TBtools - an integrative toolkit developed for interactive analyses of big biological data

Chengjie Chen[1,2,3], Hao Chen[4], Yi Zhang[1,2,3], Hannah R. Thomas[5], Margaret H. Frank[5], Yehua He[2,3], Rui Xia[1,2,3*]

[1]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, [2]Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in South China, Ministry of Agriculture, [3]College of Horticulture, South China Agricultural University, Guangzhou 510642, China

[4]Oilseed Crops Institute, Hunan Agricultural University, Changsha 410128, China

[5]Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA

*To whom correspondence should be addressed: rxia@scau.edu.cn

Abstract

The rapid development of high-throughput sequencing (HTS) 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 100 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, with a new plotting engine ("JIGplot") developed to maximum their interactive ability, which allows quick point-and-click modification to almost every graphic feature.

TBtools is a 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.

The Rationale for TBtools development

Developed for wet-lab biologists

The exponential growth of biological data has come with the rapid development and renovation of high-throughput sequencing (HTS) techniques. Managing big-data and decoding the underlying bio-information effectively and efficiently presents a significant challenge to wet-lab biologists. Various bioinformatics software, 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 is often underutilized when experimental biologists run into programming roadblocks. Here, we present TBtools, a Toolkit for Biologists integrating various biological data handling tools with a user-friendly interface; out aim is to accelerate discoveries by providing an out-of-the-box solution to the data-handling dilemma of biologists. TBtools contains an extensive collection of functions, which integrate into a graphic user interface (GUI) that can be easily navigated using point-and-click icons. For each function in TBtools, we designed its GUI panel according to the most straightforward IOS logic, i.e., Set Input Data, Set Output Path if Required and Click Start Button. This interface makes the handling of big-data a more pleasant and efficient experience.

Developed as an integrative toolkit

In order to handle large biological data, researchers are currently required to work under a command-line environment, and use several, even dozens of independent tools sequentially. For instance, to identify homologous genes of a specific gene family from a species, users must access genomic data which are commonly available as two separate files, a genome sequence file in FASTA format and a gene structure annotation file in GFF3/GTF format; then they should firstly invoke “gffread” (Trapnell et al., 2013) which only works under Unix-like

operating system to retain gene sequences, secondly apply BLAST program (Camacho et al., 2009) to get an ID list of homologous genes, and finally use home-brew scripts or other tools like “seqkit” (Shen et al., 2016) to extract sequences of homologous genes. In TBtools, we aim to integrate all of these functions in a Java Run Time Environment (version >= 1.6), which is compatible across the three major operating systems (Windows, Machintosh, Unix). All of the functions can be implemented through simple point-and-click 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 or achieved by invoking cross-platform programs (e.g., BLAST). To date, more than 100 functions are available in TBtools, covering the most commonly used tools for bioinformatic analyses. These include big-data preview, data format conversion, basic sequence management, interactive data visualization in numerous forms that span from simple Venn diagrams to sophisticated synteny plots (Fig. 1). Notably, the development of TBtools was highly collaborative and greatly motivated by the true needs of wet-lab biologists. In the past five 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 and features of TBtools.

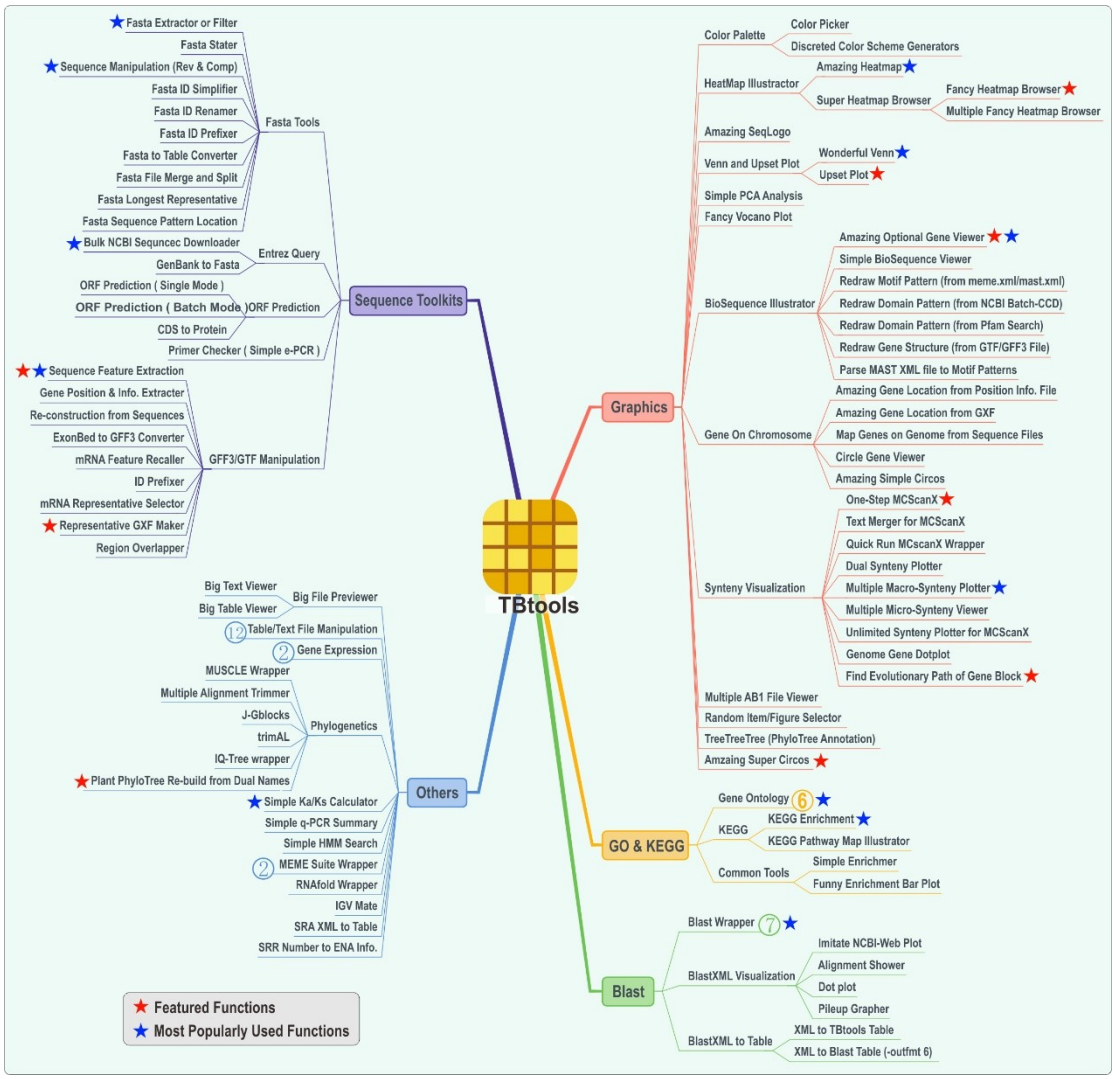


Figure 1. Functional outline of TBtools.

There are five core catalogs that contain over 100 functions in TBtools. Featured that are most popularly used are highlighted with stars. Numbers in circles denote the number of subfunctions contained under each function.

Developed to interactively present data

Data visualization and presentation are indispensable parts of bioinformatic 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 (Fig. 2). Users can easily trigger the preset functions by double-clicking on graphic elements or right-click to modify visual aspects, such as color,

shape, stroke size, text, etc. 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 re-plotting pre-structured datasets. Another distinguishing feature of JIGplot is its unique function for coordinate transformation. For example, the coordinates of graphs can be easily switched from cartesian to polar, which allows more information to be displayed through circular plotting (Fig. 2). Besides, all graphs prepared in TBtools can be exported in both high-resolution bitmap and vector formats to allow maximum flexibility for the user's end.

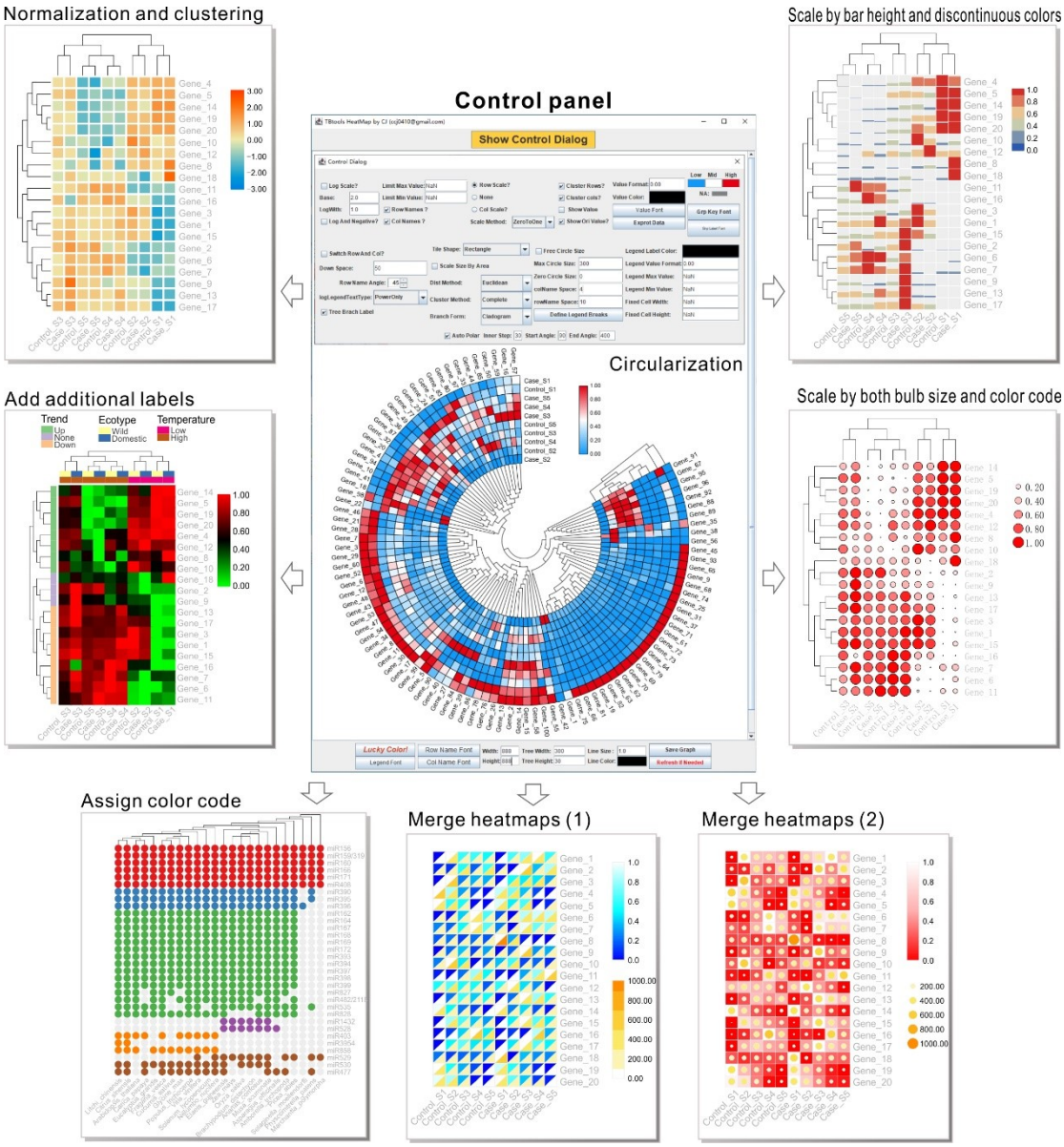


Figure 2. The powerful plotting engine 'JIGplot' displays great interactability in TBtools

A variety of heatmaps generated by TBtools are used for the demonstration of its great interactive ability and functional diversity.

Overview of TBtools

Function outline

All the functions in TBtools are grouped under five main catalogs (Fig. 1): (1) Sequence Toolkits. This includes all functions related to sequence management, such as sequence extraction, format conversion, and ORF prediction. (2) BLAST. A simplified BLAST wrapper was implemented in TBtools (Camacho et al., 2009); this allows sequence comparison on a local

115 machine with optimized settings. Functions facilitating BLAST result visualization and
116 management are also provided under this menu. (3) GO&KEGG. Enrichment analysis is a
117 common approach to investigate the biological significance of specific gene sets. In TBtools,
118 we have developed several functions for enrichment analysis of gene ontologies (GO)
119 (Ashburner et al., 2000) and KEGG pathways (Kanehisa and Goto, 2000). Result files can be
120 quickly visualized with an easy-to-use bar plot function, and sub-grouped genes to each
121 specific GO/KEGG term can be extracted for further analyses. (4) Graphics. A wide variety of
122 graphs can be rapidly prepared in TBtools, including Venn diagrams, heatmaps, Circos graphs
123 (Connors et al., 2009), eFP figures (Winter et al., 2007), and so on. Most graphic features (e.g.,
124 color, shape, label) can be personalized according to individual preference. (5) Others. the
125 Other class contains many additional functions that are commonly used during big-data
126 exploration. It includes tools used for previewing big files, editing phylogenetic trees, and
127 calculating Ka/Ks values.

128 ## Stellar functions

129 Functions in TBtools have been improved, expanded, and optimized based on demand and
130 feedback from our user base which includes over 15,000 stable users worldwide, many of
131 whom are actively involved in the improvement of TBtools. A series of stellar functions are
132 widely used by users; the results and/or graphs prepared using TBtools have been featured
133 in hundreds of peer-reviewed publications. These functions include eFP Browser, Interactive
134 Heatmap, Simple Circos, Gene Family Tool, and many more (Fig. 3). Although most of these
135 functions were not originally invented in TBools, they are optimized, upgraded, and simplified
136 according to the philosophy of "simple is best", ensuring accessible usage for biologists.
137

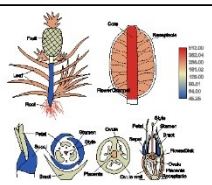
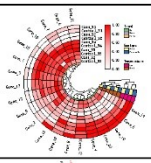
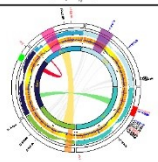
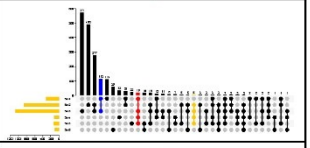
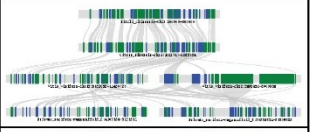
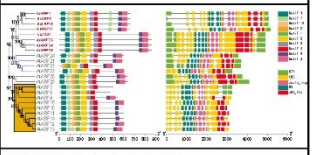
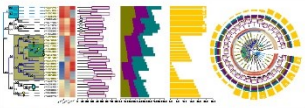
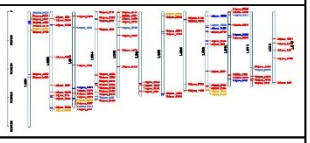
	Function	Description	Exemplative graph
1	eFP Browser	eFP Browser provides a straight-forward solution to visualize expression data onto a pictographic representation of an organism (e.g. plant or animal). It is particularly useful to provide an overview of gene expression levels in many different tissues 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 description of prevalently used (top 10) functions in TBtools

Conclusion

High-throughput sequencing techniques have generated vast amounts of biological data. For efficient and effective handling of this data, we have developed TBtools, a user-friendly toolkit integrated with 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.

Acknowledgments

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) and the Special Support Program of Guangdong Province (2019TX05N193). Support to M. H. F. comes from the NSF Faculty Early Career Development Program (IOS-1942437). 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.

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.

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.

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.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**:27–30.

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**:0163962.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**:46–53.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G. V., and Provart, N. J. (2007). An "electronic fluorescent pictograph" Browser for exploring and analyzing large-scale biological data sets. *PLoS One* **2**:e718.