TFutils: Data Structures for Transcription Factor Bioinformatics

Shweta Gopaulakrishnan[1] **and Vincent Carey**[1]

[1]**Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School**

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 content to enhance the precision and utility of integrative analyses that have components involving TFs.

Keywords

Bioinformatics, DNA transcription, Transcription factors

Introduction

A central concern of genome biology is improving understanding of gene transcription. 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 TFs, and opportunities for doing so in systems biology contexts, are reviewed in Weirauch et al. [1].

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 3.6. The package can be installed with

```
library(BiocInstaller) # use source("http://www.bioconductor.org/biocLite.R") if not available
biocLite("TFutils")
```

Enumerating transcription factors

Various sources of human tfs

```
## 'select()' returned 1:many mapping between keys and columns
```

We have four basic enumerations of TFs with diverse forms of metadata.

```
TFs_GO
```

```
## TFutils TFCatalog instance GO.0003700
##  1068 native Ids, including
## 165 ... 110354863
##  935 unique HGNC tags, including
## AEBP1 AHR ... ZNF765-ZNF761 ZNF660-ZNF197
```

```
TFs_MSIG
```

```
## TFutils TFCatalog instance MsigDb.TFT
##  615 native Ids, including
## AAANWWTGC_UNKNOWN ... GCCATNTTG_YY1_Q6
##  196 unique HGNC tags, including
## MYOD1 TCF3 ... USP7 YY1
```

```
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
```

```
TFs_HOCO
```

```
## TFutils TFCatalog instance hocomoco11
##  771 native Ids, including
## AHR_HUMAN.H11MO.0.B ... ZSCA4_HUMAN.H11MO.0.D
##  680 unique HGNC tags, including
## AHR AIRE ... ZSCAN31 ZSCAN4
```

GO: 820 HOCOMOCO: 680 CIS-BP: 1734 (how many map to HGNC)? MSigDb TFclass

A simple way of enumerating genes coding for TFs is to interrogate Gene Ontology Annotation. In Bioconductor 3.6, the annotations are derived from the November 2017 latest-lite table. The number of distinct gene symbols annotated to the term *DNA binding transcription factor activity* is found as

```
1
```

```
## [1] 1
```

These annotations are accompanied by evidence codes.

Another relevant resource is the HOCOMOCO project (Kulakovskiy et al. [2]). In the conclusion of the 2018 *Nucleic Acids Research* paper, these authors indicate that their database identifies 680 human TFs.

Enumerating TF targets

The Broad Institute MSigDb (Subramanian et al. [3]) includes a gene set collection devoted to cataloging TF targets. We have used Bioconductor's *GSEABase* package 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)
```

Names of TFs for which target sets are assembled are encoded in a somewhat systematic way. We attempt to decode with string operations:

```
tftn = names(TFutils::tftColl)
stftn = strsplit(tftn, "_")
```

So there are some exact matches between components of the MSigDb TF target collection names and the HOCOMOCO TF names. However, we observe some peculiarity in nomenclature in the MSigDb labels:

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

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

Some manual curation will be in order to improve the precision with which MSigDb TF target sets can used.

Quantitative data on TF binding sites

References

[1] Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J.M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431–1443, 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.08.009. URL `http://linkinghub.elsevier.com/retrieve/pii/S0092867414010368`.

[2] Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Ruslan N. Sharipov, Alla D. Fedorova, Eugene I. Rumynskiy, Yulia A. Medvedeva, Arturo Magana-Mora, Vladimir B. Bajic, Dmitry A. Papatsenko, Fedor A. Kolpakov, and Vsevolod J. Makeev. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, 2018. ISSN 13624962. doi: 10.1093/nar/gkx1106.

[3] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506580102. URL `http://www.pnas.org/content/102/43/15545`.