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

Chengjie Chen[1,2,3*], Rui Xia[1,2,3], Hao Chen[4], Yehua He[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

*The corresponding Author.

YH: heyehua@hotmail.com

RX: rxia@scau.edu.cn

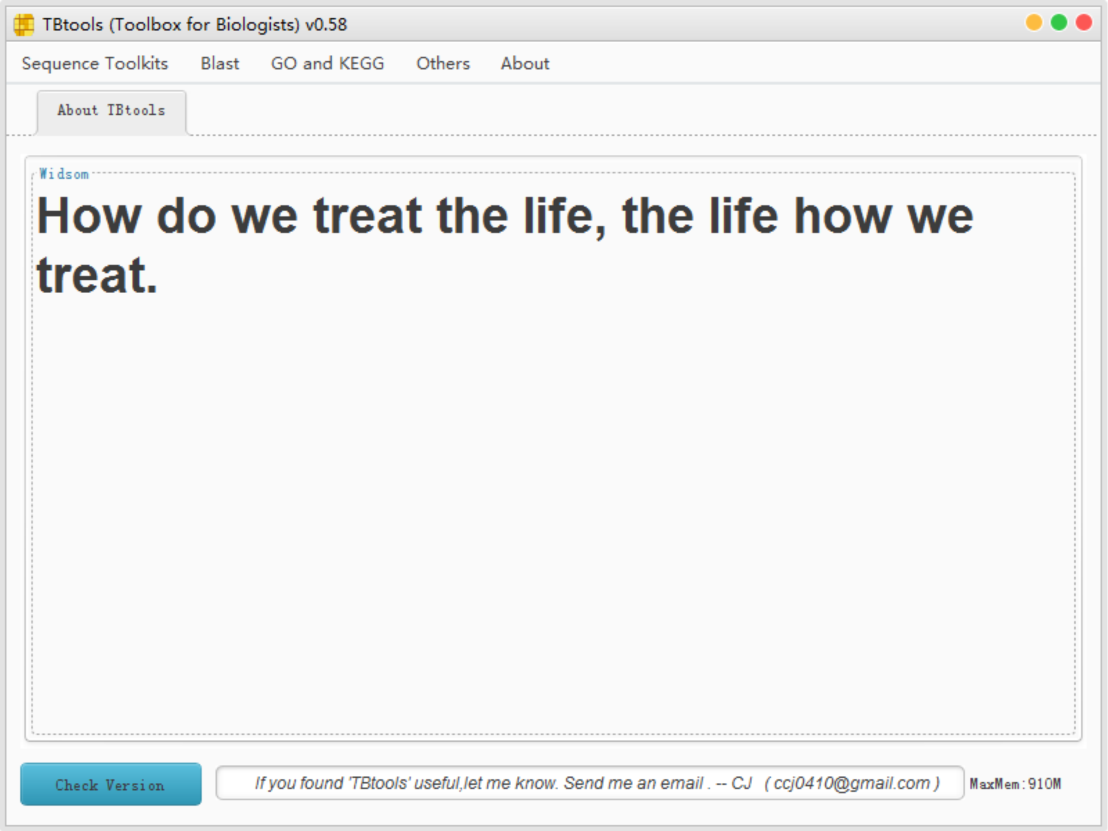
CC: ccj0410@gmail.com

Abstract

Various softwares or pipelines have been developed for biological information mining from high-throughput sequencing (HTS) data, and most of them relies on programming and command-line environment with which most biologists are unfamiliar. Bioinformatic tools with an user-friendly interface are preferred by wet-lab biologists. Here, we describe TBtools, a Toolkit for Biologists integrating various HTS-data handling tools with a user-friendly interface. It includes a large collection of functions, which facilitate many simple, routine but elaborate tasks working on HTS data, such as bulk sequence extraction, gene set functional enrichment, venn diagram and etc. TBtools can run under all operating systems with JRE1.6 and is freely available at github.com/CJ-Chen/TBtools. Since its development, it has been used by many researchers. It will be a useful toolkit for wet-lab biologists to work on all kinds of high-throughput data.

1. Introduction

The rapid development of high-throughtput sequencing techniques leads to the boom of sequencing data for biologists, all kinds of softwares, packages, and pipelines have been developed to meet various analysis needs. Most of them relies on programming and command-line environment with which most biologists are unfamiliar. Few user-friendly tools focus on common but elaborate tasks such as bulk sequences extraction, gene set functional enrichment and blast alignments visualization[1]. Many web-based applications, R packages or pipelines require user to either upload big data or work under command-line environment. Running small scripts or a few commands, which might be easy for computation specialists or bioinformaticians to accomplish in seconds, may take non-specialist users much time for the first try and even more time in future. Thus, we here present TBtools, a toolset that sets out to save time for wet-lab biologists from daily sequence analysis work. To date, it includes more than 40 little functions classified into five groups.



How do we treat the life, the life how we treat.

If you found 'TBtools' useful,let me know. Send me an email . -- CJ (ccj0410@gmail.com)

Figure 1. Overview of the main menu of TBtools. More than 40 functions are classified into five groups, which are showed on the main menu.

2. Manual of TBtools

TBtools is an easy-to-use toolset containing two running mode: GUI mode for out-of-the-box usage and command-line mode for batch analyses. A bundle of functions have been developed in the current version of TBtools (Figure 1, GUI mode) for frequent needs in daily molecular analyses, and more tools can be added with the increase of demands. The “Check Version” could be used to check whether an update of TBtools is available.

2.1 Sequence Toolkits

Fasta Tools are used for the manipulation of sequence files in fasta format.

Amazing Fasta Extrator can accomplish almost all fasta sequences extraction tasks with the utilization of pre-built index.

Quick Fasta Extractor or Filter extracts/filters sequences in a quick mode, without pre-built index.

Fasta Extractor extract sequences from a fasta file in a slow fasta record parsing mode. (Decrepited)

Fasta Subseq is similar to Fasta Extrator, but is used to extract subsequences of fasta records. (Decrepited)

Fasta Stat generates summary information of each fasta record and the whole fasta set, such as number of record, length of sequences, GC content and N50 and so on.

Fasta simplifier simplifies ID of all fasta records.

Fasta Table Converter converts sequence file between fasta format and tab-delimited table.

Fasta Merge And Split merges fasta records from several files or splits fasta records into several files.

Fasta Sequence Manipulator can do Reverse sequence, Complement sequence, and DNA to RNA conversion.

NCBI Seq Downloader downloads sequences from NCBI via their API in Fasta/GenBank format.

Get Complete ORF (Open Reading Frame) predicts complete ORF from input sequence(s).

Check Primers (Simple e-PCR) can be used to check the specificity of primers via hamming distance.

Gtf/Gff3 Sequences Extractor extracts sequences from genome sequences according to the gtf/gff3 file, like complete set of cds sequences.

2.2 Blast Wrapper

Auto Blast Tools contains three auto-blast wrapper, **Blast Several Sequences 2** (to) **A Big Database**, **Blast Compare 2 Seq [Sets] and Blast Compare 2 Seq [Sets] <Big File>**, which can automatically determine the molecular type of input sequences.

Reciprocal Blast is used for conducting reciprocal blast between two fasta files. It can also be used for ID matching between two separate transcriptome assembly results.

Blast Result Visualization contained three little application to visualize blast results in different format "Alignment Graph", "Dot plot" and "Pileup Graph" (Figure 2).

Blast XML File Converter converts xml file into tab-delimited table, same as Blast+ "-outfmt 6" and the other collapsed all HSP into a single line.

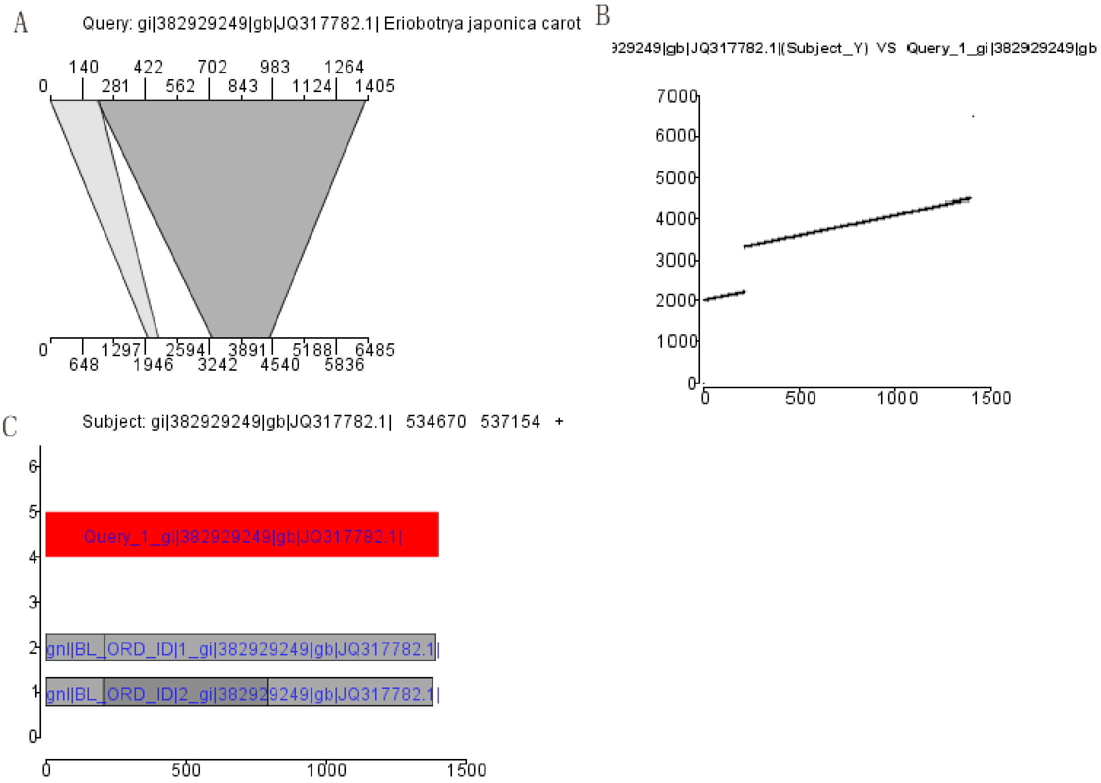


Figure 2 Three ways of visualization of Blast XML file. A) Alignment Graph; B) Dot plot; C) Pileup Graph

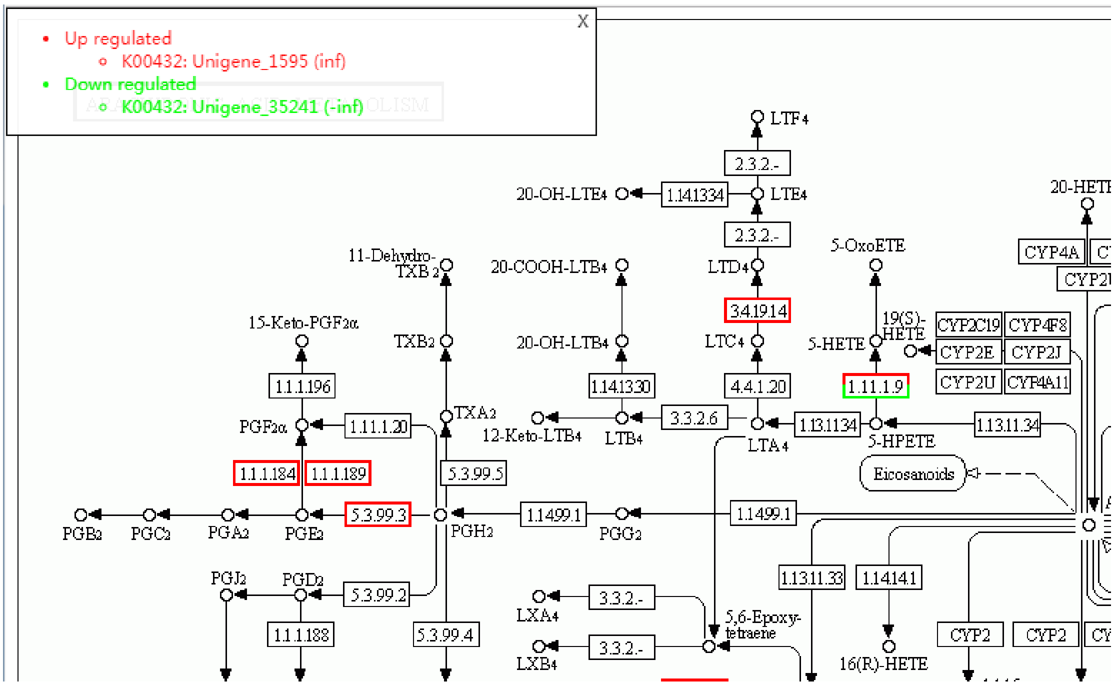
2.3 GO and KEGG

GO Annotation is an application for mapping NCBI gi/accession and Uniprot ID to Gene Ontology ID based on ID matchning information from ftp://ftp.pir.georgetown.edu/databases/idmapping/idmapping.tb.gz.

GO Enrichment conducts GO term enrichment analysis based on hypergeometric distribution.

KEGG Enrichment conducts KEGG pathway enrichment analysis based on hypergeometric distribution.

KEGG Pathway Map Drawer colors pathway maps and generates interactive pathway file (.html, Figure 3).



- Up regulated
  - K00432: Unigene_1595 (inf)
- Down regulated
  - K00432: Unigene_35241 (-inf)

Figure 3 Interactive Graph generated by TBtools.

2.4 Others

Table Manipulator provides two applications (Table ID Extractor and Filter) to manipulate **big** tab-delimited file, like extraction, filtering or sorting of rows and columns.

Wonderful Venn generates interactive venn plot allowing up to six ID sets (Figure 4A).

Map Gene on Genome produces a graph showing positions of genes on chromosomes based on input sequences file or position information (Figure 4B).

Dual Synteny Plotter visualizes results from MCScanX[2] in an interactive mode (Figure 4C).

Domain/Motif Pattern Drawer visualizes motif/domain information predicted by the MEME suite, Pfam and NCBI CD-search database. Besides, we also integrate a function for gene structure (exon/intron) presentation(Figure 5), which accepts a gtf/gff file and a list of gene IDs. It has no limit to the size of input file and is easier to use than GSDS[3].

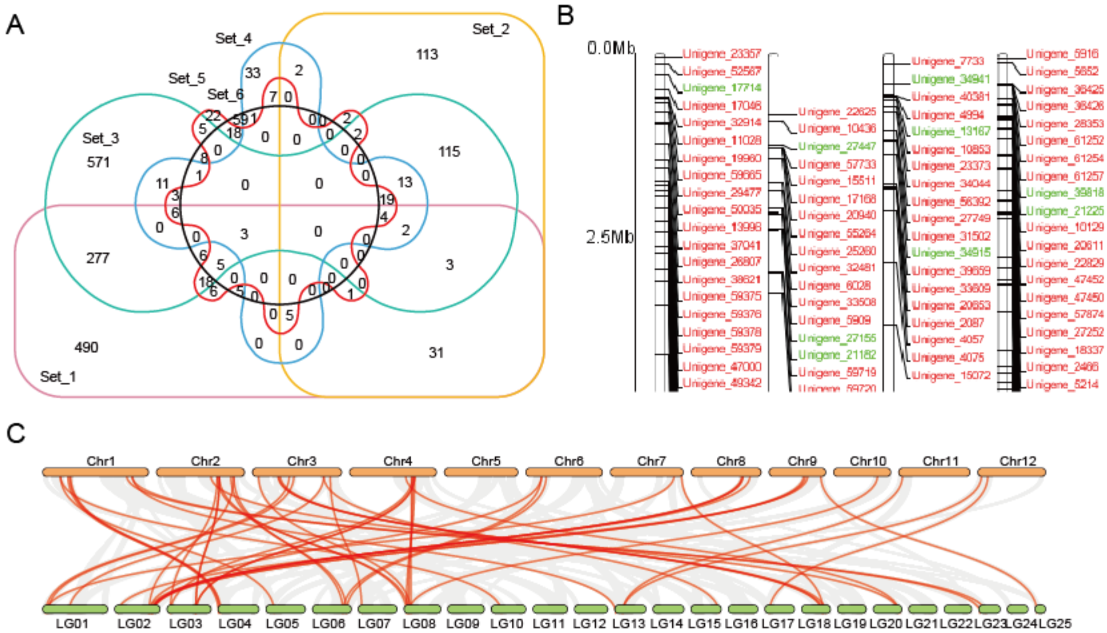


Figure 4. Example graph of wonderful venn, map gene on genome and dual synteny plotter.

3. Discussion and conclusion

Recent year, bioinformatics develops fast and has been emerging as a critical discipline for biology study. Major efforts have been put into the development of pipelines, which emphasize on the performance on 'big' bio-data analyses such as genome assembly, annotation, expression analysis, easy-to-use tools with user-friendly interface is much less available. We present TBtools, a toolset to meet the needs of wet-lab biologists for routine data work, who are unfamiliar to programing or command-line environments. TBtools integrates most frequently used functions into one handy executable jar file. Compare to existing softwares or pipelines, such as venny, kalama and comprehesive R packages, TBtools is lighter and easy-to-use, with no requirement of data upload, network connection and special computational skills. On the other hand, TBtools contains a homemade graphic module 'JJplot2' (now JIGplot) which is a java implementation of 'Grammer of Graphic', making it very suitable for the production of publishable graphs. We believe that TBtools will be a handy and useful toolset to facilitate downstream utilization of HTS data, especially for web-lab biologists.

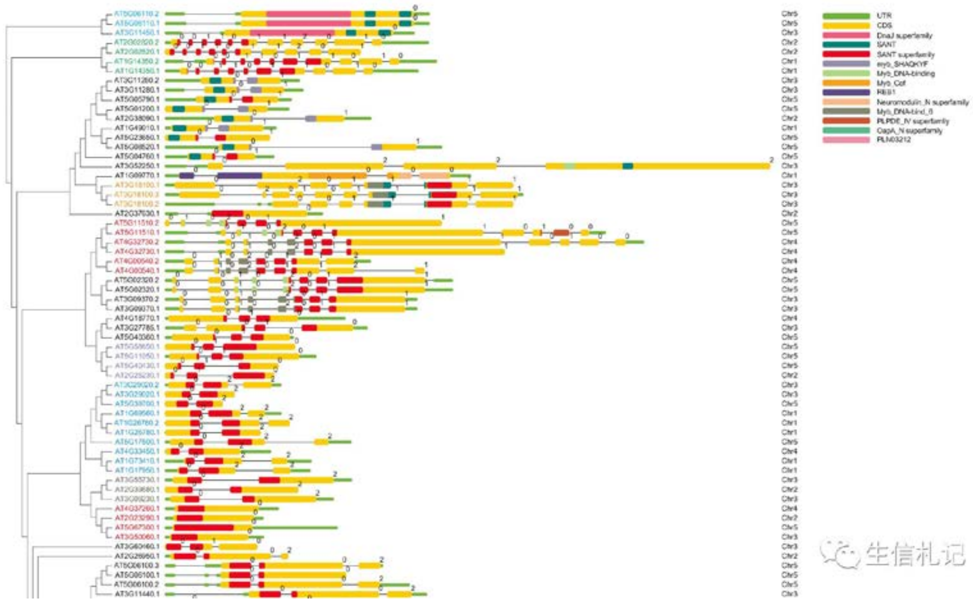


Figure 5. Examples graph of Motif/Doman Patter Drawer

4. Availability and requirements

TBtools is available as a jar file on https://github.com/CJ-Chen/TBtools. JRE verison not lower than 1.6 is needed, which is pre-installed in Linux or Mac. Window users may have installed it before using TBtools. Blast+ needs to be added into environment path if user want to use the **Blast wrapper** application.

Competing interests

No competing interest is delared.

Authors' contributions

CC and HC designed the software. CC did the coding jobs. YH and RX initiated and supervised the project.

Acknowledgments

We thank all valuable suggestions and feedbacks from hundreds of TBtools users from QQ Chatting Group: bioinformatics*中国(276151571) and TBtools 使用交流群 (553679029).

1. Neumann, R., Kumar, S., Haverkamp, T. H. & Shalchian-Tabrizi, K. BLASTGrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive BLAST data. *BMC Bioinformatics* **15,** 128 (2014).

2. Wang, Y. *et al.* MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40,** 1–14 (2012).

3. Guo, A. Y., Zhu, Q. H., Chen, X. & Luo, J. C. [GSDS: a gene structure display server]. *Yi Chuan* **29,** 1023–1026 (2007).