SOFTWARE TOOL ARTICLE

REVISED TFutils: Data structures for transcription factor bioinformatics [version 2; peer review: 2 approved, 1 not approved]

Benjamin J. Stubbs[1], Shweta Gopaulakrishnan[1], Kimberly Glass[1], Nathalie Pochet[2,3], Celine Everaert[2,3], Benjamin Raby[1,4], Vincent Carey[1]

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

v2 **First published:** 05 Feb 2019, **8**:152 (https://doi.org/10.12688/f1000research.17976.1)
Latest published: 17 May 2019, **8**:152 (https://doi.org/10.12688/f1000research.17976.2)

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, including an interface to results of 690 ENCODE experiments. 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.

Open Peer Review

Reviewer Status ✓ ✓ ✗

	Invited Reviewers		
	1	**2**	**3**
REVISED **version 2** published 17 May 2019	✓ report	✓ report	✗ report
version 1 published 05 Feb 2019	? report	? report	

1 **Lihua Julie Zhu**, University of Massachusetts Medical School (UMMS), Worcester, USA
Haibo Liu, Iowa State University, Ames, USA
Rui Li, University of Massachusetts Medical School, Worcester, USA

2 **Giovanna Ambrosini**, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
Philipp Bucher, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

3 **Kevin Ernst**, Cincinnati Children's Hospital Medical Center, Cincinnati, USA

Matthew T. Weirauch, Cincinnati Children's Hospital Medical Center (CCHMC) , Cincinnati, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Vincent Carey (stvjc@channing.harvard.edu)

Author roles: Stubbs BJ: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Gopaulakrishnan S**: Conceptualization, Data Curation, Investigation, Methodology, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Glass K**: Conceptualization, Data Curation, Investigation, Methodology, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Pochet N**: Conceptualization, Funding Acquisition, Investigation, Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Everaert C**: Conceptualization, Investigation, Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Raby B**: Conceptualization, Funding Acquisition, Investigation, Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Carey V**: Conceptualization, Funding Acquisition, Investigation, Methodology, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Support for the development of this software was provided by the National Institutes of Health [U01 CA214846 to VC, U24 CA180996], the Chan Zuckerberg Initiative [DAF 2018-183436 to VC, R01 NHLBI HL118455 to BR] and NIH/NCI/ITCR R21 CA209940, NIH/NIAID R03 AI131066, U01 CA214846 collaborative set aside to NP.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Stubbs BJ, Gopaulakrishnan S, Glass K *et al.* **TFutils: Data structures for transcription factor bioinformatics [version 2; peer review: 2 approved, 1 not approved]** F1000Research 2019, **8**:152 (https://doi.org/10.12688/f1000research.17976.2)

First published: 05 Feb 2019, **8**:152 (https://doi.org/10.12688/f1000research.17976.1)

REVISED **Amendments from Version 1**

The two reviewer reports raised valid concerns related to the clarity of presentation.

A reviewer has noted that the interpretation of Figure 3 is difficult, and we concur. Additional work on the relationship between sequence-based and in vitro evidence of TF binding, specifically with respect to the combinatorial aspects of binding suggested by Figure 3, is warranted.

To demonstrate interplay of TFutils with existing Bioconductor tools, Figure 5 is new, and makes use of motifStack and MotifDb.

See referee reports

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

```
# 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*.

A complete reference manual enumerating all functions and data sets in the package is available at: http://bioconductor.org/packages/release/bioc/manuals/TFutils/man/TFutils.pdf

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 TFT (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

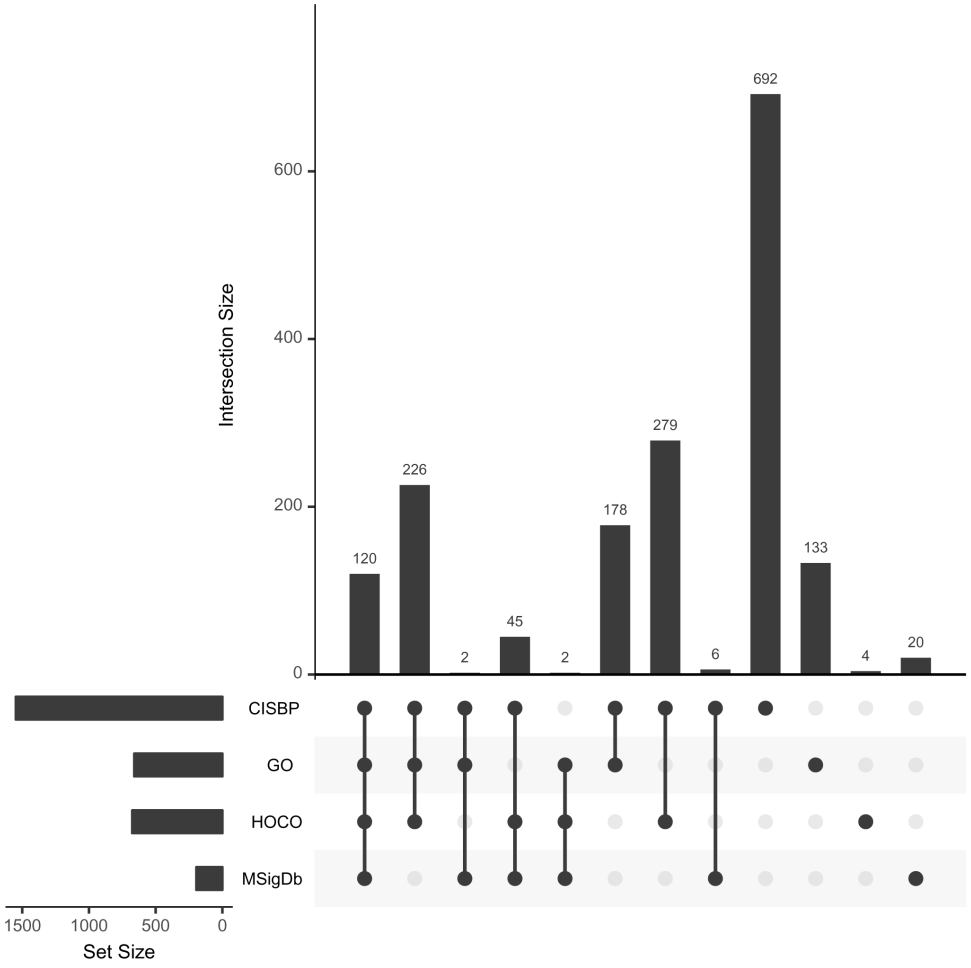

Figure 1. Sizes of transcription factor (TF) catalogs and of intersections based on HGNC (HUGO Gene Nomenclature Committee) symbols for TFs.

(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.

```
TFutils::tftColl
```

```
## GeneSetCollection
##   names: AAANWWTGC_UNKNOWN, AAAYRNCTG_UNKNOWN, ..., GCCATNTTG_YY1_Q6 (615 total)
##   unique identifiers: 4208, 481, ..., 56903 (12774 total)
##   types in collection:
##     geneIdType: EntrezIdentifier (1 total)
##     collectionType: NullCollection (1 total)
```

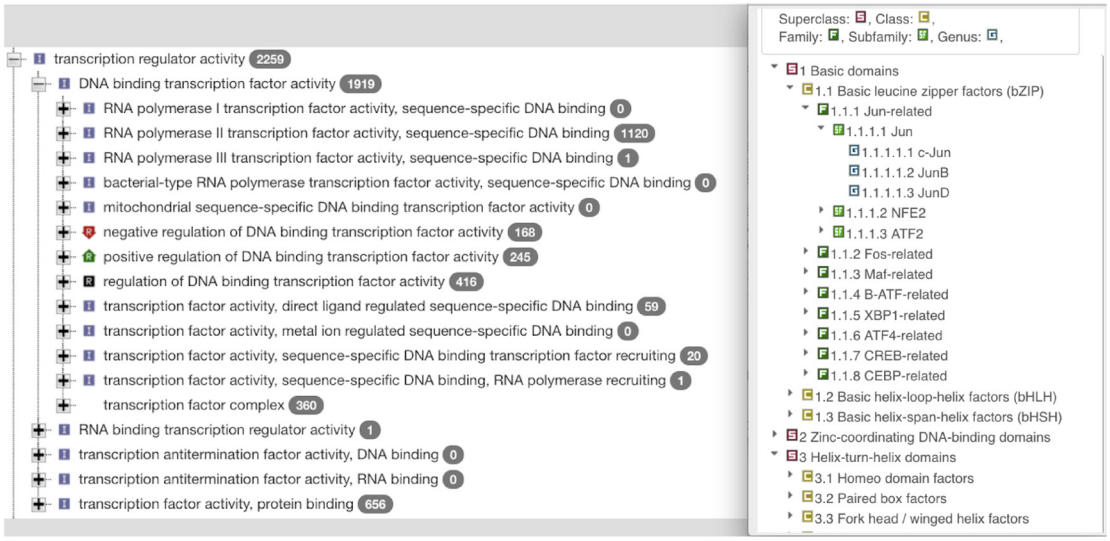

Figure 2. Screenshots of AmiGO and TFClass hierarchy excerpts.

Table 1. Most frequently represented transcription factor (TF) classes in CISBP and HOCOMOCO. The number of unique human TF_Name entries in CISBP is 1734. The number of unique Transcription factor entries in HOCOMOCO (Sept. 2018 version) is 678. Entries in columns Nc (Nh) are numbers of distinct TFs annotated to classes in columns CISBP (HO-COMOCO) respectively. Entries are ordered top to bottom by frequency of occurrence. There is no substantive correspondence between entries on a given row. Harmonization of class terminology is beyond the scope of this paper.

CISBP	Nc	HOCOMOCO	Nh
C2H2 ZF	655	More than 3 adjacent zinc finger factors{2.3.3}	106
Homeodomain	199	HOX-related factors{3.1.1}	41
bHLH	104	NK-related factors{3.1.2}	36
bZIP	66	Paired-related HD factors{3.1.3}	35
Unknown	49	Factors with multiple dispersed zinc fingers{2.3.4}	30
Forkhead	48	Forkhead box (FOX) factors{3.3.1}	27
Sox	48	Ets-related factors{3.5.2}	25
Nuclear receptor	46	Three-zinc finger Krueppel-related factors{2.3.1}	20
Myb/SANT	30	POU domain factors{3.1.10}	18
Ets	27	Tal-related factors{1.2.3}	18

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.*[8]) was used to score the human reference genome for TF binding affinity for 689 motif matrices to which genes are associated. Full details of the execution of FIMO are provided in Sonawane *et al*[9]. Sixteen (16) tabix-indexed BED files are lodged in an AWS S3 bucket for illustration purposes.

```
library(GenomicFiles)
data(fimo16)
fimo16
```

```
## GenomicFiles object with 0 ranges and 16 files:
## files: M0635_1.02sort.bed.gz, M3433_1.02sort.bed.gz, ..., M6159_1.02sort.
bed.gz, M6497_1.02sort.bed.
## detail: use files(), rowRanges(), colData(), ...
```

```
head(colData(fimo16))
```

```
## DataFrame with 6 rows and 2 columns
##          Mtag        HGNC
##   <character> <character>
## 1     M0635_1      DMRTC2
## 2     M3433_1       HOXA3
## 3     M3467_1        IRF1
## 4     M3675_1      POU2F1
## 5     M3698_1        TP53
## 6     M3966_1       STAT1
```

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.

```
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.

```
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.

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

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

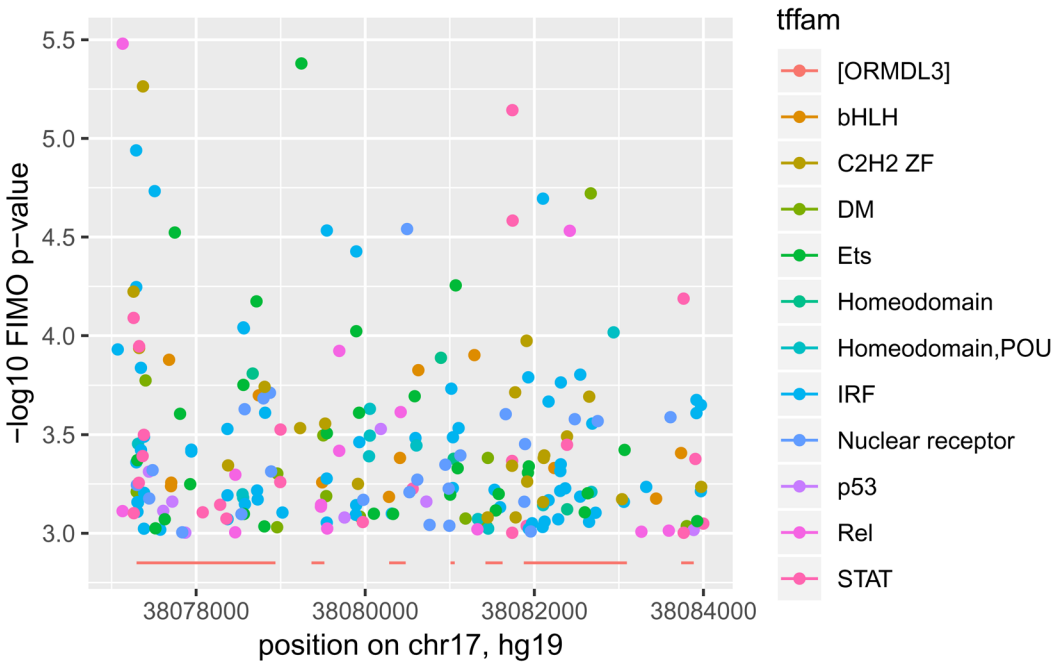

Figure 3. TF binding in the vicinity of gene *ORMDL3*. Points are -log10-transformed FIMO-based p-values colored according to TF class as annotated in CISBP. Segments at bottom of plot are transcribed regions of *ORMDL3* according to UCSC gene models in build hg19.

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 *AnnotationHub* (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*.

Visualization of motif relationships in a family of transcription factors

Inspired by a referee's suggestion, we created functions that couple the HOCOMOCO TFclass enumeration with Bioconductor's MotifDb[10] and motifStack[11] package resources. Figure 5 is the output of example(tffamCirc.plot), available in version 1.5.1 of TFutils.

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.

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. A complete enumeration of functions and data available in the package is provided in the reference manual at http://bioconductor.org/packages/release/bioc/manuals/TFutils/man/TFutils.pdf

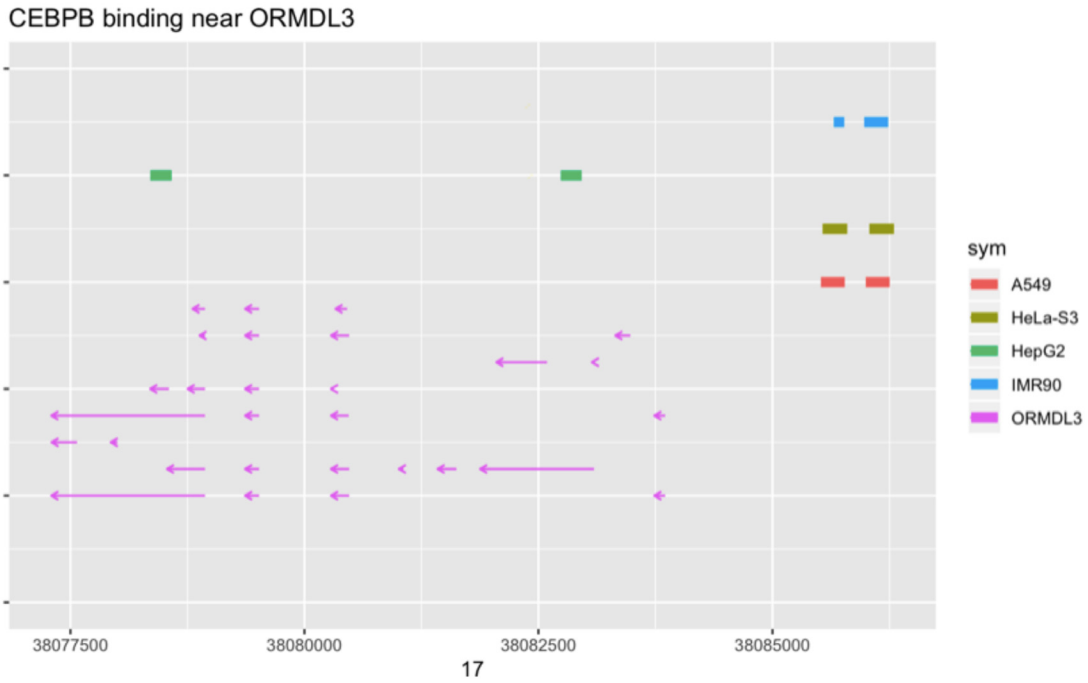

Figure 4. Binding of CEBPB in the vicinity of *ORMDL3* derived from ChIP-seq experiments in four cell lines reported by ENCODE. Colored rectangles at top are regions identified as narrow binding peaks, arrows in bottom half are exons in *ORMDL3*. Arrows sharing a common vertical position are members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

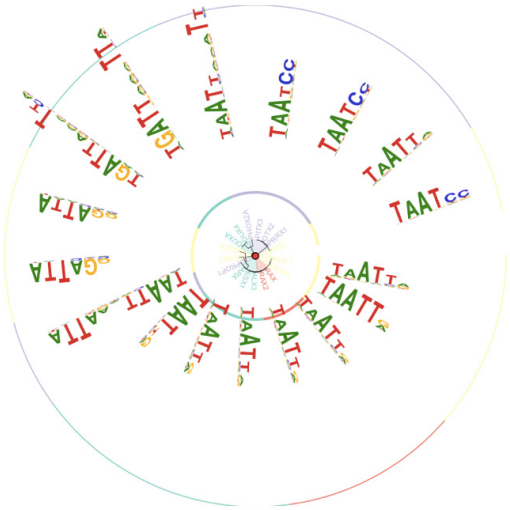

Figure 5. A circos display of motifs of transcription factors in the TFclass 3.1.3 (paired-related homeodomain factors).

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 GeneSetCollection 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 `TFtargs` 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 6 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.

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.

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

## TFutils TFCatalog instance CISBP.info
##  7592 native Ids, including
##    T004843_1.02 ... T153733_1.02
##  1551 unique HGNC tags, including
##    TFAP2B TFAP2B ... ZNF10 ZNF350
```

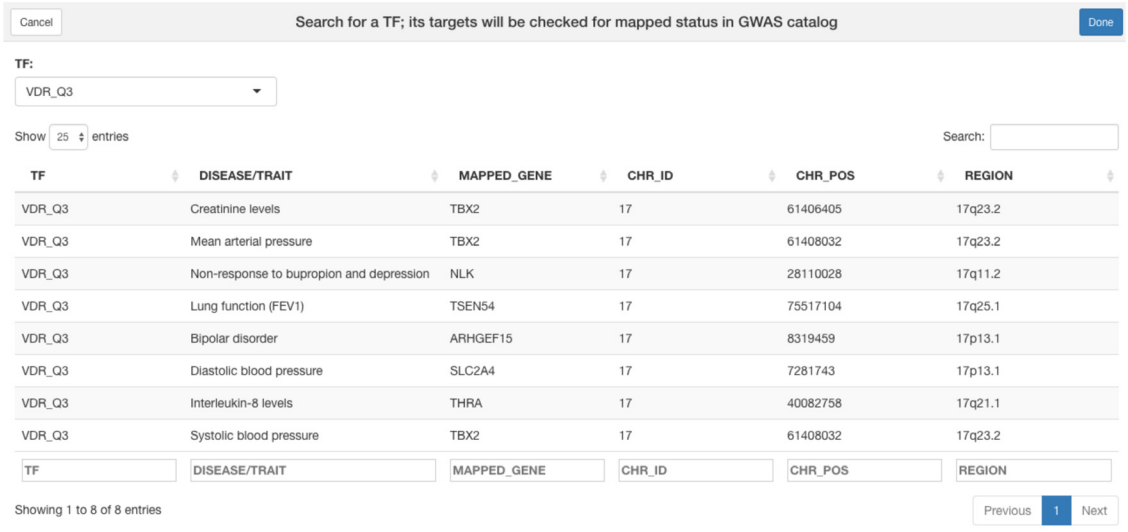

Figure 6. TFtargs screenshot. This example reports on recent EBI GWAS catalog hits on chromosome 17 only.

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.

Find 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".

```
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
```

Retrieve 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.

```
tt = topTraitsOfTargets("MTF1", TFutils::tftColl, ebicat37)
```

```
## remapping identifiers of input GeneSetCollection to Symbol...
```

```
## done
```

```
head(tt)
```

```
##                            DISEASE.TRAIT MAPPED_GENE       SNPS CHR_ID
## 1                       Atopic dermatitis        TNXB rs41268896      6
## 2                       Atopic dermatitis        TNXB rs12153855      6
## 3                       Atopic dermatitis       KIF3A  rs2897442      5
## 4 Attention deficit hyperactivity disorder      SEMA3A   rs797820      7
## 5 Attention deficit hyperactivity disorder        DNM1  rs2502731      9
## 6 Attention deficit hyperactivity disorder        GPC6  rs7995215     13
##     CHR_POS
## 1  32102292
## 2  32107027
## 3 132713335
## 4  83979723
## 5 128214278
## 6  93756253
```

```
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).

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 6 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 underling data is sourced from Sonawane *et al*. 2017 https://doi.org/10.1016/j.celrep.2017.10.001[9]

Software availability

Source code is available from GitHub: https://github.com/vjcitn/TFutils

Archived source code: https://doi.org/doi:10.18129/B9.bioc.TFutils[12]

Licence: Artistic License 2.0

Grant information

Support for the development of this software was provided by the National Institues of Health [U01 CA214846 to VC, U24 CA180996], the Chan Zuckerberg Initiative [DAF 2018-183436 to VC, R01 NHLBI HL118455 to BR] and NIH/NCI/ITCR R21 CA209940, NIH/NIAID R03 AI131066, U01 CA214846 collaborative set aside to NP.

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

References

1. Lambert SA, Jolma A, Campitelli LF, *et al.*: **The Human Transcription Factors.** *Cell.* 2018; **172**(4): 650–665.
PubMed Abstract | Publisher Full Text

2. Weirauch MT, Yang A, Albu M, *et al.*: **Determination and inference of eukaryotic transcription factor sequence specificity.** *Cell.* 2014; **158**(6): 1431–1443.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Ashburner M, Ball CA, Blake JA, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet.* 2000; **25**(1): 25–29.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Kulakovskiy IV, Vorontsov IE, Yevshin IS, *et al.*: **HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis.** *Nucleic Acids Res.* 2018; **46**(D1): D252–D259.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Subramanian A, Tamayo P, Mootha VK, *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A.* 2005; **102**(43): 15545–15550.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Gertz J, Savic D, Varley KE, *et al.*: **Distinct properties of cell-type-specific and shared transcription factor binding sites.** *Mol Cell.* 2013; **52**(1): 25–36.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Wingender E, Schoeps T, Haubrock M, *et al.*: **TFClass: expanding the classification of human transcription factors to their mammalian orthologs.** *Nucleic Acids Res.* 2018; **46**(D1): D343–D347.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif.** *Bioinformatics.* 2011; **27**(7): 1017–1018.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Sonawane AR, Platig J, Fagny M, *et al.*: **Understanding Tissue-Specific Gene Regulation.** *Cell Rep.* 2017; **21**(4): 1077–1088.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Shannon P, Richards M: ***MotifDb*: An Annotated Collection of Protein-DNA Binding Sequence Motifs.** R package version 1.26.0.
Publisher Full Text

11. Ou J, Wolfe SA, Brodsky MH, *et al.*: **motifStack for the analysis of transcription factor binding site evolution.** *Nat Methods.* 2018; **15**(1): 8–9.
PubMed Abstract | Publisher Full Text

12. Carey V, Gopaulakrishnan S: **TFutils: TFutils.** R package version 1.2.0. 2018.
http://www.doi.org/10.18129/B9.bioc.TFutils

Open Peer Review

Current Peer Review Status: ✔ ✔ ✘

Version 2

Reviewer Report 18 July 2019

https://doi.org/10.5256/f1000research.21137.r49980

✘ **Kevin Ernst**
Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
Matthew T. Weirauch
Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center (CCHMC) , Cincinnati, OH, USA

The authors present TFutils, a Bioconductor package for analyze TFs and their binding sites. TFutils combines information and analysis capabilities from several popular sources into one interface. The authors present use-cases and tutorials for how TFutils can be used to ask and answer several biological questions.

Major Comments

1. From what I can tell, the manuscript is basically a copy-pasted version of the “vignette” from their TFutils R package on Bioconductor, probably written by the (co)author of the package. If the purpose is to teach people how to use their TFutils package, or even compare/contrast it with other existing solutions, in my opinion the manuscript falls very short of this goal. It’d be different if the vignette (or the paper) made a clear case for what TFutils does right up front. Apart from saying it doesn’t do sequence logo visualizations, the introduction lacks any discussion of what good the package is, except that it “assembles various resources intended to clarify and unify approaches to working with TF concepts in bioinformatic analysis”. A helpful remedy for this would be to have a table of included functions, data structures, and sample datasets, similar to the auto-generated listing from their own package, which is already publicly-available on the web. The “Intro” section (actually the whole paper) lacks any straightforward enumeration of the contents of the package like “we include data structures (name them) that wrap X, Y, and Z databases, and functions alpha, beta, and gamma to do (whatever).”

2. The manuscript itself is very hard to follow. It would be much better if there were with a common, practical theme through the examples – as written, they are quite disjoint and not very well explained. There is no unifying thread to any of these examples, no focus on telling a “story” with a real-life research inquiry, or making the case for how TFutils could make that inquiry easier. If people in my lab (a transcription factor bioinformatics lab) have trouble following it, then there is almost no chance that a non-expert could follow it. One thing that would really help would be more

human language in the code samples parts (or code comments), in order to aid in reader understanding. As it stands, the text surrounding the code is choppy, and does little to illuminate what's going on in the code samples. The examples and the text surrounding them are sometimes painfully disjoint, to the point where I was wondering how the text and code and figures right next to each other were related.

3. I would strongly urge the authors to use the Lambert *et al.* collection of 1,639 human TFs as their basis. Bigger is not always better. Although the other databases currently used collectively sum to many more human TFs, there are many false positives. For example, GO includes things like kinases in this category. This is exactly the reason that Lambert and colleagues went to the painstaking efforts of collecting all human TF candidates and manually curating the list one by one.

4. Table 1 is comparing apples to oranges. Cis-BP contains all human TFs, regardless of their motif status (i.e. even if a TF does not have a known DNA binding motif, it is still included in the database). HOCOMOCO, on the other hand, only includes TFs with motifs. So, it does not really make sense to compare their TF members, which seems to be the major point of Table 1. This also applies to the following comment in the Discussion, which also does not really make sense based on these facts: "is noteworthy that the four resources used for Figure 1 agree on names of only 119 TFs. The fact that CIS-BP 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." The only reason AHRR is not in the other databases is probably because it has not had its motif directly determined through experimentation.

5. The example functionality shown in Figure 6 does not make sense to me. As I understand it, a "TF target" is from MSigDB, which is simply a predicted binding site for a TF (here, VDR), in the promoter of a gene (e.g., TBX2 in the top row of the figure). TBX2 is associated with, e.g. Creatine Levels. But, we do not know where the GWAS signal is located relative to TBX2. The GWAS signal could be 20,000 bases away from TBX2, or inside its intron, etc. So, what is the connection here between binding of VDR to the promoter of TBX2 and the GWAS signal? Unless I am misunderstanding how this part of the tool is working, this analysis seems very misleading to me.

Minor comments

1. Introduction: "typically near gene promoter regions" – I would add "and enhancers", since this is actually where the majority of the TF binding sites are located.

2. There was just a new release of Cis-BP (version 2.0) – if its not too much work, I would urge the authors to update their system, since it is a major update. I realize that this can be a lot of work, so I will leave it to the authors to decide if this is worth immediate action.

3. I would be very careful about calling the MSigDB collection "TF targets" – these are not experimentally determined binding events (e.g., through ChIP-seq). These are simply the result of scanning motifs in promoters. I think these should therefore only be referred to as "predicted targets". This seems nitpicky, but I think there is a very important distinction here.

4. It looks like the motifs from FIMO are Cis-BP motifs (these are incorporated into FIMO), which is fine. But according to the example shown, it looks like the IDs might be truncated in your database – for example, the top one is called "M3433_1", which could lead to ambiguities, since it could correspond to "M3433_1.02", "M3433_1.01", "M3433_1.00", etc.

5. It is usually not mentioned whether a function or data structure comes from TFutils or some other R package - a simple inline comment or note in the accompany text would really clear this up.

6. Dependencies are left up to the reader to figure out, which is a bit of a nuisance for someone that is trying to decide whether or not a given package is worth exploring.

7. Cis-BP should be spelled “Cis-BP” (as it is in Determination and inference of eukaryotic transcription factor sequence specificity - Weirauch *et al.* (2014)[1]).

8. Some of the sentence structure and word choices (data is “lodged” in Amazon S3, “harvesting” insights from the data) are inappropriate.

9. There are numerous “typesetting” problems in the paper, most of which, I assume originate with the original R Markdown vignette.

10. Monospace font not used consistently for variable or function names (numerous occurrences) - in many places throughout the text.

11. A bold font is used where a 4th-level headline (####) should be, which when rendered as HTML, creates a run-on with the first sentence of the intended section; e.g.:
Annotated reference to ENCODE ChIP-seq results.encode690
Find TFs that are direct GWAS hits for a given trait.directHitsInCISBP

12. Additional individual nit-picks (there are many) can be found in a marked-up version of v2 of the paper, found here.

References

1. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJM, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR: Determination and inference of eukaryotic transcription factor sequence specificity.*Cell*. 2014; **158** (6): 1431-1443 PubMed Abstract | Publisher Full Text

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

No

Is the description of the software tool technically sound?

No

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?

No

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Gene regulation, bioinformatics, genomics, functional genomics, disease genetics.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 27 June 2019

https://doi.org/10.5256/f1000research.21137.r48671

Giovanna Ambrosini
Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
Philipp Bucher
Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

The authors have followed up on several of our suggestions and this (in our opinion) has significantly improved the article.

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

Is the description of the software tool technically sound?
Partly

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?
Partly

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

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Epigenetics, ChIP-seq, regulatory region annotation, motif analysis, database design, web tools.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 28 May 2019

https://doi.org/10.5256/f1000research.21137.r48672

Lihua Julie Zhu
Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School (UMMS), Worcester, MA, USA
Rui Li
University of Massachusetts Medical School, Worcester, MA, USA

We think that the authors have addressed our major concerns, and we would like to change the status to Approved for the revised version.

There are a couple of typos in the revised version that the authors might want to correct.

1. Figure 4 legend states that there are four cell lines while the figure seems to depict data for 5 cell lines.
2. The authors probably meant “four” instead of “for” in the following sentence under the Discussion section.
   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.”

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

Is the description of the software tool technically sound?
Partly

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?
Partly

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

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 26 April 2019

https://doi.org/10.5256/f1000research.19660.r45003

?

Giovanna Ambrosini
Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
Philipp Bucher
Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

TFutils is a Bioconductor package meant to help users study TF binding in the human genome. The tool integrates several resources such as Gene Ontology (GO), CISBP, HOCOMOCO, and MSigD (the Molecular Signature Database). The paper describes how to tackle basic problems users are faced with when trying to work with TFs, in particular the TF classification, the gene targets identification, and, ultimately, the prediction of TF binding affinities.

This article looks more like a software tutorial than a scientific article. As software has to evolve in order keep up with user needs, the text will have to be updated on a regular basis in the future, in order to remain up-to-data as well. This is fine if F1000Research accepts updates and supports versioning of articles. Otherwise, another format should be chosen for presenting this tool.

Just by reading the article, we didn't get a clear impression of what is inside TFutils. Going through the command examples was helpful in this respect. Nevertheless, we have doubts whether we would be able to use this package in a productive manner in the future. The promise that "TFutils lowers the barriers of usage of key findings of TF biology" holds only for expert users of Bioconductor, who are already familiar with all the other packages mentioned in this article and necessary to reproduce the results.

The current manuscript has several shortcomings. At a general level, it is not very transparent to the naïve reader what is actually new from this package and what functionalities are provided by the many other Bioconductor packages referred to in the text. Fortunately, we found a well-organized reference manual for TFutils version 1.2.0 on the internet, which clarified this issue for us. A URL to this document should have been included in the article.

As this is a tutorial-style document, it would be helpful to provide complete R code for reproducing Figures 3 and 4. While Figure 3 is relatively easy to generate, it took us at least half a day to reproduce Figure 4. A major limitation is that the fimo16 object, upon which Figure 3 is based, contains only TF affinity data for 16 out of 689 scanned TF motif matrices.

Figure 3 shows the predicted binding sites for 16 TFs in a selected genomic region. Already with such a small number of TFs, the Figure is pretty crowded with dots. One wonders what it would look like if all 689 FIMO-scanned motif matrices were considered. In view of the density of motif matches it seems doubtful whether any biological insights can be gained from such a plot. Some guidance for the interpretation is needed.

Figure 4 shows ENCODE binding peaks for CEBPB in the same genomic region that was used for Figure 3. Naturally, we were curious to know whether the peaks seen in this Figures co-localize with corresponding motif matches in Figure 3. Unfortunately, CEBPB is not included in the fimo16 collection. To exemplify the power of the tool, it would have been preferable to choose an example where the reader can crosscheck the consistency between predictions and experiments via comparison of Figure 3 with Figure 4.

Overall, our impression is that TFutils is a useful package albeit for a restricted community of users already familiar with the other Bioconductor packages mentioned in the article. However, the manuscript could benefit from major revisions as pointed out above.

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?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Epigenetics, ChIP-seq, regulatory region annotation, motif analysis, database design, web tools.

We confirm that we have read this submission and 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.

Author Response 08 May 2019

Vincent Carey, Brigham and Women's Hospital, Harvard Medical School, Boston, USA

We are very appreciative of the effort underlying this review. Important questions were raised and we endeavor to answer them fully below. Reviewer comments are prefaced by "**QUERY**" and our replies are prefaced by "**RESPONSE**".

QUERY: This article looks more like a software tutorial than a scientific article. As software has to evolve in order keep up with user needs, the text will have to be updated on a regular basis in the future, in order to remain up-to-data as well. This is fine if F1000Research accepts updates and supports versioning of articles. Otherwise, another format should be chosen for presenting this tool.

RESPONSE: Because F1000Research accepts updates and supports versioning of articles, we believe that this format is an acceptable one for the work we have described.

QUERY: Just by reading the article, we didn't get a clear impression of what is inside TFutils. Going through the command examples was helpful in this respect. Nevertheless, we have doubts whether we would be able to use this package in a productive manner in the future. The promise that "TFutils lowers the barriers of usage of key findings of TF biology" holds only for expert users of Bioconductor, who are already familiar with all the other packages mentioned in this article and necessary to reproduce the results.

RESPONSE: We are glad that the examples in the paper were useful to the reviewer. It is true that Bioconductor software generally requires acquaintance with and use of multiple interrelated packages. In this sense, it is different from a relatively common approach of command-line utility implementation of bioinformatic analysis tools, where a given tool may be understood and used in isolation. We do not agree that "expert" level of understanding of Bioconductor is necessary to make use of the material described, although some facility with and enthusiasm for the R language would be necessary to make headway. The paper is published in the Bioconductor channel of F1000research, and it is expected that the readership will have an acquaintance with the resources and limitations of the Bioconductor software and data ecosystem.

QUERY: The current manuscript has several shortcomings. At a general level, it is not very transparent to the naïve reader what is actually new from this package and what functionalities are provided by the many other Bioconductor packages referred to in the text. Fortunately, we found a well-organized reference manual for TFutils version 1.2.0 on the internet, which clarified this issue for us. A URL to this document should have been included in the article.

RESPONSE: This is a useful observation. We have now included a reference to http://bioconductor.org/packages/release/bioc/manuals/TFutils/man/TFutils.pdf in the implementation section of the paper.

QUERY: As this is a tutorial-style document, it would be helpful to provide complete R code for reproducing Figures 3 and 4. While Figure 3 is relatively easy to generate, it took us at least half a day to reproduce Figure 4. A major limitation is that the fimo16 object, upon which Figure 3 is based, contains only TF affinity data for 16 out of 689 scanned TF motif matrices.

RESPONSE: We appreciate the effort taken here. The production of Figure 4 was complicated and we have created an app and associated github repository to clarify the basic issues. The repository is https://github.com/vjcitn/encdemo and the face page of the repo has a screenshot of the app, which runs at https://vjcitn.shinyapps.io/encdemo/. Our point in the visualization of Figure 4 is not the visualization per se, which can be accomplished with standard genome browsers, with suitable commands. Rather, we use the visualization to give concrete demonstration of the immediate programmatic availability (to users of this package) of the relevant experimental results and annotations. We concur that the limitation of fimo16 to a small number of TFs is disappointing. Comprehensive presentation of the scan scores to our user base/readership would require computational resources that we have not yet been able to muster. A local deployment requires close to a terabyte of indexed storage.

QUERY: Figure 3 shows the predicted binding sites for 16 TFs in a selected genomic region. Already with such a small number of TFs, the Figure is pretty crowded with dots. One wonders what it would look like if all 689 FIMO-scanned motif matrices were considered. In view of the density of motif matches it seems doubtful whether any biological insights can be gained from such a plot. Some guidance for the interpretation is needed.

RESPONSE: As noted just above we do not have a mechanism for providing all 689 scans. We have added text after Figure 3 acknowledging the challenge of interpretation, specifically with respect to combinatorics of TF binding.

QUERY: Figure 4 shows ENCODE binding peaks for CEBPB in the same genomic region that was used for Figure 3. Naturally, we were curious to know whether the peaks seen in this Figures co-localize with corresponding motif matches in Figure 3. Unfortunately, CEBPB is not included in the fimo16 collection. To exemplify the power of the tool, it would have been preferable to choose an example where the reader can crosscheck the consistency between predictions and experiments via comparison of Figure 3 with Figure 4.

RESPONSE: We agree that unification of concepts underlying Figure 3 and 4 would be quite desirable. Figure 3 is based on the analysis of motifs in reference sequence, while Figure 4 is a severe reduction of cell-type specific information from in vitro experiments. The data underlying the encdemo app noted above should be useful for beginning surveys across TFs and across cell types essential for a full understanding of cell-type specific combinatorics of TF binding.

QUERY: Overall, our impression is that TFutils is a useful package albeit for a restricted community of users already familiar with the other Bioconductor packages mentioned in the article. However, the manuscript could benefit from major revisions as pointed out above.

RESPONSE: We appreciate the effort taken in this review and we have endeavored to answer the questions raised.

Competing Interests: None.

Reviewer Report 20 February 2019

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

?

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

1. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B: JASPAR: an open-access database for eukaryotic transcription factor binding profiles.*Nucleic Acids Res*. 2004; **32** (Database issue): D91-4 PubMed Abstract | Publisher Full Text
2. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüss M, Reuter I, Schacherer F: TRANSFAC: an integrated system for gene expression regulation.*Nucleic Acids Res*. 2000; **28** (1): 316-9 PubMed Abstract
3. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J: hPDI: a database of experimental human protein-DNA interactions.*Bioinformatics*. 2010; **26** (2): 287-9 PubMed Abstract | Publisher Full Text
4. Newburger DE, Bulyk ML: UniPROBE: an online database of protein binding microarray data on protein-DNA interactions.*Nucleic Acids Res*. 2009; **37** (Database issue): D77-82 PubMed Abstract | Publisher Full Text
5. Ou J, Wolfe SA, Brodsky MH, Zhu LJ: motifStack for the analysis of transcription factor binding site evolution.*Nat Methods*. 2018; **15** (1): 8-9 PubMed Abstract | Publisher Full Text

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 confirm that we have read this submission and 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.

Author Response 08 May 2019

Vincent Carey, Brigham and Women's Hospital, Harvard Medical School, Boston, USA

We appreciate the careful reading of this paper, which has inspired additional functionality that will be added to the package. Reviewer comments (isolated with **QUERY**: tag) are addressed below (with **RESPONSE**: tag).

QUERY: 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 JASPAR1, Transfac2, HDPI3 and uniPro4?

RESPONSE: According to http://cisbp.ccbr.utoronto.ca/summary.php?by=4&orderby=MSource_Identifier JASPAR and TRANSFAC entries are included in CIS-BP. UniProbe is an interesting resource but does not provide a mapping from TF or motif to target genes. Interfacing to UniProbe is beyond the scope of the tasks intended for this paper. The same issue appears to us to apply to hPDI. Resources that do not enumerate "gene level" TF targets are beyond the scope of current work.

QUERY: 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 motifs5 and MotIV. It will be helpful to present a few lines of code to show how to integrate

data from TFutils to aforementioned pipelines.

RESPONSE: We will add an example of how diversity in TF reference motif data leads to a structure that may be visualized with MotifDb/motifStack. Specifically the 1.5.x+ versions of TFutils will include a function that uses motifStack and MotifDb to generate displays like Figure 5 of the revised paper. This is the result of example(tffamCirc.plot)in the current devel branch of TFutils.

QUERY: 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.

RESPONSE: Some of these changes are made. Our approach to the narrative attempts to follow the F1000Research schema.

QUERY: 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?

RESPONSE: The package will be updated according to the Bioconductor release protocol. The main pages give explicit information on provenance of information underlying serialized data structures. Community input on the utility of the various sources will be important in determining the frequency of content updates.

QUERY: 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.

RESPONSE: The code is open source. Pull requests are welcome. If there are specific features of interest to the reviewers we will consider how to incorporate them in future versions of the package/manuscript.

QUERY: 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.

RESPONSE: Totals are added in the caption of Figure 1.

QUERY: 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

RESPONSE: The python package addresses "tensorflow", which is not related to TFs in bioinformatics. We do not believe that the risk of confusion is high, but will consider renaming if events of confusion are observed.

QUERY: 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.

RESPONSE: The presentation follows the suggested F1000Research format.

QUERY: 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.
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.

RESPONSE: The last paragraph of the introduction is used to pre-empt potential reader disappointment early in the presentation, so we prefer to leave it where it is. The third paragraph in the discussion does include content of general and possibly introductory significance, but that content is embedded in concrete illustrations that depend upon actual details of package use. The "summary" element is provided to give the reader a break before plunging into the obligatory Methods/Implementation material. We concur with your basic aesthetic preferences but have organized our text in what we consider to be a rational way.

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

QUERY: 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.

RESPONSE: metadata(encode690) provides details. The ENCODE data are derived from Bioconductor's AnnotationHub package. These facts are noted in the vicinity of Figure 4. We have added a phrase to the abstract that mentions the ENCODE interface.

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

RESPONSE: These authors describe how FIMO was used to obtain sequence-based binding affinity scores; the main text is slightly modified to clarify the role of this reference.

QUERY: 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.

RESPONSE: We agree that the gene model displays are sub-optimal. However, the visualizations are not central to the package. Ideally, Gviz, ggbio, or karyploteR infrastructure would be used and we will pursue these improvements in updates to the package. Further discussion of the visualization is conducted with the other referee report. We note that the interactive "app" at https://vjcitn.shinyapps.io/encdemo/ can be used to interactively view binding sites for a small

number of TFs in a small number of cell types. Such visualizations can be better accomplished with standard genome browsers. The visualizations in the paper are provided to make concrete the immediate programmatic availability of these resources and concepts to package users. In particular, Figure 3 just scratches the surface of the concept that the combinatorics of TF binding are complex. As noted by the other reviewer, the display is impossible to parse meaningfully in detail, but the overall interpretation is that binding events form a complex ensemble and that data structures and programming patterns are needed to develop compelling interpretations.

QUERY: Please include the used R packages in the citation.

RESPONSE: After running the Rmarkdown document in TFutils/vignette/TFutils_f1000 that conducts all the computations presented in the paper, sessionInfo() can be run to enumerate all packages in use. There are 37 packages attached, and 60 loaded but not attached. It does not seem reasonable to burden the paper with such an accounting. Users can run sessionInfo() and then query the DESCRIPTION files of packages of interest for provenance information.

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

RESPONSE: Done.

QUERY: 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 …”.

RESPONSE: Done.

Competing Interests: No competing interests were disclosed.