SOFTWARE TOOL ARTICLE

TFutils: Data structures for transcription factor bioinformatics

[version 1; peer review: 1 approved with reservations]

Benjamin J. Stubbs¹, Shweta Gopaulakrishnan¹, Kimberly Glass¹, Nathalie Pochet²,³, Celine Everaert²,³, Benjamin Raby¹,⁴, Vincent Carey¹

¹Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA
²Broad Institute, Cambridge, MA, 02142, USA
³Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA
⁴Pulmonary Genetics Center, Children's Hospital Boston, Boston, MA, 02115, USA

Abstract

DNA transcription is intrinsically complex. Bioinformatic work with transcription factors (TFs) is complicated by a multiplicity of data resources and annotations. The Bioconductor package TFutils includes data structures and functions to enhance the precision and utility of integrative analyses that have components involving TFs. TFutils provides catalogs of human TFs from three reference sources (CISBP, HOCOMOCO, and GO), a catalog of TF targets derived from MSigDb, and multiple approaches to enumerating TF binding sites. Aspects of integration of TF binding patterns and genome-wide association study results are explored in examples.

Keywords

Transcription factors, Gene expression, Gene regulation, Bioconductor

This article is included in the Bioconductor gateway.
Introduction
A central concern of genome biology is improving understanding of gene transcription. In simple terms, transcription factors (TFs) are proteins that bind to DNA, typically near gene promoter regions. The role of TFs in gene expression variation is of great interest. Progress in deciphering genetic and epigenetic processes that affect TF abundance and function will be essential in clarifying and interpreting gene expression variation patterns and their effects on phenotype. Difficulties of identifying functional binding of TFs, and opportunities for using information of TF binding in systems biology contexts, are reviewed in Lambert et al.¹ and Weirauch et al.².

This paper describes an R/Bioconductor package called TFutils, which assembles various resources intended to clarify and unify approaches to working with TF concepts in bioinformatic analysis. Computations described in this paper can be carried out with Bioconductor version 3.8. The package can be installed with

```r
# use install.packages("BiocManager") if not already available
library(BiocManager)
install("TFutils")
```

In the next section we describe the basic concepts of enumerating and classifying TFs, enumerating TF targets, and representing genome-wide quantification of TF binding affinity. This is followed by a review of the key data structures and functions provided in the package, and an example in cancer informatics.

The present paper does not deal directly with the manipulation or interpretation of sequence motifs. An excellent Bioconductor package that synthesizes many approaches to these tasks is universalmotif.

Basic concepts of transcription factor bioinformatics

Enumerating transcription factors

Given the importance of the topic, it is not surprising that a number of bioinformatic research groups have published catalogs of transcription factors along with metadata about their features. Standard nomenclature for TFs has yet to be established. Gene symbols, motif sequences, and position-weight matrix catalog entries have all been used as TF identifiers.

In TFutils we have gathered information from four widely used resources, focusing specifically on human TFs: Gene Ontology (GO, Ashburner et al.³, in which GO:0003700 is the tag for the molecular function concept “DNA binding transcription factor activity”), CISBP (Catalog of Inferred Sequence Binding Preferences) (Weirauch et al.²), HOCOMOCO (Homo sapiens Comprehensive Model Collection) (Kulakovskiy et al.⁴), and the “c3 TFF (transcription factor target)” signature set of MSigDb (Molecular Signatures Database) (Subramanian et al.⁵). Figure 1 depicts the sizes of these catalogs, measured using counts of unique HGNC gene symbols. The enumeration for GO uses Bioconductor’s org.Hs.eg.db (version 3.7.0) package to find direct associations from GO:0003700 to HGNC symbols. The enumeration for MSigDb is heuristic and involves parsing the gene set identifiers used in MSigDb for exact or close matches to HGNC symbols. For CISBP and HOCOMOCO, the associated web servers provide easily parsed tabular catalogs.

Classification of transcription factors

As noted by Weirauch et al.², interpretation of the “function and evolution of DNA sequences” is dependent on the analysis of sequence-specific DNA binding domains. These domains are dynamic and cell-type specific (Gertz et al.⁶). Classifying TFs according to features of the binding domain is an ongoing process of increasing intricacy. Figure 2 shows excerpts of hierarchies of terms related to TF type derived from GO (on the left) and TFclass (Wingender et al.⁷). There is a disagreement between our enumeration of TFs based on GO in Figure 1 and the 1919 shown in AmiGO, as the latter includes a broader collection of receptor activities.

Table 1 provides examples of frequently encountered TF classifications in the CISBP and HOCOMOCO catalogs. The numerical components of the HOCOMOCO classes correspond to TFClass subfamilies (Wingender et al.⁷).

Enumerating TF targets

The Broad Institute MSigDb (Subramanian et al.⁵) includes a gene set collection devoted to cataloging TF targets. We have used Bioconductor’s GSEABase package (version 1.45.0) to import and serialize the gmt representation of this collection.
Figure 1. Sizes of transcription factor (TF) catalogs and of intersections based on HGNC (HUGO Gene Nomenclature Committee) symbols for TFs.

Figure 2. Screenshots of AmiGO and TFClass hierarchy excerpts.
Names of TFs for which target sets are assembled are encoded in a systematic way, with underscores separating substrings describing motifs, genes, and versions. Some peculiarity in nomenclature in the MSigDb labels can be observed:

grep("NFK", names(TFutils::tftColl), value=TRUE)

# [1] "NFKAPPAB65_01" "NFKAPPAB_01" "NFKB_Q6"
# [4] "NFKB_C" "NFKB_Q6_01" "GGGNNTTTCC_NFKB_Q6_01"

Manual curation will be needed to improve the precision with which MSigDb TF target sets can be associated with specific TFs or motifs.

Quantitative predictions of TF binding affinities
In this subsection we address representation of putative binding sites. First we illustrate how to represent sequence-based affinity measures and the binding site locations implied by these. We then discuss use of results of ChIP-seq experiments for cell-type-specific binding site enumeration.

Affinity scores based on reference sequence. The FIMO algorithm of the MEME suite (Grant et al.) was used to score the human reference genome for TF binding affinity for 689 motif matrices to which genes are associated. Full details are provided in Sonawane et al. Sixteen (16) tabix-indexed BED files are lodged in an AWS S3 bucket for illustration purposes.
We harvest scores in a genomic interval of interest (bound to `fimo16` in the `rowRanges` assignment below) using `reduceByFile`. This yields a list with one element per file. Each such element holds a list of scanTabix results, one per query range.

```r
library(BiocParallel)
register(SerialParam()) # important for macosx?
rowRanges(fimo16) = GRanges("chr17", IRanges(38.077e6, 38.084e6))
rr = GenomicFiles::reduceByFile(fimo16, MAP=function(r,f)
  scanTabix(f, param=r))
```

scanTabix produces a list of vectors of text strings, which we parse with `data.table::fread`. The resulting tables are then reduced to a genomic location and -log10 of the p-value derived from the binding affinity statistic of FIMO in the vicinity of that location.

```r
asdf = function(x) data.table::fread(paste0(x, collapse="\n"), header=FALSE)
gg = lapply(rr, function(x) {
  tmp = asdf(x[[1]][[1]])
  data.frame(loc=tmp$V2, score=-log10(tmp$V7))
})
for (i in 1:length(gg)) gg[[i]]$tf = colData(fimo16)[i,2]
```

It turns out there are too many distinct TFs to display names individually, so we label the scores with the names of the associated TF families as defined in CISBP.

```r
matchcis = match(colData(fimo16)[,2], cisbpTFcat[,2])
fann = cisbpTFcat[matchcis,]$Family_Name
for (i in 1:length(gg)) gg[[i]]$tfam = famn[i]
nn = do.call(rbind, gg)
```

A simple display of predicted TF binding affinity near the gene `ORMDL3` is provided in Figure 3.

**TF binding predictions based on ChIP-seq data from ENCODE.** The ENCODE project provides BED-formatted reports on ChIP-seq experiments for many combinations of cell type and DNA-binding factors. TFutils includes a table `encode690` that gives information on 690 experiments involving pairs formed from 91 cell lines and 161 TFs for which results have been recorded as GRanges instances that can be acquired with the `Annotation-Hub` (version 2.15.4) package. Positional relationships between cell-type specific binding sites and genomic features can be investigated. An illustration is given in Figure 4, in which is it suggested that in HepG2 cells, CEBPB exhibits a distinctive pattern of binding in the vicinity of `ORMDL3`.

**Summary**

We have compared enumerations of human transcription factors by different projects, provided access to two forms of binding domain classification, and illustrated the use of cloud-resident genome-wide binding predictions. In the next section we review selected details of data structures and methods of the `TFutils` package.
Figure 3. TF binding in the vicinity of gene ORMDL3. Points are -log10-transformed FIMO-based p-values colored according to TF class as annotated in CISBP. Segments at bottom of plot are transcribed regions of ORMDL3 according to UCSC gene models in build hg19.

Figure 4. Binding of CEBPB in the vicinity of ORMDL3 derived from ChIP-seq experiments in four cell lines reported by ENCODE. Colored rectangles at top are regions identified as narrow binding peaks, arrows in bottom half are exons in ORMDL3. Arrows sharing a common vertical position are members of the same transcript as cataloged in Ensembl version 75.
Methods
Implementation

The TFutils package is designed to lower barriers to usage of key findings of TF biology in human genome research. TFutils is supplied as a conventional R package distributed with, and making use of, the Bioconductor software ecosystem. TFutils includes ready-to-use reference data, tools for visualizing binding sites, and tools that simplify integrative use of TF binding information with GWAS findings.

Data resources

Catalogs. Two reference resources have been collected into the TFutils package as data.frame instances. These are cisbpTFcat (CISBP: 7592 x 28), and hocomoco.mono.sep2018 (mononucleotide models, full catalog, 769 x 9). These data.frames are snapshots of the CISBP and HOCOMOCO catalogs

Indexed BED in AWS S3. As described above fimo16 provides programmatic access to FIMO scores for 16 TFs, using the GenomicFiles (version 1.19.0) protocol.

Annotated reference to ENCODE ChIP-seq results. encode690 simplifies programmatic access to TF:cell-line combinations available in Bioconductor AnnotationHub (version 2.15.4).

TF targets enumerated in MsigDb. The c3-TFT (TF targets) subset from MSigDb is provided as a GeneSet-Collection instance as defined in GSEABase.

Illustrative GWAS records. The full EBI/EMBL GWAS catalog is available in the gwascat package (version 2.15.0); for convenience, an excerpt focusing on chromosome 17 is supplied with TFutils as gwascat_hg19_chr17.

Infrastructure for interacting with components of TFutils

Interactive enumeration of TF targets implicated in GWAS. The TTargs function runs a shiny app that permits selection of a TF in the nomenclature of the MsigDb c3/TFT gene set collection. The app will search an object provided by the gwascat package for references in the MAPPED_GENE field that match the targets of the selected TF. Figure 5 gives an illustration.

The TFCatalog S4 class. Reference catalogs for TF biology are structured with the TFCatalog S4 class. Two essential components for managing a catalog are the native TF identifier for the catalog and the HGNC gene symbol typically used to name the TF. The TFCatalog class includes a name field to name the catalog, and a character vector with elements comprised of the native identifiers for catalogued TFs.

Figure 5. TTargs() screenshot. This example reports on recent EBI GWAS catalog hits on chromosome 17 only.
For example, CISBP uses T004843_1.02 to refer to motifs associated with gene TFAP2B. There are five such motifs, three derived from SELEX, one from Transfac, and one from Hocomoco.

A `data.frame` instance that has an obligatory column named `HGNC` can include any collection of fields that offer metadata about the TF in the specified catalog. Here is how we construct and view a TFCatalog object using the CISBP reference data.

```r
data(cisbpTFcat)
TFs_CISBP = TFCatalog(name="CISBP.info",
    nativeIds=cisbpTFcat[,1],
    HGNCmap = cisbpTFcat)
```

```
## 7592 native Ids, including T004843_1.02 ...
## 1551 unique HGNC tags, including TFAP2B ...
```

**Operation: Installation**
The TFutils package can be installed in any version of R subsequent to 3.5.0, and therefore will be usable on Unix, Windows, or Mac platforms. The preferred method of installation employs the CRAN package BiocManager, through the R command `BiocManager::install("TFutils")`. All necessary dependencies will be installed through this process.

**Operation: Use cases**
In this section we consider applications of the tools in genetic epidemiology. First we look for TFs that may harbor variants associated with traits in the EBI GWAS catalog. Then we show how to enumerate traits associated with targets of a selected TF.

**TFs that are direct GWAS hits for a given trait.** `directHitsInCISBP` accepts a string naming a trait, and returns a `data.frame` of TFs identified as “mapped genes” for the trait, with their TF “family name”.

```r
library(dplyr)
library(magrittr)
library(gwascat)
data(ebicat37)
directHitsInCISBP("Rheumatoid arthritis", ebicat37)
```

```
## Joining, by = "HGNC"
##  HGNC Family_Name
## 1  ARID5B ARID/BRIGHT
## 7  EOMES T-box
## 15 GATA3 GATA
## 35 JAZF1 C2H2 ZF
## 37 MECP2 MBD
## 45 MTF1 C2H2 ZF
## 57 REL Rel
## 65 STAT4 STAT
## 79 AIRE SAND
## 82 IRF5 IRF
```

**Traits mapped to genes that are targets of a given TF**
`topTraitsOfTargets` will acquire the targets of a selected TF, check for hits in these genes in a given GWAS catalog instance, and tabulate the most commonly reported traits.

```r
tt = topTraitsOfTargets("MTF1", TFutils:::tftColl, ebicat37)
```
## remapping identifiers of input GeneSetCollection to Symbol...

## done

```r
head(tt)
```

<table>
<thead>
<tr>
<th>DISEASE.TRAIT</th>
<th>MAPPED_GENE</th>
<th>SNPS</th>
<th>CHR_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atopic dermatitis</td>
<td>TNXB</td>
<td>rs41268896</td>
<td>6</td>
</tr>
<tr>
<td>Atopic dermatitis</td>
<td>TNXB</td>
<td>rs12153855</td>
<td>6</td>
</tr>
<tr>
<td>Atopic dermatitis</td>
<td>KIF3A</td>
<td>rs2897442</td>
<td>5</td>
</tr>
<tr>
<td>Attention deficit hyperactivity disorder</td>
<td>SEMA3A</td>
<td>rs797820</td>
<td>7</td>
</tr>
<tr>
<td>Attention deficit hyperactivity disorder</td>
<td>DNM1</td>
<td>rs2502731</td>
<td>9</td>
</tr>
<tr>
<td>Attention deficit hyperactivity disorder</td>
<td>GPC6</td>
<td>rs7995215</td>
<td>13</td>
</tr>
</tbody>
</table>

```r
table(tt[,1])
```

- Atopic dermatitis: 3
- Attention deficit hyperactivity disorder: 3
- Height: 7
- Menarche (age at onset): 4
- Obesity-related traits: 11
- Rheumatoid arthritis: 3

**Discussion**

Sources and consequences of variations in DNA transcription are fundamental problems for cell biology, and the projects we have made use of for cataloging transcription factors are at the boundaries of current knowledge.

It is noteworthy that the four resources used for Figure 1 agree on names of only 119 TFs. The fact that CISBP distinguishes 475 TFs that are not identified in any other source should be better understood. We observe that the ascription of TF status to AHRR is based on its sharing motifs with AHR (see [http://cisbp.ccbr.utoronto.ca/TFreport.php?searchTF=T014165_1.02](http://cisbp.ccbr.utoronto.ca/TFreport.php?searchTF=T014165_1.02)).

Figure 2 and Table 1 show that the classification of TFs is now fairly elaborate. Use of the precise terminology of the TFClass system to label TFs of interest at present relies on associations provided with the HOCOMOCO catalog.

As population studies in genomic and genetic epidemiology grow in size and scope, principles for organizing and prioritizing loci associated with phenotypes of interest are urgently needed. Figure 5 shows that loci associated with phenotypes related to kidney function, lung function, and IL-8 levels are potentially unified through the fact that the GWAS hits are connected with genes identified as targets of VDR (vitamin D receptor). This example limited attention to hits on chromosome 17; the Tftargs tool permits ad libitum exploration of phenotype-locus-gene-TF associations. Our hope is that the tools and resources collected in TFutils will foster systematic development of evidence-based mechanistic network models for transcription regulation in human disease contexts, thereby contributing to the development of personalized genomic medicine.
Data availability
With the exception of the FIMO scoring data (fimo16), all data underlying the results are available as part of the article and no additional source data are required.

fimo16 links to indexed bed files in a public S3 bucket funded by the Bioconductor foundation. The underlying data is sourced from Sonawane et al. 2017 https://doi.org/10.1016/j.celrep.2017.10.001

Software availability
Source code is available from GitHub: https://github.com/vjcitn/TFutils
Archived source code: https://doi.org/10.18129/B9.bioc.TFutils
Licence: Artistic License 2.0

Grant information
Support for the development of this software was provided by the National Institutes of Health [U01 CA214846 to VC, U24 CA180996] and the Chan Zuckerberg Initiative [DAF 2018-183436 to VC, R01 NHLBI HL118455 to BR].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Open Peer Review

Current Referee Status: ?

Version 1

Referee Report 20 February 2019

https://doi.org/10.5256/f1000research.19660.r44033

Lihua Julie Zhu 1, Haibo Liu 2, Rui Li 3
1 Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School (UMMS), Worcester, MA, USA
2 Iowa State University, Ames, IA, USA
3 University of Massachusetts Medical School, Worcester, MA, USA

The TFutils package provides useful, convenient, integrated data structures for TF-related bioinformatics analyses, by incorporating the basic information of human transcription factors (TFs), such as TF classification, known TF targets, genome-wide TF binding sites and binding affinity scores, which might be used to prioritize candidate genetic variants and help understand gene transcriptional regulatory mechanisms. Importantly, it also provides an interactive interface to query TFs and TF targets implicated in human traits as discovered by many GWASs.

In a quick test, all demo code in this paper worked. However, to make sure TFutils is more useful to the bioinformatics community, a few questions may need to be addressed. Here are our detailed comments and questions.

1. TFutils includes resources from CISBP, HOCOMOCO, GO and MSigDb. There are additional human TF resources. Is there any reason not to include those resources such as JASPAR 1, Transfac 2, HDPI 3 and uniPro 4?

2. There are potential packages that will likely import TFutils such as TFBSTools for the analysis of transcription factor binding sites manipulation, motifStack for graphic representation of multiple motifs 5 and MotIV. It will be helpful to present a few lines of code to show how to integrate data from TFutils to aforementioned pipelines.

3. The section “Basic concepts of transcription factor bioinformatics” includes lots of background information, such as existing TF-related data sources/bases, TF classification, and how TFutils incorporates and access those resources. To make it easy to follow, we suggest break this part into the Introduction section and the Method section. The author may move the background information and TF classification to the Introduction section, and include an Implementation section in the Methods section to describe how TFutils incorporates all these data sources and how to retrieve the relevant information in TFutils and how to integrate with other packages as mentioned in 2, where the R script snippets can be displayed.

4. To maintain/increase the user base, it is important to keep the data up to date. Currently, the data were snapshots of the CISBP and HOCOMOCO catalogs. If the resources are not updated regularly, it’s unlikely that users will use TFutils after 2-3 years. Is there a plan in place to have the resources assembled by TFutils be update regularly? How often is the update going to be? Is it going to be automatic or manually?
5. Flexibility of the data structure is also important, as users may want to expand the utility of TFutils. Suggest authors describe how to add features to the current data structures in TFutils in the manuscript.

6. It will be useful to add information on the numbers of TFs and targets included in the assembled resources, as well as in the original databases.

7. There is a python package having the same name “tfutils” which is very popular. If it is not too hard to do, we suggest authors change the package name to avoid confusion.

8. Installation and running environments of the TFutils was described twice, once in the Introduction section, the other time in the Methods section: Operation: Installation. It is better to only describe this once in the Methods section.

9. There are many short paragraphs consisting of one or two sentences and related information are scattered into different sections. For instances, the last paragraph of the Introduction section about the limitations of TFutil might be moved to the Discussion part; whereas the third paragraph in the Discussion section might be moved to somewhere at the beginning of the Introduction section or where it is appropriate.

10. Page 6, the Summary section might be better moved to between the data availability section and the discussion section to summarize the implemented functionality of TFutil.

Besides those major issues, we also have a few minor questions:

1. Currently the abstract only mentions TF targets derived from the MSigDb. Considering that the ENCODE TF ChIP-seq data is one of the most significant resources for TF targets information as mentioned in the main text, suggest authors add how the ENCODE ChIP-seq data were incorporated into TFutils in the abstract.

2. Page 5, please clarify the type of details in the sentence “Full details are provided in Sonawane et al”.

3. Gene structure can be better depicted in Fig. 3 and Fig. 4, perhaps adopting the gene structure visualization in most genome viewers, showing exon/intron structure and gene transcription direction.

4. Please include the used R packages in the citation.

5. “TFtargs()” in Figure 5 legend needs to be edited.

6. For the subtitles under the Use cases section, suggest add find before “TFs that are direct GWAS …” and retrieve before “Traits mapped to genes that …”.

References


Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioinformatics, ChIP-seq, CRISPR technology, RNA-seq, annotation, ATAC-seq, motif analysis, shRNA/CRISPR screening, visualization, machine learning and database application

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com