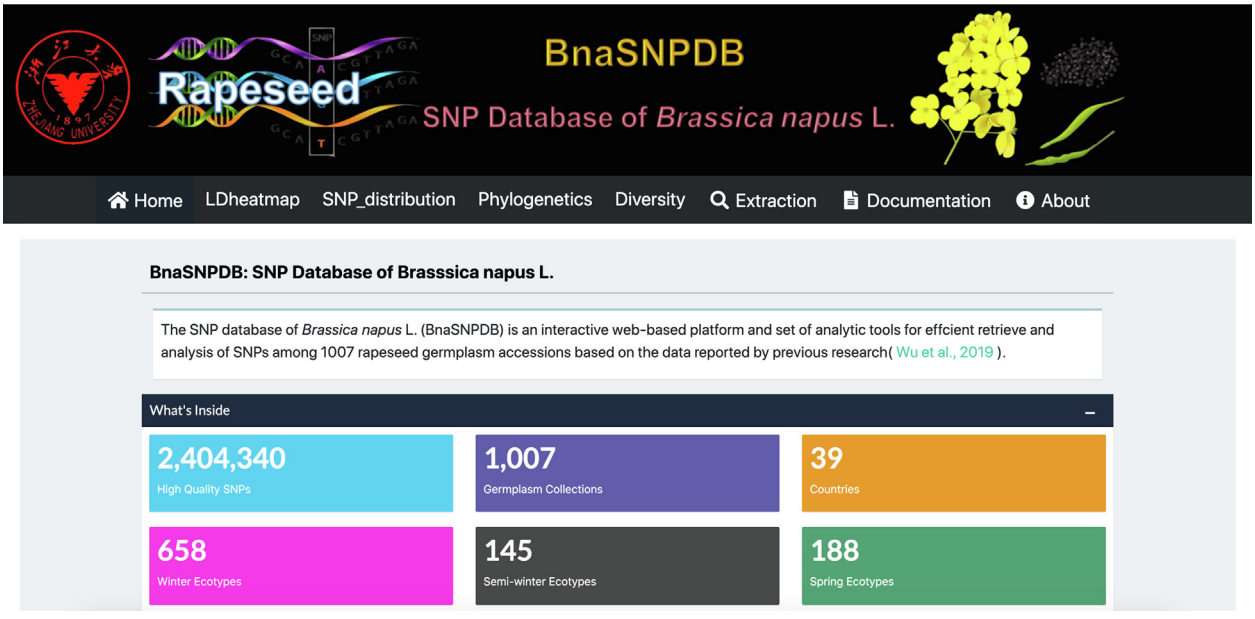

Fig. 1. The graphical interface of the BnaSNPDB database. A total of 8 menus are implemented in the database.

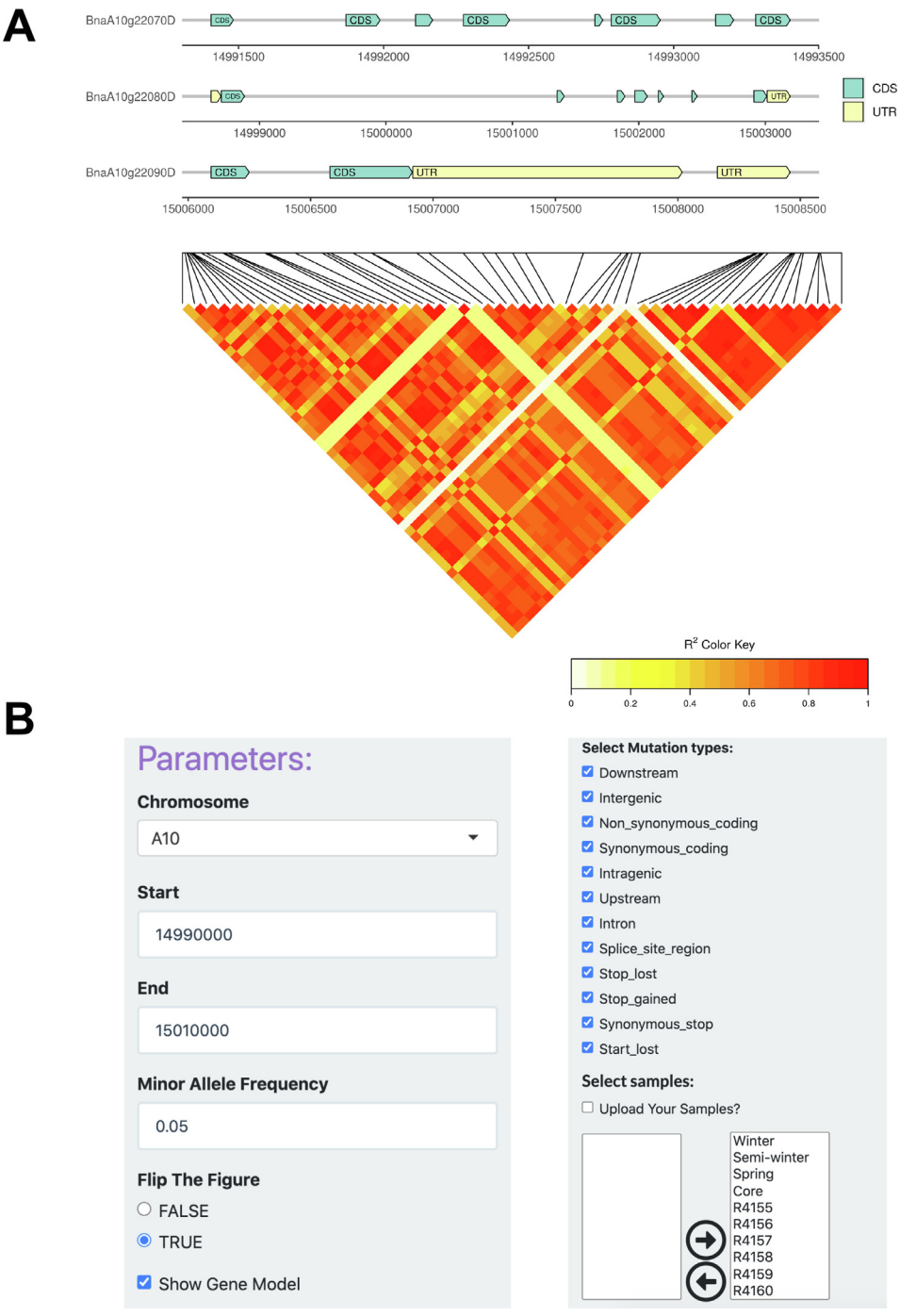

Fig. 2. An example of a linkage disequilibrium heat map generated using the LDheatmap function of BnaSNPDB. A. LD heat map of the ~10-kb *FLC* ortholog (*BnaA10g22080D*) region. Linkage disequilibrium measured in r^2 is represented as an inverted-triangle heat map. The color key indicates r^2 values. The structure of annotated gene models is shown on top of the heat map. B. A uniform, flexible interface for manipulating the query parameters are provided. Users can choose a specific genomic region, SNP effects, samples, etc.

supports different parameters for filtering sites, such as MAF and SNP effects, which can be used to restrict the SNP sites used in calculation and visualization and be more intuitive in the visualization of SNP sites distribution (Fig. 3B). By default, SNP sites are shown in columns, and samples are displayed along rows. Curcuminoid, blue, and red colors indicate the SNPs homozygous for the reference allele (0), heterozygous SNPs (1), and the SNPs homozygous for the non-reference allele (2), respectively. This function can be useful for exploring group-specific haplotypes or genotype patterns. For example, using this function the conserved SNPs specific to the three ecotype groups were found in the putative promoter region of the *FLC* ortholog (*BnaA10g22080D*) and the *FLOWERING LOCUS T* (*FT*) ortholog (*BnaA02g12130D*; Fig. 3A; Supplementary Fig. S3) [4].

The Phylogenetic function is designed to perform phylogenetic analysis and explore the genetic distances and evolutionary history based on high-density SNP data (Fig. 4A). The distance matrix is calculated on the basis of the genetic distance between accessions in the user-specified genomic region (Fig. 4B). The R package APE is then used to construct a neighbor-joining (NJ) tree based on

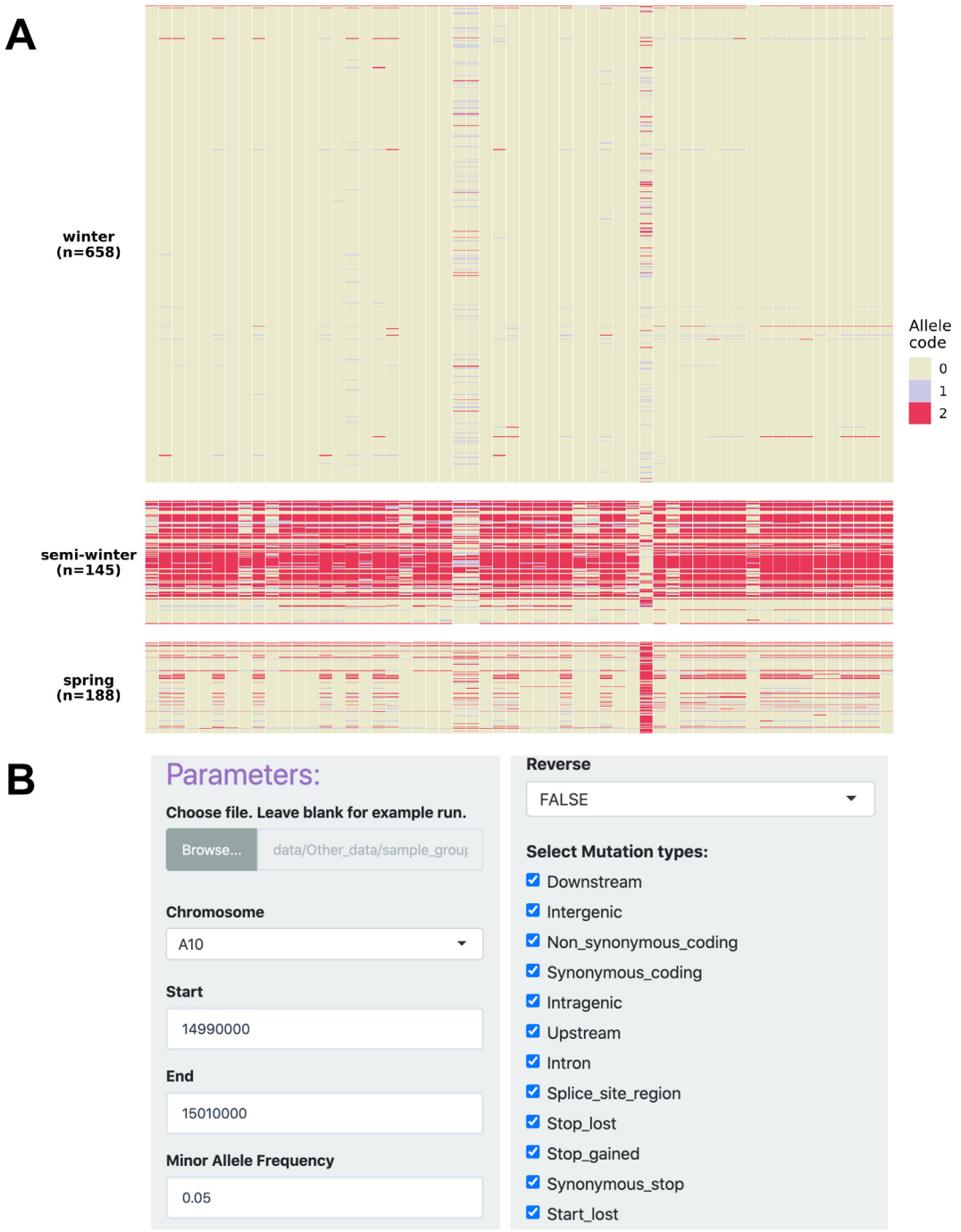

Fig. 3. An example of the SNPdistribution function. A. Comparison of conserved SNPs specific to the three ecotype groups in the *FLC* ortholog (*BnaA10g22080D*) region. The curcuminoid, blue, and red colors indicate SNPs homozygous for the reference allele (0), heterozygous SNPs (1), and SNPs homozygous for the non-reference allele (2), respectively. B. A uniform, flexible interface for manipulating the query parameters are provided. Users can choose a specific genomic region and SNP effects upload samples in one or more group lists. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

genetic distances [21]. The NJ tree is visualized in a circular format by using the ggtree package [22]. This function helps users visualize the distances between the samples of a genomic region.

The Diversity function allows users to calculate nucleotide diversity among groups of rapeseed accessions in a user-specified genomic region (Fig. 5A). This function supports some parameters for filtering SNP sites, such as MAF and SNP effects (Fig. 5B). A user-specified genomic region is split into non-overlapping genomic regions, and each genomic region contains 10 SNP sites. The nucleotide diversity among rapeseed accessions belonging to a specific subgroup is calculated for each genomic region with the R package pegas [23] and visualized with the R package ggplot2 [24]. This function allows users to select the number of SNPs in each window and decide whether or not gene models should be shown in the specified genomic region.

The Extraction function provides an interface for extracting SNPs, genes, and accession information from BnaSNPDB. The genes of four reference genomes annotated on the basis of different databases can be retrieved and downloaded. The detailed information and geographical distribution of user-chosen rapeseed accessions

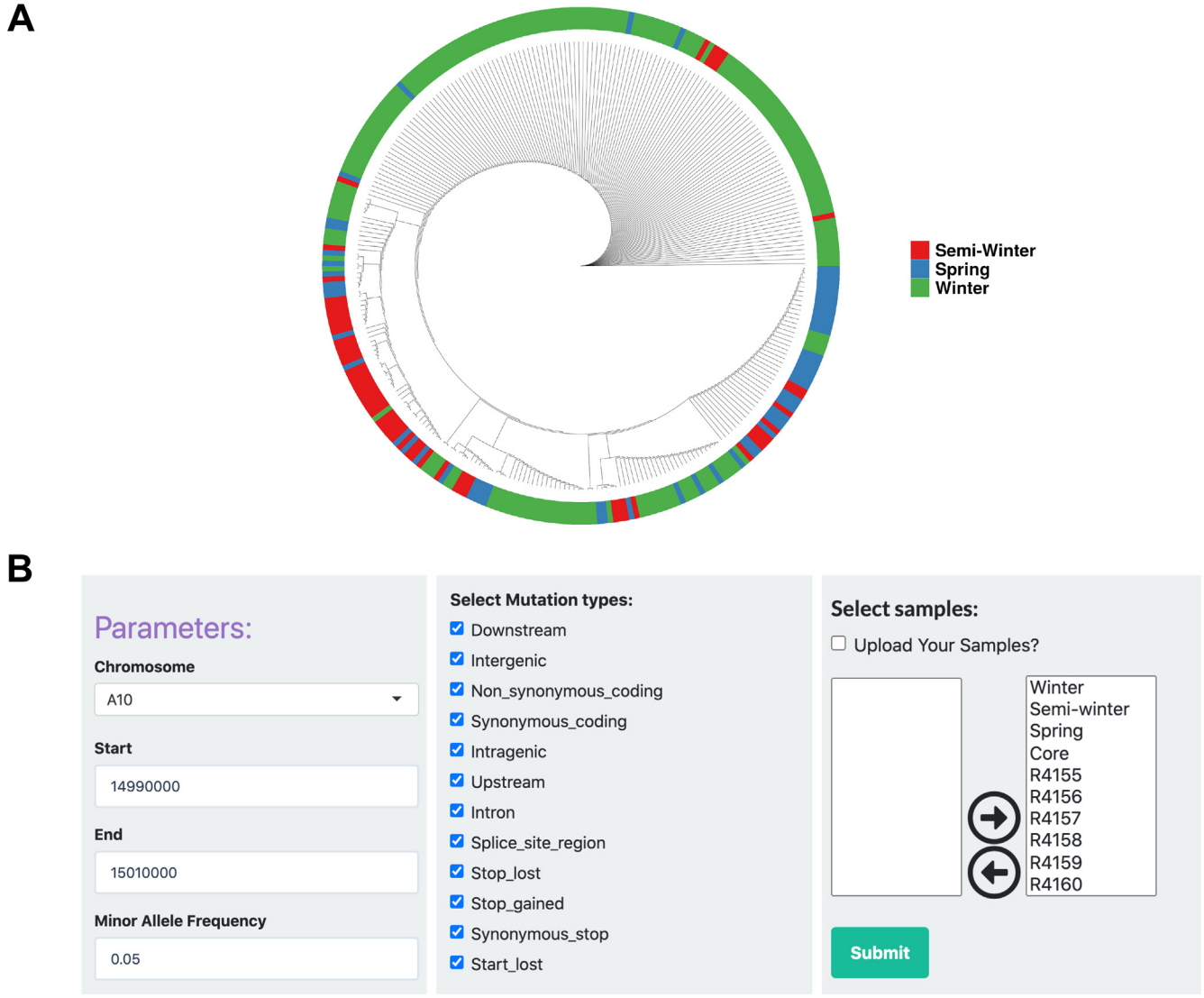

Fig. 4. An example of the Phylogenetic function. A. NJ tree constructed based on SNPs around the *FLC* ortholog (*BnaA10g22080D*) using the Phylogenetic function of BnaSNPDB. The NJ tree is visualized in the circular format using the ggtree package. Each tip of the tree represents a rapeseed accession. B. A uniform, flexible interface for manipulating the query parameters are provided. Users can choose a specific genomic region, SNP effects, samples, etc.

are displayed in figures and tables, respectively, which are available for download (Supplementary Figs. S4-S5).

To make it easier for users to make better use of this database, we have uploaded a brief but informative tutorial on Youtube (https://www.youtube.com/watch?v=8mBHzsKVotc), and on Bilibili (https://www.bilibili.com/video/BV1654y1C7hU) as well

3. Discussion and conclusion

With the increasing affordability of genome sequencing, large amounts of sequencing data are produced; consequently, biology has reached the big data and post-genomic era. The availability of the *B. napus* reference genome [1–3] and the emergence of high-throughput sequencing technologies have led to the generation of large rapeseed genomic variation datasets [4]. Obstacles in the analysis of such data can be resolved by using various bioinformatics tools, but doing so relies heavily on programming skills, which pose a particular challenge to most wet-lab biologists and plant breeders. Meeting the requirements of uniform, user-friendly, and powerful platforms, such as web servers, is difficult, although they can facilitate fast and efficient access to massive rapeseed genomic variation data in a local area and in a centralized location. In this context, BnaSNPDB is developed, and its SNP dataset is the largest and the most comprehensive genomic variation dataset of rapeseed so far. It can be flexibly used to perform genetic analysis of a specific genomic region. It can also be utilized to easily produce, decorate, and download publication-quality figures with various format preferences. In a traditional framework for building a database, multiple programming languages and database management system are involved; as such, establishing the database in different operating systems and local servers is difficult. By contrast, BnaSNPDB is purely constructed using the R code, which is compatible with different operating systems, such as Linux, MacOS, and Microsoft Windows. A detailed installation guide for users is provided at the https://github.com/YTLogos/BnaSNPDB. This database will be helpful for future studies on population genomics, molecular-marker-assisted rapeseed breeding, comparative genomics, evolutionary studies, and other related fields.

The SNPs of this database were obtained from the genome resequencing of a worldwide collection of 1,007 germplasm accessions. The sequencing depth of the study at the time, ~6.9× depth in aver-

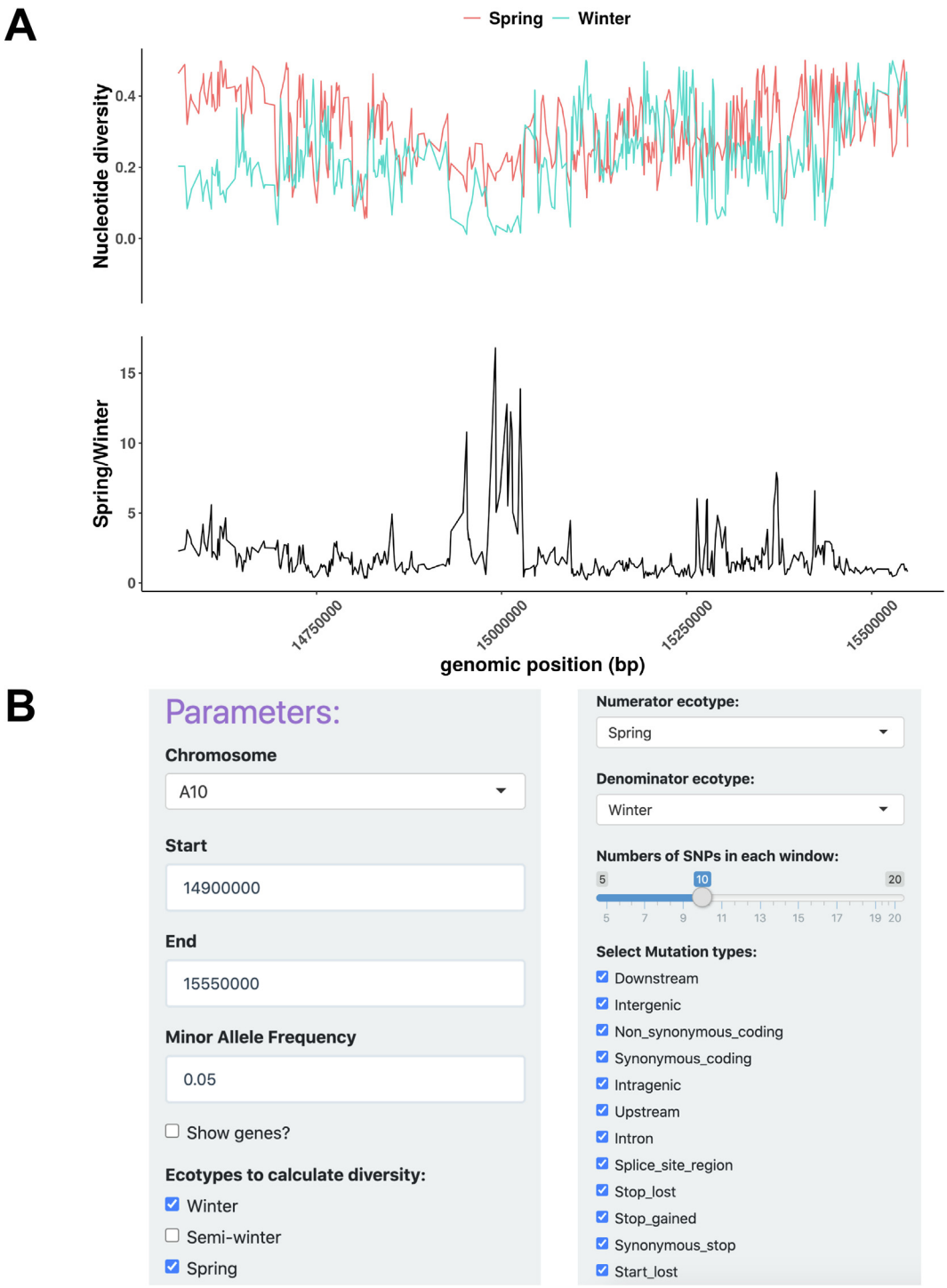

Fig. 5. An example of the Diversity function. A. The top panel displays the nucleotide diversity between winter ecotypes and spring ecotypes indicated by different colors. The bottom panel shows the division of nucleotide diversity in winter ecotypes and spring ecotypes. B. A uniform, flexible interface for manipulating the query parameters are provided. Users can choose a specific genomic region, SNP effects, samples, numbers of SNPs in each window, numerator ecotype, denominator ecotype, etc.

age, was consistent with the resequencing projects on other polyploidy species, such as upland cotton (*Gossypium hirsutum*) [25,26]. Our thought was to include as many individual genotypes as possible to attain the largest resource of genetic polymorphism. The balance between the accession number and sequencing depth did not allow deep sequencing for structural variation (SV) analysis, such as presence-absence variation (PAV) and copy-number variation (CNV) analyses. The SV analysis results obtained from a sequencing depth of less than 30× may be unreliable. Therefore, in BnaSNPDB, the analyses were based on SNP variations. *FLOWERING LOCUS C* (*BnaA10g22080D*) [4] and *SWEET4* (*BnaC07g24860D*) [20] were used as examples to show the associations between allelic SNP variations in genetic populations and agronomic traits, such as flowering time and leaf trichome, respectively, and to demonstrate the efficiency of the tools (Fig. 2A; Supplementary Fig. S2).

Studies have been conducted on SVs, particularly small-scale (30–10,000 bp) to mid-scale (10,000–30,000 bp) SVs, and their impact on evolution- and adaptation-related traits in *B. napus*

[3,25]. Long-read sequencing technologies have revealed a high level of widespread small-scale to mid-scale SVs in *B. napus* and suggested that up to 10% of all genes are influenced by these SVs. Some examples have also been presented to describe the contribution of such SVs to adaptation and disease resistance in rapeseed.

SNP and SV analyses complementarily reflect the genetic polymorphism of a species. However, building a more comprehensive interactive web portal for this polyploid species by integrating SNP and SV data and providing more analysis modules remains challenging for rapeseed researchers. Nevertheless, our approach made a significant progress toward this goal.

4. Materials and methods

4.1. Processing of genomic variation data

The genotypic data among 1,007 rapeseed accessions were obtained from a previous research [4]. The worldwide collection of 1,007 *B. napus* germplasm accessions includes 668 winter types, 149 semi-winter types, and 190 spring-types from 39 countries and has an average of ~6.6-fold coverage (Supplementary Table S2, Supplementary Fig. S1). Non-biallelic SNPs were removed using BCFtools (https://github.com/samtools/bcftools). The SNPs with a MAF lower than 5% were filtered. The SNP sites with a missing rate of higher than 50% were further removed. As a result, 2,691,432 SNPs were removed, and 2,404,340 high-quality SNPs were obtained. The SNPs were annotated on the basis of the *B. napus* v4.1 reference genome [1] by using SnpEff [27]. The rapeseed reference genome was split into non-overlapping genomic regions, and the genotypic data at SNP sites in each genomic region were stored as an R data file (.RData), which could be efficiently loaded into the memory with the R programming language [17]. The genotypic data were converted into an integer sparse matrix in the R data file to facilitate the storage and retrieval of the genomic dataset in accordance with a previously reported approach [28]. All the data used to develop BnaSNPDB can be obtained at https://github.com/YTLogos/BnaSNPDB/tree/master/data.

4.2. Building of the interactive user interface and analysis functions of BnaSNPDB

The interactive user interface is implemented using the R/Shiny framework, with powerful and convenient functions for post-processing and visualizing genotypic data. Numerous open-source bioinformatics analysis tools/pipelines are implemented with R, so BnaSNPDB utilizes the R/Shiny framework to build the interactive web portal for compatibility, extendability, and portability. All the R scripts used to develop the interactive web portal are freely available at https://github.com/YTLogos/BnaSNPDB.

5. Availability of supporting source code and requirements

Project name: BnaSNPDB.

Project home page: http://121.41.229.126:3838/bnasnpdb, http://rapeseed.zju.edu.cn:3838/bnasnpdb.

GitHub repository: https://github.com/YTLogos/BnaSNPDB.

Operating system(s): Platform independent.

Programming language: R (≥3.6.0).

R packages used in this study: shiny(1.5.0), shinydisconnect (0.1.0), ggplot2(3.3.2), stringr(1.4.0), dplyr(1.0.2), tidyr(1.1.1), forcats(0.5.0), patchwork(1.0.1), glue(1.4.1), ggpubr(0.4.0), writexl (1.3), snpStats(1.38.0), IRanges(2.22.2), LDheatmap(0.99–8), ape (5.4–1), pegas(0.13), gridExtra(2.3), grid(4.0.2), ggtree(2.2.1), shinycssloaders(1.0.0), shinysky(0.1.3), shinydashboard(0.7.1), shinyWidgets(0.5.3), gggenes(0.4.0), DT(0.15), shinythemes(1.1.2), NAM(1.7.3), adegenet(2.1.3) (Supplementary Table S3).

License: GPLv3.

Any restrictions to use by non-academics: None.

CRediT authorship contribution statement

Tao Yan: Conceptualization, Methodology, Software, Formal analysis. **Qian Wang:** Data curation, Software. **Antony Maodzeka:** Data curation. **Dezhi Wu:** Supervision, Project administration. **Lixi Jiang:** Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by Natural Science Foundation of China (Code No. 31961143008, 31671597, 31971817) and Jiangsu Collaborative Innovation Center for Modern Crop Production.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.09.031.

References

[1] Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, et al. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 2014;345:950–3. https://doi.org/10.1126/science.1253435.

[2] Bayer PE, Hurgobin B, Golicz AA, Chan C-K, Yuan Y, Lee H, et al. Assembly and comparison of two closely related Brassica napus genomes. Plant Biotechnol J 2017;15(12):1602–10. https://doi.org/10.1111/pbi.12742.

[3] Chawla HS, Lee H, Gabur I, Vollrath P, Tamilselvan-Nattar-Amutha S, Obermeier C, et al. Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. Plant Biotechnology Journal n.d.;n/a. https://doi.org/10.1111/pbi.13456.

[4] Wu D, Liang Z, Yan T, Xu Y, Xuan L, Tang J, et al. Whole-genome resequencing of a worldwide collection of rapeseed accessions reveals the genetic basis of ecotype divergence. Mol Plant 2019;12:30–43. https://doi.org/10.1016/j.molp.2018.11.007.

[5] Lu K, Wei L, Li X, Wang Y, Wu J, Liu M, et al. Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. Nat Commun 2019;10. https://doi.org/10.1038/s41467-019-09134-9.

[6] Meyer RS, Choi JY, Sanches M, Plessis A, Flowers JM, Amas J, et al. Domestication history and geographical adaptation inferred from a SNP map of African rice. Nat Genet 2016;48:1083–8. https://doi.org/10.1038/ng.3633.

[7] Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 2010;9:166–77. https://doi.org/10.1093/bfgp/elq001.

[8] Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 2012;44:32–9. https://doi.org/10.1038/ng.1018.

[9] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156–8. https://doi.org/10.1093/bioinformatics/btr330.

[10] Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, et al. Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res 2017;45:D1075–81. https://doi.org/10.1093/nar/gkw1135.

[11] Luo H, Zhao W, Wang Y, Xia Y, Wu X, Zhang L, et al. SorGSD: a sorghum genome SNP database. Biotechnol Biofuels 2016;9. https://doi.org/10.1186/s13068-015-0415-8.

[12] Gui S, Yang L, Li J, Luo J, Xu X, Yuan J, et al. ZEAMAP, a comprehensive database adapted to the maize multi-omics era. IScience 2020;23:101241. https://doi.org/10.1016/j.isci.2020.101241.

[13] Wilkinson PA, Allen AM, Tyrrell S, Wingen LU, Bian X, Winfield MO, et al. CerealsDB—new tools for the analysis of the wheat genome: update 2020. Database (Oxford) 2020;2020. https://doi.org/10.1093/database/baaa060.

[14] Tang W, Ye J, Yao X, Zhao P, Xuan W, Tian Y, et al. Genome-wide associated study identifies NAC42-activated nitrate transporter conferring high nitrogen use efficiency in rice. Nat Commun 2019;10. https://doi.org/10.1038/s41467-019-13187-1.

[15] Jaqueth JS, Hou Z, Zheng P, Ren R, Nagel BA, Cutter G, et al. Fertility restoration of maize CMS-C altered by a single amino acid substitution within the Rf4 bHLH transcription factor. Plant J 2020;101:101–11. https://doi.org/10.1111/tpj.14521.
[16] Scheben A, Verpaalen B, Lawley CT, Chan C-K, Bayer PE, Batley J, et al. CropSNPdb: a database of SNP array data for Brassica crops and hexaploid bread wheat. Plant J 2019;98(1):142–52. https://doi.org/10.1111/tpj.14194.
[17] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.R-project.org/.
[18] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson. shiny: Web Application Framework for R 2020. R package version 1.4.0.2. https://CRAN.Rproject.org/package=shiny.
[19] Shin J-H, Blay S, Graham J, McNeney B. LDheatmap: An *R* function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Soft 2006;16. https://doi.org/10.18637/jss.v016.c03.
[20] Xuan L, Yan T, Lu L, Zhao X, Wu D, Hua S, et al. Genome-wide association study reveals new genes involved in leaf trichome formation in polyploid oilseed rape (Brassica napus L.). Plant Cell Environ 2020;43:675–91. https://doi.org/10.1111/pce.13694.
[21] Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 2004;20(2):289–90.
[22] Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 2017;8:28–36. https://doi.org/10.1111/2041-210X.12628.
[23] Paradis E. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 2010;26(3):419–20.
[24] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer; 2016.
[25] Du X, Huang G, He S, Yang Z, Sun G, Ma X, et al. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. Nat Genet 2018;50(6):796–802. https://doi.org/10.1038/s41588-018-0116-x.
[26] Wang M, Tu L, Lin M, Lin Z, Wang P, Yang Q, et al. Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. Nat Genet 2017;49:579–87. https://doi.org/10.1038/ng.3807.
[27] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 2012;6:80–92. https://doi.org/10.4161/fly.19695.
[28] Yao W, Huang F, Zhang X, Tang J. ECOGEMS: efficient compression and retrieve of SNP data of 2058 rice accessions with integer sparse matrices. Bioinformatics 2019;35:4181–3. https://doi.org/10.1093/bioinformatics/btz186.