Open access • Posted Content • DOI:10.1101/289660

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

Chengjie Chen, Hao Chen, Yi Zhang, Hannah R. Thomas ...+3 more authors

Institutions: South China Agricultural University, Hunan Agricultural University, Cornell University

Published on: 09 Mar 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Interface (Java), Data visualization and Biological data

TBtools, a Toolkit for Biologists integrating various biological data handling tools with a user-friendly interface

Chengjie Chen[1,2,3], Hao Chen[4], 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

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

Abstract

Summary: Rapid development of high-throughput sequencing (HTS) techniques has led biology into the "big-data" era. Data analysis using various bioinformatics softwares or pipelines relying on programming and command-line environment is challenging and time-consuming for most wet-lab biologists. Bioinformatics tools with a user-friendly interface are preferred. Here, we present TBtools (a Toolkit for Biologists integrating various biological data handling tools), a stand-alone software with a user-friendly interface. It has powerful data handling engines for both bulk sequence processing and interactive data visualization. It includes a large collection of functions, which may facilitate much simple, routine but elaborate work on biological data, such as bulk sequence extraction, gene set enrichment analysis, Venn diagram preparation, heatmap illustration, comparative sequence visualization, etc.

Availability and implementation: 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.

Contact: rxia@scau.edu.cn

Supplementary information: Detailed instruction is provided with the software as well at https://github.com/CJ-Chen/TBtools/releases

Introduction

Exponential growth of biological data comes with the rapid development and renovation of high-throughput sequencing (HTS) techniques. Various bioinformatics softwares, pipelines, and packages have been developed to meet all kinds of analysis requests of biologists. Most of these tools are packed small scripts written in various programming languages, which are not easy-to-use for most biologists with limited computation knowledge. Some tools are web-based applications, of which efficient usage and stability rely much on internet infrastructure. A few softwares with a graphic user interface (GUI) have greatly assisted biologists in certain data analyses, such as the widely used Venny for Venn diagram drawing (Oliveros, J.C, 2007), HemI for heatmap illustration (Deng *et al.*, 2014) and GSDS2 for gene feature visualization (Hu *et al.*, 2015). However, none of a toolset integrates all these commonly used functions together. Moreover, with the fast increase of biological data, routine work of biologists, such as sequence extraction, gene annotation, file format conversion etc., becomes time-consuming and tedious. An integrated tool with emphasis on routine biological data management is in high demand as well.

Therefore, we present TBtools, a toolkit developed to save time and energy of biologists from various tasks of data analyses. It has a user-friendly interface (Fig. 1A) with two main classes of functions, bulk sequence processing and powerful data visualization. The former mainly includes batch sequence management, stand-alone BLAST wrapper, and gene set enrichment analysis; the latter contains Venn diagram preparation, heatmap illustration, comparative sequence visualization, etc. The toolkit has also incorporated with many easy-to-use features, such as drag-and-drop file input, flexible graph resizing, color change with a color picker, easy element shape manipulation, and quick graph exportation in both high-resolution bit-map and vector formats. TBtools will be a handy toolkit for biologists to work with all kinds of HTS data.

Key Features

1 Bulk Sequence Processing

1.1 Sequence Extraction and Manipulation

Sequence extraction and manipulation are routine jobs in biological data analyses. We developed a series of functions to facilitate such kind of work in TBtools. With small computation resources, users can quickly obtain statistical information of sequences, extract bulk sequences from huge genomes, modify sequence identifiers in a batch mode, split or merge sequence files and manipulate nucleotide sequences (like Reverse, Complement, DNA to RNA), etc.

Hundreds of genomes have been sequenced and published so far. Genome sequences (usually in FASTA format) and gene-structure annotation files (in GTF/GFF3 format) are easily accessible. However, it is complicated and tedious to extract sequences according to GTF/GFF3 annotation files from genome sequences. TBtools has been incorporated with a function to extract all kinds of sequences based on annotation features, such as gene, mRNA, CDS, UTR, and upstream (before the 5' UTR or start codon, also considered as "promoter region") or downstream (after the 3' UTR or stop codon) sequences with a given length specified by users.

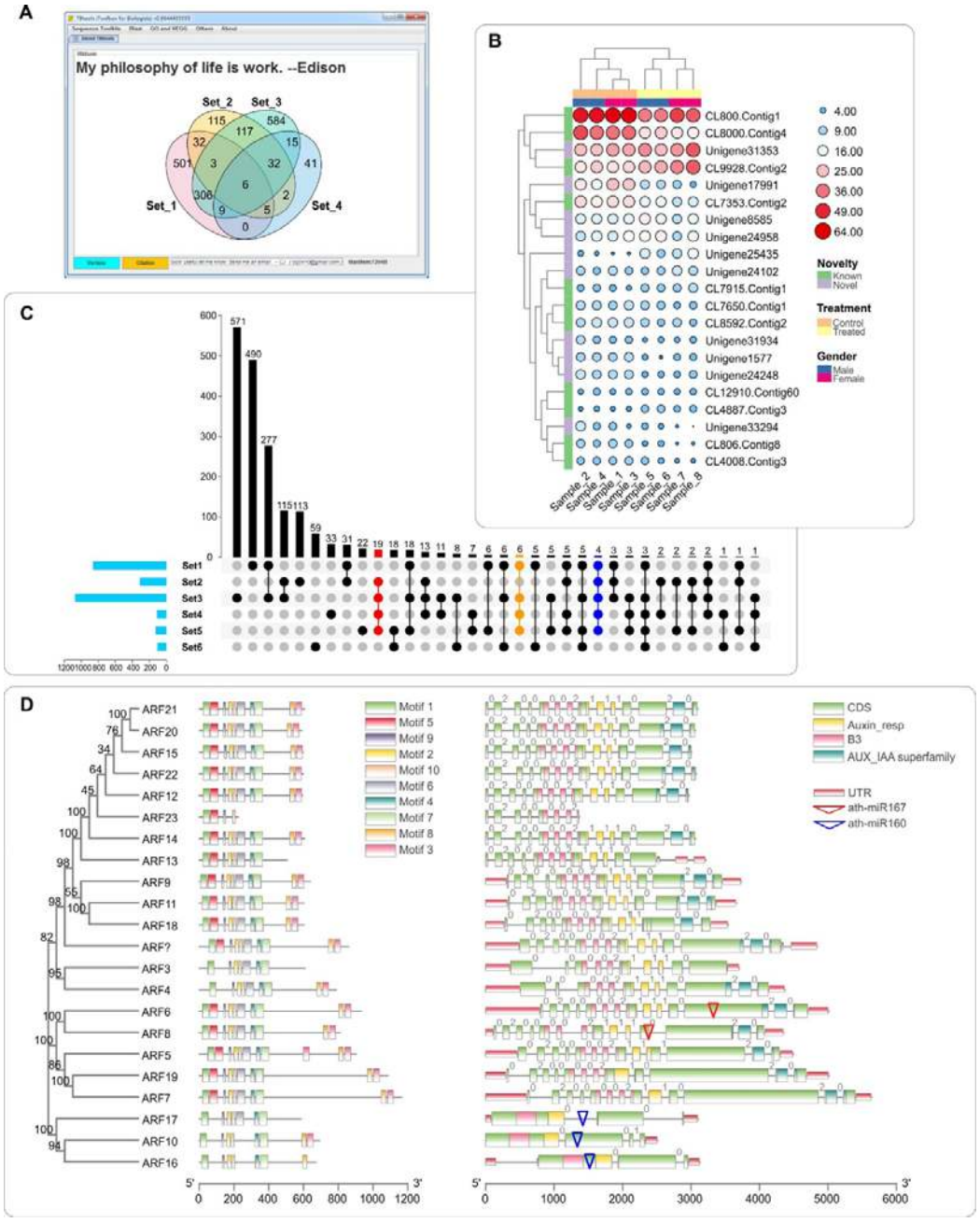

Figure 1. Interface of TBtools and representative graphs produced by the toolkit. A. The main interface of TBtools and an example of Venn diagram from four datasets. Users can obtain a gene list of an intersection by double clicking on given numbers. Right-click menu allows easy color change. B. An example of UpSet plot generated by TBtools. Color change and exportation of intersection gene lists are also supported. C. An example of heatmap generated by TBtools, with annotation information of rows and columns included. Value matrix is represented by both continuous color scheme and circle size. D. A graph simultaneously showing phylogenetic tree, MEME motifs, protein domains, gene-structures, and miRNA target sites of *ARF* (*Auxin Response Factors*) genes in *Arabidopsis*.

1.2 Local BLAST Wrapper and Result Visualization

Local BLAST package makes biologists capable of performing sequence comparison in a batch mode on a local machine. Several GUI wrappers incorporated with local BLAST have been reported, like BioEdit (Hall *et al.*, 2011). They often ask users to specify most alignment and filter parameters, which are usually unfamiliar to wet-lab biologists. In TBtools, we have implemented a simplified BLAST wrapper. Instead of asking for complicated parameters, we minimize users' operation in most cases. For example, automatic detection of molecular type is applied to avoid the step of choosing among different BLAST algorithms. In addition, TBtools also has several functions to visualize BLAST results or transform the format of result files from XML to tab-delimited text.

1.3 Gene Set Enrichment Analysis

In the biological big-data era, gene set enrichment analysis is a common approach to investigate biological significance of specific gene sets, for instance, differentially expressed genes. However, it often requires biologists to work with programs under command line environment or via web. We have developed several functions for enrichment analysis of gene ontology and KEGG pathway. In result files, corresponding genes annotated to each specific ontology term can be easily obtained for further analyses. Users can also quickly visualize results with an easy-to-use bar plot function in TBtools.

2 Powerful Interactive Graphics

2.1 Venn Diagram and Upset Plot

Various web-running applications or small tools have been developed for the preparation of Venn diagrams. In TBtools we have made Venn diagram drawing easier and more convenient. Users can either drag-and-drop input text files or paste text into TBtools and easily obtain an interactive Venn diagram (Fig. 1A). Currently, Venn diagram supports up to six datasets, but those with more than four datasets are hard to interpret. To avoid this limitation, we have adopted UpSet plot in TBtools as well (Lex and Gehlenborg, 2014) (Fig. 1C). The UpSet plot allows the interpretation of unlimited sets of data in a single plot, and the function for UpSet plot can be used the same as for the Venn diagram. Users can easily obtain interested gene list of any intersection simply by double-clicking on corresponding text in Venn diagrams or bars in UpSet plots. One-time exportation of gene lists of all intersections is supported as well.

2.2 Heatmap Illustrator

Heatmap is another common graph widely used in large data presentation. There are a few applications developed for heatmap preparation, for instance, HemI (Deng *et al.*, 2014) with a use-friendly graphic interface. In TBtools, we have developed a heatmap illustrator with a few renovations (Fig. 1C). Several color schemes are pre-set and users can change color easily using a color picker. Additionally, users can also employ circle sizes to represent values, which is useful and friendly to color-blind readers (Fig. 1C). In TBtools, users can also add annotation information to columns and rows and incorporate pre-calculated tree structure. Data can be exported according to the structure of the data matrix of the clustered heatmap.

2.3 Comparative Sequence Visualization

Comparative sequence analyses of gene structures and functional motifs/domains are often applied for studies of gene families. FancyGenes (Rambaldi and Ciccarelli, 2009) and GSDS2 (Hu *et al.*, 2015) have been commonly used for the comparative sequence visualization. Both of them are web-based applications, which rely much on internet infrastructure. The former is out of service and not maintained anymore, and the latter has a limitation of maximum gene number and asks for data in specific formats. In TBtools, we have incorporated a series of functions for comparative sequence visualization. Input files in various formats are acceptable, including GTF/GFF file of gene annotation, XML file from MEME/MAST suite (Bailey *et al.*, 2009), and tab-delimited text files. Conversion from amino acid coordinates (for proteins) to nucleotide coordinates (for mRNA/exon) is automatically implemented in the toolkit. In addition, a phylogenetic tree can be added in parallel. Using this visualization function, users can generate an integrated graph presenting phylogenetic tree, motif/domain patterns, gene structures, and miRNA target sites (Fig. 1D). If a valid GFF3 file is not available, TBtools can generate a home-made GFF3 file according to user-provided sequences of mRNAs and corresponding genome.

Conclusion

High-throughput sequencing techniques have generated huge amount of biological data. For efficient and effective handling of those huge data, we have developed TBtools, a user-friendly toolkit integrated with a large number of functions with emphasis on bulk sequence processing and data visualization. Its robustness has been validated by thousands of users. It will become a handy and useful toolkit for biologists.

Acknowledgements

Special thanks to thousands of TBtools users for their kind advices. Thanks for the help of all labmates in Xia Lab and He Lab.

Funding

This work is supported by funds from the National Natural Science Foundation of China (31872063 to R. X.), the Guangzhou Science and Technology Key Project (201804020063 to R. X.), the Innovation Team Project of the Department of Education of Guangdong Province (2016KCXTD011 to R. X.), and to Y. H. the Technology Commission of Guangdong Province (2013B020304002 to Y. H.) and the Key Research and Development Program of HuNan Province (2016JC2014 to H. C.).

References

Bailey,T.L. *et al.* (2009) MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, 202–208.

Deng,W. *et al.* (2014) HemI: A Toolkit for Illustrating Heatmaps. *PLoS One*, **9**, e111988.

Hall,T. *et al.* (2011) BioEdit: An important software for molecular biology. *GERF Bull. Biosci.*, **2**, 60–61.

Hu,B. *et al.* (2015) GSDS 2.0: An upgraded gene feature visualization server. *Bioinformatics*, **31**, 1296–1297.

Lex,A. and Gehlenborg,N. (2014) Sets and intersections. *Nat. Methods*, **11**, 779.

Oliveros, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams.

http://bioinfogp.cnb.csic.es/tools/venny/index.html

Rambaldi,D. and Ciccarelli,F.D. (2009) FancyGene: Dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics*, **25**, 2281–2282.