

TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data

Chengjie Chen^{1,2,3,4}, Hao Chen⁵, Yi Zhang^{1,2,3,4}, Hannah R. Thomas⁶, Margaret H. Frank⁶, Yehua He^{2,3,4} and Rui Xia^{1,2,3,4,*}

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

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

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

⁴College of Horticulture, South China Agricultural University, Guangzhou, Guangdong 510640, China

⁵Oilseed Crops Institute, Hunan Agricultural University, Changsha 410128, China

⁶Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA

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

<https://doi.org/10.1016/j.molp.2020.06.009>

ABSTRACT

The rapid development of high-throughput sequencing techniques has led biology into the big-data era. Data analyses using various bioinformatics tools rely on programming and command-line environments, which are challenging and time-consuming for most wet-lab biologists. Here, we present TBtools (a Toolkit for Biologists integrating various biological data-handling tools), a stand-alone software with a user-friendly interface. The toolkit incorporates over 130 functions, which are designed to meet the increasing demand for big-data analyses, ranging from bulk sequence processing to interactive data visualization. A wide variety of graphs can be prepared in TBtools using a new plotting engine (“JIGplot”) developed to maximize their interactive ability; this engine allows quick point-and-click modification of almost every graphic feature. TBtools is platform-independent software that can be run under all operating systems with Java Runtime Environment 1.6 or newer. It is freely available to non-commercial users at <https://github.com/CJ-Chen/TBtools/releases>.

Keywords: TBtools, bioinformatics, big data, data visualization, gene family

Chen C., Chen H., Zhang Y., Thomas H.R., Frank M.H., He Y., and Xia R. (2020). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant*. **13**, 1194–1202.

INTRODUCTION

The exponential growth of biological data has followed the rapid development and innovation of high-throughput sequencing (HTS) technologies. Managing big data and decoding the underlying bioinformation effectively and efficiently present a significant challenge for wet-lab biologists. Various bioinformatics software programs, pipelines, and packages have been developed to meet this challenge; however, most of these tools are packaged as scripts written in disparate programming languages and require a working knowledge of the command-line environment. This lack of easily accessible tools remains a significant obstacle for wet-lab biologists who want to process their own data but lack proficient computational skills. HTS technologies are frequently used to investigate biological phenomena on a genomic scale. Unfortunately, the big data generated are often underutilized when experimental biologists run into programming roadblocks.

When handling large data files generated by HTS, we often run across difficulties. For instance, a common task is to extract specific sequence records or take regional sequences from large Fasta files; this is difficult or often impossible to accomplish with regular text editors, especially in batch mode. Although a few scripts or software are available, for example seqkit (Shen et al., 2016) and FasPaser2 (Sun, 2018), they require users to either have a working knowledge of the command-line environment or are restricted to Windows operating systems. Heatmap is a commonly used graphic for visualizing large-scale gene expression data. Various web-based tools, such as Heatmapper (Babicki et al., 2016), allow users to generate heatmaps outside of the command-line environment; however, these programs require a stable internet connection and thus inhibit offline data analysis. Other tools like

pheatmap (<https://cran.r-project.org/web/packages/pheatmap/>) and complexHeatmap (Gu et al., 2016) are useful tools for generating various heatmap configurations but require a working knowledge of the R programming language. Gene families are groups of genes that share sequence similarity and are projected to have related functions. For comparative analyses, identification of not only sequence features, including gene structure (untranslated region [UTR], exon, and intron), conserved motifs, and potential functional domains, but also phylogenetic relationships is important to reveal the conservation and differences among members of a gene family. Tools such as GSDS2 (Hu et al., 2015) and FancyGenes (Rambaldi and Ciccirelli, 2009) have been designed to show gene structures; however, there is no single tool that is capable of simultaneously showing all of these features in a collective way. Moreover, with the increasing demand for viewing large datasets, many brilliant approaches have been developed, such as Upsetplots (Lex and Gehlenborg, 2014), Circos plots (Connors et al., 2009), and eFP browsers (Winter et al., 2007). Each of these visualization tools is packaged under different programming environments (e.g., R, Python, and Perl), which are challenging for the average wet-lab biologist to navigate. To make these analyses more complicated, many of these programs have a long list of dependencies, making software installation a major challenge. Circos, for instance, is packaged with dependency on many additional Perl modules, making the software installation intimidating for the average wet-lab biologist.

The majority of bioinformatics tools that are essential for analyzing and visualizing big biological datasets lack user-friendly interfaces that allow simple point-and-click interactions and provide interactive modification options. To conquer this deficiency, and bridge the gap between bioinformatics analyses and wet-lab data producers, we have developed and present here TBtools (a Toolkit for Biologists integrating various biological data-handling tools), a stand-alone software suite with a user-friendly interface. TBtools is an integrative and interactive toolkit with cross-platform compatibility. Our aim is to ultimately accelerate discoveries by providing an out-of-the-box solution to the data-handling dilemma of biologists.

RESULTS AND DISCUSSION

The Rationale for TBtools Development

TBtools is developed for wet-lab biologists and contains an extensive collection of functions, which are integrated into a user-friendly graphic user interface (GUI) that can be easily navigated using point-and-click icons. For each function in TBtools, we designed a GUI panel according to the most straightforward **IOS** logic, i.e., Set **I**nput Data, Set **O**utput Path if Required, and Click **S**tart Button. This interface makes the handling of big data a more pleasant and efficient experience.

TBtools is developed as an integrative toolkit. Most bioinformatics tools or scripts are developed for specific tasks, and researchers are usually required to use several, or even dozens of independent tools sequentially to get an analysis done. In TBtools, we aim to integrate all of the commonly used functions together in a Java Run Time Environment (version ≥ 1.6), which is compatible across the three major operating systems (Win-

dows, Macintosh, Linux). All of the functions can be implemented with simple point-and-click functions using a mouse. Sequential steps for data analyses and visualizations have been integrated into a single IOS workflow. Functions in TBtools are, in most cases, coded from scratch using Java, with the exception of a few cross-platform programs, including BLAST (Camacho et al., 2009), MEME (Bailey et al., 2009), HMMER (Mistry et al., 2013), MUSCLE (Edgar, 2004), trimAl (Capella-Gutiérrez et al., 2009), IQ-TREE (Nguyen et al., 2015), and ViennaRNA (Lorenz et al., 2011). To date, more than 130 functions are available in TBtools, covering the tools most commonly used for bioinformatics analyses. These include big-data preview/view, data format conversion, basic sequence management, and interactive data visualization in numerous forms that range from simple Venn diagrams to sophisticated synteny plots (Figure 1). Notably, the development of TBtools was highly collaborative and greatly motivated by the true needs of wet-lab biologists. In the past 5 years, the tool has attracted over 15 000 stable users. Many of these users actively provide informative feedback and suggestions, which has significantly enhanced the functionality, robustness, and features of TBtools.

TBtools is developed for interactive data presentation. Data visualization and presentation are indispensable parts of bioinformatics analyses. In contrast to regular graph generators, which usually produce uneditable figures, TBtools generates interactive graphs full of editable features. A newly developed plotting engine named “JIGplot” (Java Interactive Graphics) is incorporated into TBtools, enabling rapid modifications to various graph features (Figure 2). Users can easily trigger the preset functions by double-clicking on graphic elements or right-clicking to modify visual aspects, such as shape, color, angle, stroke size, and text. Different graphic panels (e.g., phylogenetic tree, sequence alignment, gene structure, heatmaps) can be conveniently combined or arranged according to the user's needs. Thus, the interactive nature of JIGplot allows users to simultaneously visualize data and prepare publishable graphs without investing extra time replotting pre-structured datasets. Another distinguishing feature of JIGplot is its unique coordinate transformation function, inspired by “The Grammar of Graphics” (Wilkinson, 2005). For example, the coordinates of graphs can be easily switched from Cartesian to polar, which allows more information to be displayed through circular plotting (Figure 2). Furthermore, all graphs prepared in TBtools can be exported in both high-resolution bitmap and vector formats to provide the user with maximum flexibility.

Overview of Functions in TBtools

All the functions in TBtools are grouped under five main catalogs (Figure 1). (1) Sequence Toolkits. This includes all functions related to sequence management, such as sequence extraction, format conversion, and open reading frame prediction. (2) BLAST. A simplified BLAST wrapper that enables BLAST to be implemented in TBtools (Camacho et al., 2009) and allows sequence comparison on a local machine with optimized settings. In addition, functions facilitating the visualization and management of BLAST results are provided under this menu. (3) GO&KEGG. Enrichment analysis is a common approach to investigate the biological significance of specific gene sets. In TBtools, we have developed several functions for enrichment

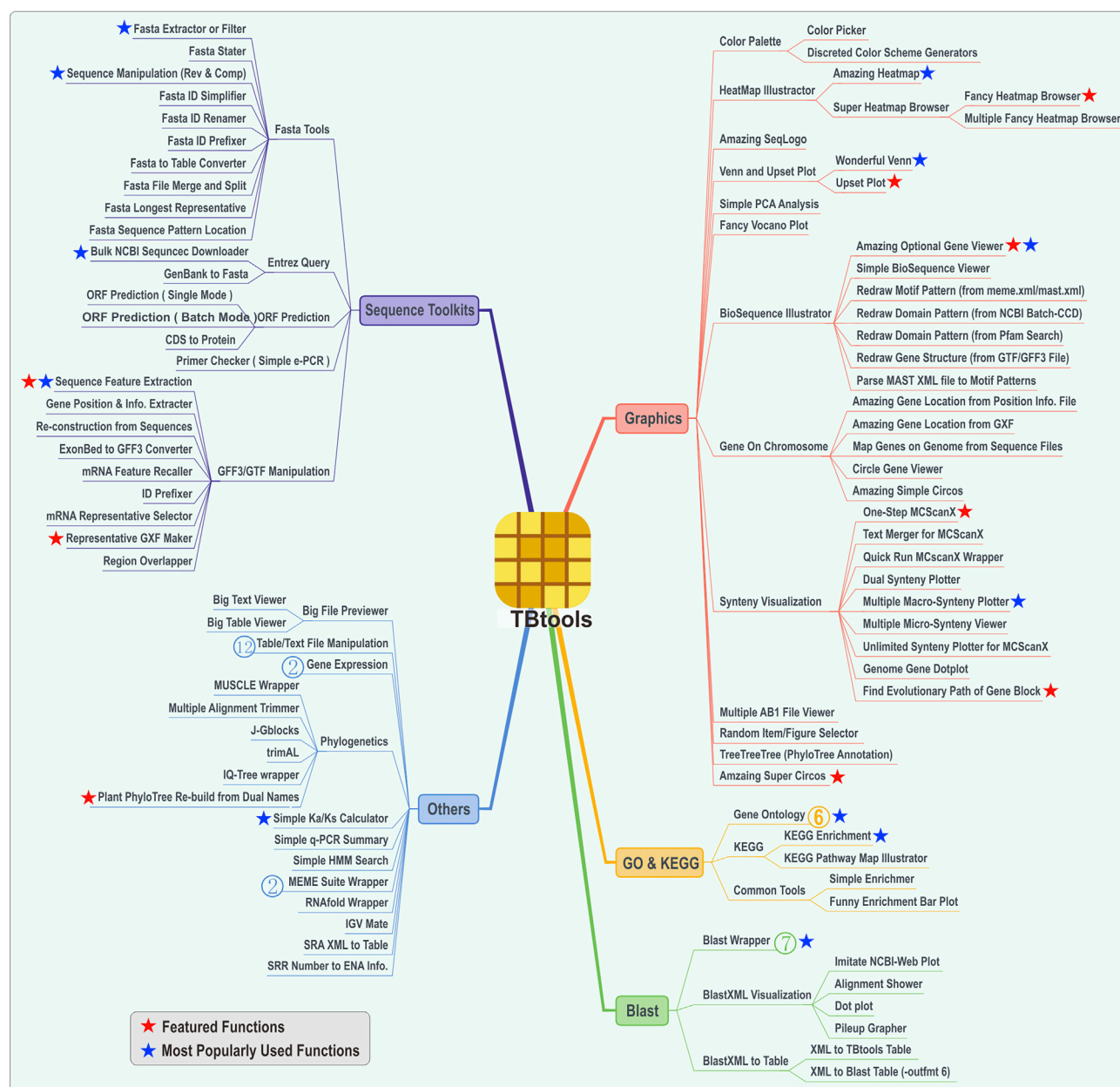


Figure 1. Outline of Functions in TBtools.

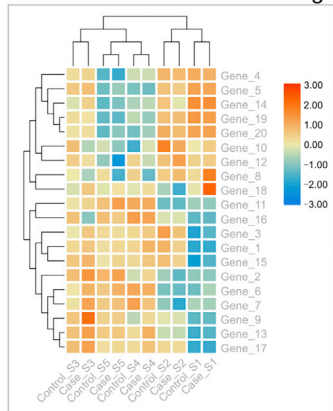
Five core function catalogs in TBtools contain over 130 functions. Featured functions that are most popularly used are highlighted with stars. Numbers in circles denote the number of subfunctions contained under each function.

analysis of gene ontologies (GO) (Ashburner et al., 2000) and KEGG pathways (Kanehisa and Goto, 2000). Result files can be quickly visualized with an easy-to-use barplot function, and subgroups of genes annotated to each specific GO/KEGG term can be extracted for further analyses. (4) Graphics. A wide variety of graphs can be rapidly prepared in TBtools, including Venn diagrams, heatmaps, Circos graphs (Connors et al., 2009), and eFP figures (Winter et al., 2007). Most graphic features (e.g., color, shape, label) can be personalized according to individual preference. (5) Other. The Other class contains many additional functions that are commonly used during big-data exploration. It includes tools used for previewing big files, editing phylogenetic trees, and calculating K_a/K_s values.

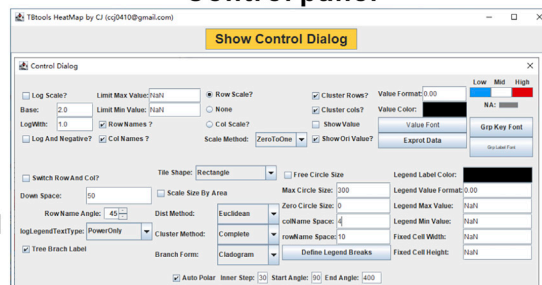
Stellar Functions

Functions in TBtools have been improved, expanded, and optimized based on demand and feedback from our user base, which includes over 15 000 stable users worldwide, many of whom are actively involved in the improvement of TBtools. A series of stellar functions are widely used by users; the results and/or graphs prepared using TBtools have been featured in hundreds of peer-reviewed publications. These functions include eFP Browser, Interactive Heatmap, Simple Circos, Gene Family Tool, and many more (Figure 3). Although most of these functions were not originally invented as part of TBtools, they have been optimized, upgraded,

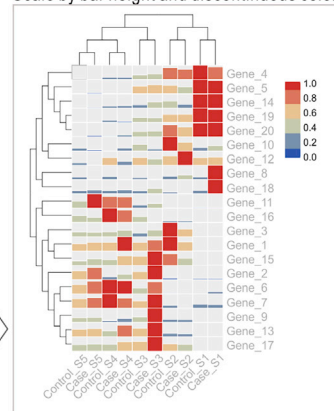
Normalization and clustering



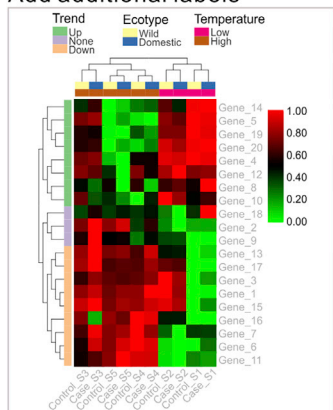
Control panel



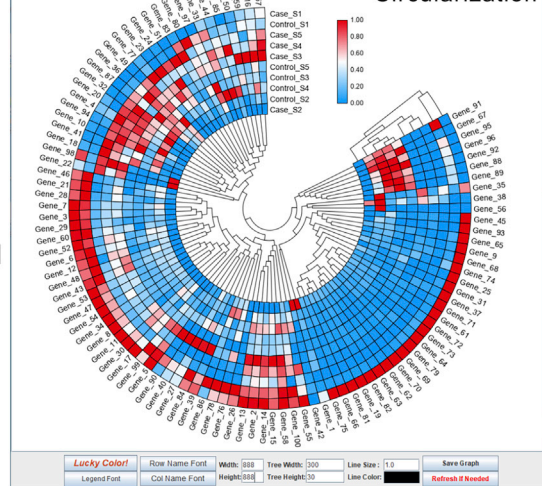
Scale by bar height and discontinuous colors



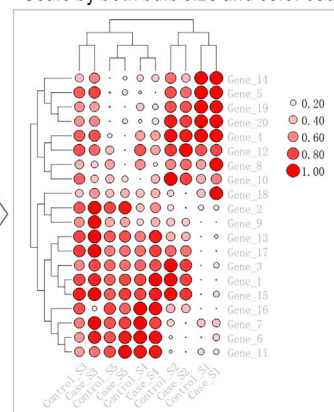
Add additional labels



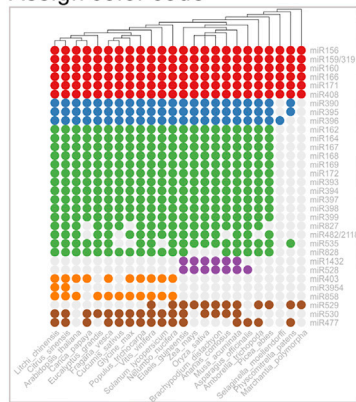
Circularization



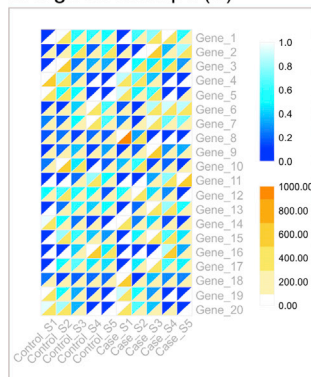
Scale by both bulb size and color code



Assign color code



Merge heatmaps (1)



Merge heatmaps (2)

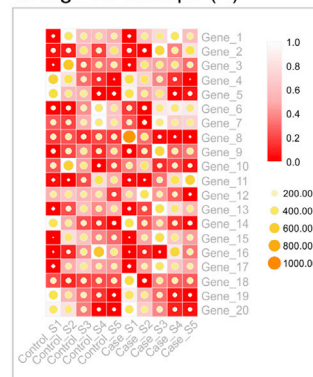


Figure 2. The Powerful Plotting Engine “JIGplot” in TBtools Displays Great Interactability.

A variety of heatmaps generated by TBtools are shown to demonstrate its great interactive ability and functional diversity.

and simplified according to the philosophy that "simple is best," ensuring their accessibility to a wide variety of biologists.

In summary, HTS technologies have generated vast amounts of biological data. For efficient and effective handling of these big data, we have developed TBtools, a user-friendly toolkit that integrates a large number of functions with an emphasis on bulk data processing and visualization. Its robustness has been validated

by tens of thousands of users, making it a handy and useful toolkit for biologists.

METHODS

Tutorials and Examples

Big File Preview/View

HTS technologies and bioinformatics analyses typically generate large data files that can range in size from hundreds of megabytes to several gigabytes.

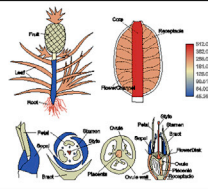
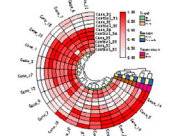
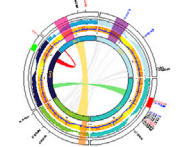
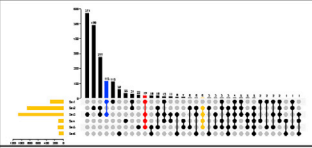
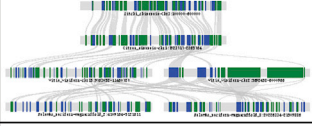
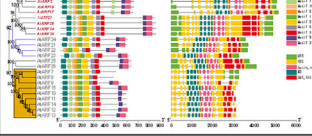
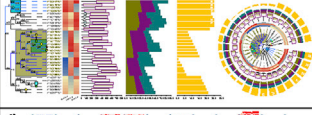
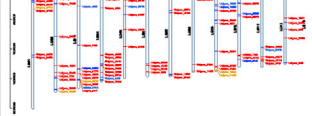
	Function	Description	Exemplative graph
1	eFP Browser	eFP Browser provides a straight-forward solution to visualize expression data onto a pictographic representation of an organism (e.g. plant or animal). It is particularly useful to provide an overview of gene expression levels in many different tissues or under different conditions.	
2	Interactive Heatmap	A wide variety of practical features in Interactive Heatmap distinguish it from many other tools, including subfunctions like interactive edition, annotation addition, graph circularization, and graph combination (Fig. 2)	
3	Simple Circos	Circos plots are common graphs used for comparative genomic analyses, but drawing circos plot with perl-Circos in command-line environment is challenging for wet-lab biologists. Simple Circos provides a simple way enabling users to make Circos plot easily.	
4	Upset Plot	Upset plot is used for the visualization of intersecting datasets, especially when the quantity of datasets is more than six, which is hard to show with regular Venn diagram.	
5	Synteny Browser	Synteny shows the conservation of blocks of genes within two sets of chromosomes that are being compared with each other. Synteny Browser enables the parallel visualization of two or more syntenic blocks which are likely derived from a single ancestral genomic region.	
6	Gene Family Tool	TBtools provides a compelling function for the presentation of phylotree, motif/domain pattern, and exon/intron gene structures simultaneously. It has been widely used in the community.	
7	Tree Annotation	Phylogenetic trees can be viewed and edited according to personal needs, like tip label, branch highlight, branch coloration. The tree can be easily combined with other graphs, and graph circularization is also supported.	
8	Genes on Chromosome	The Genes on Chromosome function maps genes or genomic features (e.g. molecular marker) onto chromosomes according to their genomic positions. Text labels or shapes are supported. Syntenic relations could also be indicated by connecting lines.	
9	Bulk sequence processing	A set of tools are provided in TBtools for huge data file processing, such as big file previewer, bulk sequence extraction/filtering, sequence/ID manipulation, and data format conversion.	
10	GFF3/GTF file manipulation	Genome annotations are always stored in files in GFF3/GTF format. Several functions have been developed for an easy usage of GFF3/GTF files for quick sequence extraction of certain regions (e.g., promoter), or for other purposes.	

Figure 3. A List of the Most Commonly Used (Top 10) Functions in TBtools.

These files are cumbersome and nearly impossible to handle using ordinary office software. Currently, biologists have to navigate through a command-line environment to take a quick look at the content, structure, and quality of these files. We have created a function that allows users to look at large data files in an interactive format: the “Big File Preview/View” function in TBtools (Figure 4A). Detailed steps are described as follows.

1. Navigate to the “Big Text View” function: **Others** → **Big File Preview** → **Big Text View** (Figure 4A).
2. Drag and drop a file to the panel or click the panel to invoke the file selection option.
3. The content of the input file will be instantly displayed in the window (Figure 4A).

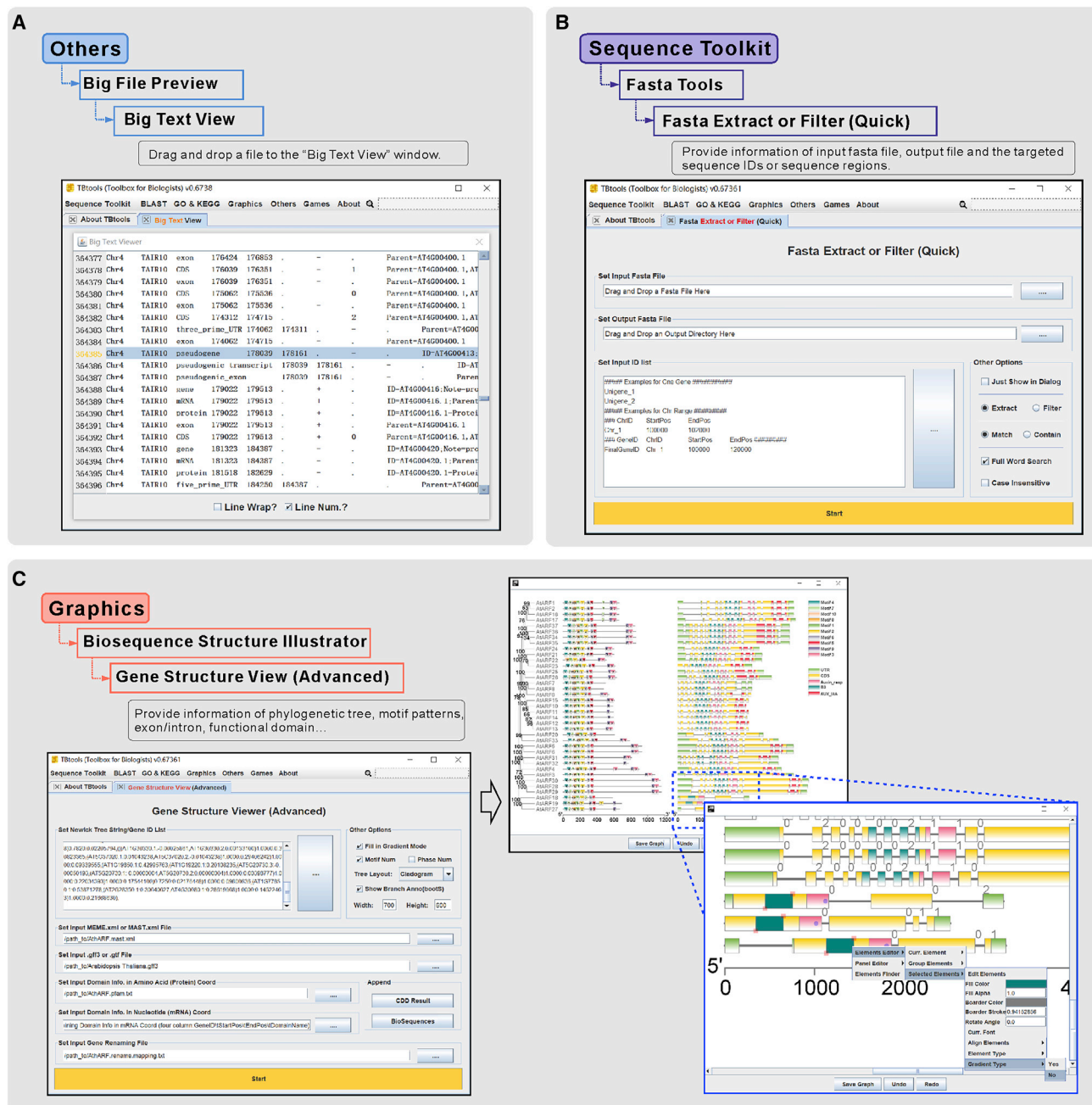


Figure 4. Demonstration of the "Big File Preview/View," "Fasta Extract," and "Gene Structure View" Functions in TBtools.

Large data files are easy to navigate and quickly preview using the "Big File Preview/View" function (A); sequences can be searched and extracted from Fasta files using the "Fasta Extract" function (B); and structural features and phylogenetic relationships among gene family members are easy to visualize using the "Gene Structure View" function (C).

Bulk Fasta Extract

Fasta sequence extraction is a common activity in biological data management. Several functions of TBtools are developed for bulk Fasta sequence extraction. For instance, the "Fasta Extract or Filter (Quick)" function can be employed to extract or filter out sequences according to a list of identifiers.

1. Navigate to the "Fasta Extract or Filter (Quick)" function: **Sequence Toolkit** → **Fasta Tools** → **Fasta Extract or Filter (Quick)** (Figure 4B).
2. Set an input sequence file in Fasta format.

3. Set the path for an output file.
4. Provide a list of sequence identifiers or regions of sequences.
5. Click "Start" and the selected sequences will be extracted into the assigned output file (Figure 4B).

Gene Family Tools

Gene structure (i.e., UTR, exon, intron) and functional sequence elements (i.e., sequence motifs and protein domains) are important features of genes and proteins. TBtools provides an integrated function for visualizing these features. To do this, the user needs to:

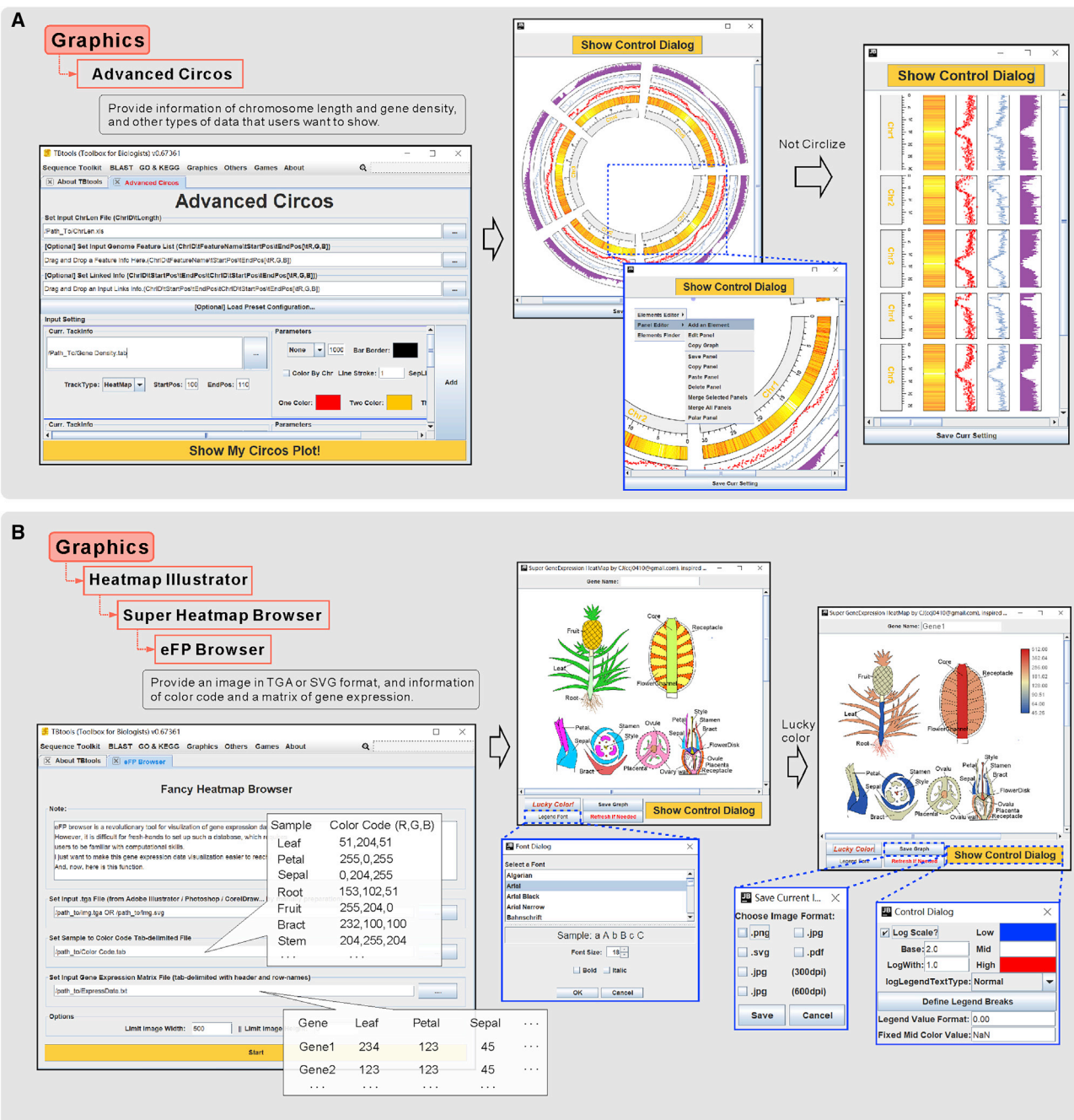


Figure 5. Examples Demonstrating the “Advanced Circos” and “eFP Browser” Functions in TBtools.

Flexible Circos graphics can be quickly generated using the “Advanced Circos” function (A); organ-specific expression can be plotted for any species of interest using the “eFP Browser” function (B).

1. Navigate to the “Gene Structure View (Advanced)” function: **Graphics** → **BioSequence Structure Illustrator** → **Gene Structure View** (Figure 4C).
2. Upload either a gene phylogeny in Newick format or a set of gene identifiers.
3. Select an XML file storing motif pattern information; this can be generated using the MEME suite (Bailey et al., 2009).
4. Select a gene-structure annotation file in GFF3/GTF format.
5. Select a file containing functional domain information with protein coordinates, formatted as four tab-delimited columns: “Gene

ID,” “Start Pos,” “End Pos,” and “Name of Domain.” This file can be prepared using CD-search (Marchler-Bauer et al., 2011).

6. Select a file containing identifier renaming information, formatted as two tab-delimited columns, “Original ID” and “New ID.”
7. Click “Start” to generate a gene structure plot with a phylogenetic tree, motif patterns, and conserved domains (Figure 4C). Each graphic element can be modified according to personal preference.

Note: Steps 2 to 6 are all optional depending on the desired user output.

Circos Plot

Circos plots are widely used to visualize large datasets on a whole-genome scale (Connors et al., 2009). TBtools provides a convenient function to generate Circos plots (Figure 5A). The required input files include:

1. A genome sequence file in Fasta format
2. A gene-structure annotation file in GFF3/GTF format

Users can either manually prepare a file containing information regarding chromosome lengths or use TBtools to generate this file. To generate the chromosome size file in TBtools:

3. Navigate to the “Fasta Stats” function: **Sequence Toolkit** → **Fasta Tools** → **Fasta Stats**.
4. Set the above genome sequence file as input.
5. Set the path for an output file.
6. Click “Start” and the chromosome length information will be saved in the output file.
7. Open the output file with Microsoft Excel or an analogous software program, keep the two columns listing chromosome identifiers and lengths, and save.

To initiate a basic Circos plot:

8. Navigate to the “Advanced Circos” function: **Graphics** → **Advanced Circos** (Figure 5A).
9. Set the file containing the chromosome lengths as input.
10. Click “Start” and generate a Circos plot frame.

To prepare a file containing genomic gene density information:

11. Navigate to the “Gene Density Profile” function: **Sequence Toolkit** → **GFF3/GTF Manipulate** → **Gene Density Profile**.
12. Set the above gene structure annotation file (the GFF3/GTF file) as input.
13. Set the path for an output file (named “Gene Density.tab” in this example).
14. Click “Start” and the gene density information will be saved into the output file (“Gene Density.tab”).

To add gene density tracks to the initial Circos plot:

15. Starting from the “Advanced Circos” window (setup in steps 8–10), click “Add” to add a new data track.
16. Set the gene density file (“Gene Density.tab”) as input.
17. Select the track type, for example, “Heatmap,” “Point,” “Line,” or “Bar.”
18. (Optional) Change the track color as needed.
19. To add additional data tracks, repeat steps 15 to 18.
20. Click the “Show My Circos Plot!” button to instantly generate a user-defined Circos plot (Figure 5A). Additional features can be added by right-clicking on the plot, and numerous aesthetic modifications are available using the pop-up menu.

eFP Graph

Development of the eFP function in TBtools was inspired by the original eFP Browser (Winter et al., 2007). This function is useful when users want to present gene expression data in a personalized, pictographic way (Figure 5B).

1. Prepare an image in TAG or SVG format, in which different tissues or organs can be color-coded.
2. Prepare a color code file, formatted with two columns, “Sample” and “Color Code.”

Note: the “Color Picker” function can be employed to generate the color code file.

3. Prepare a file containing a matrix of gene expression values; this is a standard result from a transcriptome analysis.
4. Navigate to the “eFP Browser” function: **Graphics** → **Heatmap Illustrator** → **Super Heatmap Browser** → **eFP Browser** (Figure 5B).
5. Set input files as described in the notes for each field.
6. Click “Start” and an initial graph will be generated.
7. Enter a gene identifier and press Enter to explore gene expression in a pictographic format. The color scheme can be user curated using the pop-up menu, or the “Lucky Color” button in the panel can be used to quickly flip through different color schemes (Figure 5B).

FUNDING

This work was funded by the National Key Research and Developmental Program of China (2018YFD1000104). This work is also supported by awards to R.X., Y.H., and H.C. from the National Key Research and Developmental Program of China (2017YFD0101702, 2018YFD1000500, 2019YFD1000500), the National Science Foundation of China (#31872063), the Special Support Program of Guangdong Province (2019TX05N193), the Key-Area Research and Development Program of Guangdong Province (2018B020202011), and the Guangzhou Science and Technology Key Project (201804020063). Support to M.H.F. comes from the NSF Faculty Early Career Development Program (IOS-1942437).

AUTHOR CONTRIBUTIONS

C.C., H.C., Y.H., and R.X. conceived the project; C.C. and R.X. designed the functions of the toolkit; C.C. performed all the Java coding. Y.Z., H.R.T., and M.H.F. tested the functions and helped with the preparation of the tutorial manual. C.C., M.H.F., and R.X. prepared the figures and wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Drs. Jian Ren and Qi Zhao for their encouragement of using Java for TBtool development, Dr. Guangchuang Yu for his inspirations on computational graph visualization and Ya Wu for her help on TBtools’ logo and GUI designing. We thank all labmates in the Xia lab and He lab for their generous help. We are also grateful for the kind advice from 15,000+ TBtools users, especially the >30 advanced users. No conflict of interest declared.

Received: February 29, 2020

Revised: April 4, 2020

Accepted: June 10, 2020

Published: June 22, 2020

REFERENCES

- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**:25–29.
- Babicki, S., Arndt, D., Marcu, A., Liang, Y., Grant, J.R., Maciejewski, A., and Wishart, D.S. (2016). Heatmapper: web-enabled heat mapping for all. *Nucleic Acids Res.* **44**:W147–W153.
- Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**:202–208.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421.
- Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973.

- Connors, J., Krzywinski, M., Schein, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A.** (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**:1639–1645.
- Edgar, R.C.** (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**:1792–1797.
- Gu, Z., Eils, R., and Schlesner, M.** (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**:2847–2849.
- Hu, B., Jin, J., Guo, A.Y., Zhang, H., Luo, J., and Gao, G.** (2015). GSDD 2.0: an upgraded gene feature visualization server. *Bioinformatics* **31**:1296–1297.
- Kanehisa, M., and Goto, S.** (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**:27–30.
- Lex, A., and Gehlenborg, N.** (2014). Sets and intersections. *Nat. Methods* **11**:779.
- Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L.** (2011). ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**:26.
- Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., et al.** (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**:225–229.
- Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M.** (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**:e121.
- Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q.** (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**:268–274.
- Rambaldi, D., and Ciccarelli, F.D.** (2009). FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics* **25**:2281–2282.
- Shen, W., Le, S., Li, Y., and Hu, F.** (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**:e0163962.
- Sun, Y.B.** (2018). FasParser2: a graphical platform for batch manipulation of tremendous amount of sequence data. *Bioinformatics* **34**:2493–2495.
- Wilkinson, L.** (2005). *The Grammar of Graphics* (New York: Springer-Verlag).
- Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., and Provat, N.J.** (2007). An “electronic fluorescent pictograph” Browser for exploring and analyzing large-scale biological data sets. *PLoS One* **2**:e718.