Bioinformatics, 38(20), 2022, 4806–4808
https://doi.org/10.1093/bioinformatics/btac561
Advance Access Publication Date: 24 August 2022
Applications Note

Genome analysis

GTFtools: a software package for analyzing various features of gene models

Hong-Dong Li [1,2,*,†], Cui-Xiang Lin[1,2,†] and Jiantao Zheng[1,2]

[1]School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, P.R. China and [2]Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha, Hunan 410083, P.R. China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Tobias Marschall

Received on January 9, 2022; revised on July 30, 2022; editorial decision on August 9, 2022; accepted on August 12, 2022

Abstract

Motivation: Gene-centric bioinformatics studies frequently involve the calculation or the extraction of various features of genes such as splice sites, promoters, independent introns and untranslated regions (UTRs) through manipulation of gene models. Gene models are often annotated in gene transfer format (GTF) files. The features are essential for subsequent analysis such as intron retention detection, DNA-binding site identification and computing splicing strength of splice sites. Some features such as independent introns and splice sites are not provided in existing resources including the commonly used BioMart database. A package that implements and integrates functions to analyze various features of genes will greatly ease routine analysis for related bioinformatics studies. However, to the best of our knowledge, such a package is not available yet.

Results: We introduce GTFtools, a stand-alone command-line software that provides a set of functions to calculate various gene features, including splice sites, independent introns, transcription start sites (TSS)-flanking regions, UTRs, isoform coordination and length, different types of gene lengths, etc. It takes the ENSEMBL or GENCODE GTF files as input and can be applied to both human and non-human gene models like the lab mouse. We compare the utilities of GTFtools with those of two related tools: Bedtools and BioMart. GTFtools is implemented in Python and not dependent on any third-party software, making it very easy to install and use.

Availability and implementation: GTFtools is freely available at www.genemine.org/gtftools.php as well as pyPI and Bioconda.

Contact: hongdong@csu.edu.cn

1 Introduction

Computing or extracting various features of gene models such as splice sites, independent introns, merged exons and different types of gene lengths, is frequently encountered in routine bioinformatics analysis. Such computation is essential for subsequent analysis. For example, independent introns represent the intron that does not overlap with any exon of any transcript isoforms of any gene, which is used for detecting intron retention events from RNA-seq data (Broseus *et al.*, 2020; Li *et al.*, 2020, 2021; Pimentel *et al.*, 2015; Wu *et al.*, 2022). As independent introns do not overlap with any exon, using independent introns as the reference ensures that detected intron retention events are unambiguously from intron regions (i.e. not from overlapping exons). Merged exons (also called union exons in literature) are obtained by merging exons of multiple transcript isoforms of the same gene (Zhang *et al.*, 2016). Merged exons are used as the reference to map RNA-seq reads and the length of merged exons is used to calculate reads per kilobase of exon per million mapped fragments (RPKM) (Zhang *et al.*, 2016). Other analysis includes obtaining flanking regions around transcription start sites (TSS) (Young *et al.*, 2011), and gene expression normalization by converting RNA-seq read counts to fragments per kilobase of exon per million mapped fragments (FPKM) where exonic gene length needs to be calculated (Wang *et al.*, 2009). A tool that can accomplish part of these tasks is BioMart (Smedley *et al.*, 2009). However, it is dependent on database querying and could be slow when downloading large amounts of data. In addition, some features such as independent introns and splice sites are not provided in BioMart.

Gene models provide basic annotations of genes, such as exons, introns and untranslated regions (UTRs), which are often annotated in gene transfer format (GTF) files. Other features of genes could be derived from gene models. Therefore, an efficient way for analyzing various modes of gene features would be directly interrogating gene models. The gene models provided in the ENSEMBL or GENCODE database (Harrow *et al.*, 2012) are the most comprehensive and

most widely used in practice. Here, we introduce GTFtools, a standalone command-line software that provides a set of functions to extract features from gene models. It is not dependent on any existing bioinformatics tools and thus is easy to install and use. GTFtools provides a new tool for facilitating routine bioinformatics analysis.

2 Methods and software implementation

GTFtools is a software package to calculate or extract multiple features related to genes based on gene models defined in GTF files. The input is a GTF file, which can be obtained from the ENSEMBL or GEOCODE database. The output data are written in text files. The input parameters specified by users are parsed by the '*argparse*' package, which hence needs to be installed for running GTFtools.

The main features of genes that can be calculated with GTFtools are illustrated in Figure 1. These features mainly include merged exons of splice isoforms of genes, splice sites (Yeo *et al.*, 2004), independent introns (Li *et al.*, 2020), four types of gene length (the mean, median and max of lengths of isoforms of a gene, and the length of merged exons of isoforms of a gene), UTRs and TSS-flanking regions. These features are frequently used in routine bioinformatics analysis. For example, the splice site region is used for studying splicing strength; TSS-flanking regions are used to search for transcription factor binding sites; gene length is required for calculating RNA-seq-based gene expression measured as FPKM; by searching UTRs sequence, miRNA-binding sites can be identified; independent introns are used for detecting intron retention events.

GTFtools is implemented as a command-line software in Python3 and works on both Linux and Mac operating systems. To obtain the usage of GTFtools, users can issue the command *python gtftools.py -h* from the command line. In doing so, the input options and their meanings will be shown on screen. We illustrate the use of GTFtools as follows. For example, by running *python gtftools.py -t tss_output.bed demo.gtf*, GTFtools will read the GTF file *demo.gtf* and output TSS-flanking regions into a bed-format file *tss_output.bed* specified by the *-t* option. As a second example, by using *python gtftools.py -l gene_length.txt demo.gtf*, three different types of gene lengths will be calculated. In addition to these two examples, we provide an example analysis for each parameter option of GTFtools, which is available in the table on the website https://www.genemine.org/gtftools.php. Note that although the demo.gtf is from human gene models, GTFtools can also be applied to non-human gene models like the lab mouse.

GTFtools uses the MIT license and is freely available on www.genemine.org/gtftools.php as well as on pyPI (using the command *pip install gtftools* to install) and Conda (using the command *conda install -c bioconda gtftools* to install).

3 Comparison with related tools

Next, we compare the utilities of GTFtools with those of two related tools: Bedtools and BioMart (Smedley *et al.*, 2009).

First, Bedtools was 'motivated by a need for fast, flexible tools with which to compare large sets of genomic features' as stated on the official website (https://bedtools.readthedocs.io/en/latest/content/overview.html). Thus, one main utility of Bedtools is to compare two or more genomic intervals, which are represented in bed or GTF format; examples include calculating the intersection of two sets of intervals (using the 'intersect' sub-command), merging two sets of intervals (using the 'merge' sub-command), subtracting one set of intervals from another set (using the 'subtract' sub-command) and identifying the complementary intervals of a given set of intervals (using the 'complement' sub-command). By far, Bedtools has been evolved into a very powerful suit of tools that can perform many additional functions, such as computing coverage over an entire genome, manipulating fasta files, file format conversion and calculating the Jaccard statistic between two sets of intervals.

Compared to Bedtools, the key difference is that GTFtools does not compare two sets of intervals but rather takes as input a single gene model in GTF format to extract various features of genes, such as gene lengths, splice isoforms, exons, introns, UTR, etc. Therefore, GTFtools well complements Bedtools.

Second, BioMart is a tool to extract various features of genes such as gene symbols, gene coordinates, transcript coordinates, exons, peptides, related GO terms, etc. GTFtools and BioMart share a few utilities, including extracting the coordinates of genes and splice isoforms. Most of the utilities of these two tools are different. For example, extracting independent introns and splice site regions are unique to GTFtools while obtaining peptides and homologs are unique to BioMart.

In addition, we compare the speed of GTFtools with BioMart based on overlapping functionalities. Retrieving gene and transcript isoform features such as chromosome and gene symbols are common to both tools and are therefore considered here for the comparison. The same GTF file (human gene annotation, Ensembl version 106) is used for testing. We perform the test on a MacBook Pro machine with 2.5 GHz Quad-Core Intel Core i7 and 16GB memory. For BioMart, we use the biomaRt R package (v2.52.0) and test the

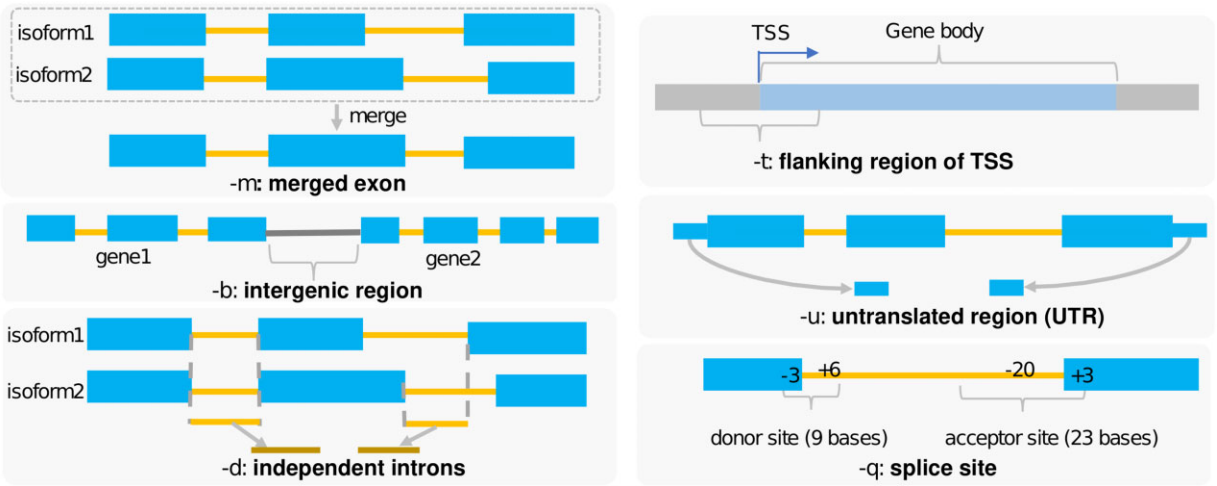

Fig. 1. Illustration of some gene features that can be obtained by GTFtools, including merged exons, TSS-flanking regions, intergenic regions, UTRs, independent introns and splice sites. *Note*: independent introns mean the intronic regions that do not overlap with any exons of any isoform of a gene, which can be used to detect unambiguous intron retention events; the splice site region is defined based on the MaxEntScan method (Yeo *et al.*, 2004), where the donor site is 9 bases long with 3 bases in exon and 6 bases in intron, and the 3′ acceptor site is 23 bases long with 20 bases in the intron and 3 bases in the exon

four different mirror databases ('www', 'useast', 'uswest' and 'asia'); the time cost corresponding to the fastest connection is recorded for comparison. For gene feature extraction (including chromosome, gene ID, gene symbol, start position, end position and strand), GTFtools and BioMart take about 4.8 and 12.0 s, respectively. For transcript feature extraction, GTFtools and BioMart take about 6.8 and 17.5 s.

4 Conclusion

A python package GTFtools for computing or extracting various features of gene models in GTF format has been developed in this article. It takes GTF files downloaded from the ENSEMBL or GENCODE database as input. Some features such as independent introns and splice site regions are unique to our tool. GTFtools is easy to install and use. It is expected that GTFtools would become widely used and facilitate routine bioinformatics analysis.

Funding

This work was supported by the National Key Research and Development Program of China [2022ZD0213700].

Conflict of Interest: none declared.

References

Broseus,L. *et al.* (2020) Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput. Struct. Biotechnol. J.*, **18**, 501–508.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.

Li,H.-D. *et al.* (2020) iREAD: a tool for intron retention detection from RNA-seq data. *BMC Genomics*, **21**, 128.

Li,H.-D. *et al.* (2021) Integrative functional genomic analysis of intron retention in human and mouse brain with Alzheimer's disease. *Alzheimer's Dement.*, **17**, 984–1004.

Pimentel,H. *et al.* (2015) Keep me around: intron retention detection and analysis. arXiv:1510.00696v1.

Smedley,D. *et al.* (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Wu,Z.-P. et al. (2022) DeepRetention: a deep learning approach for intron retention detection. *Big Data Mining Anal.* https://doi.org/10.26599/BDMA.2022.9020023.

Yeo,G. *et al.* (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.

Young,M.D. *et al.* (2011) ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.*, **39**, 7415–7427.

Zhang,C. *et al.* (2016) Bioinformatics tools for RNA-Seq gene and isoform quantification. *Next Generat. Sequenc. Appl.*, **3**, 3.