Supplementary Notes

for Liska O, Bohár B, Hidas A, Korcsmáros T, Papp B, Fazekas D & Ari E: “TFLink: An integrated gateway to access transcription factor - target gene interactions for multiple species”

1. Table 1: Species-specific summary statistics of TFLink data

Organism	Scale	Nr. of TFs[1]	Nr. of target genes	Nr. of interactions	Nr. of binding sites[2]	Nr. of binding sequences
Homo sapiens	small-scale	839	4,680	16,634	35,445	35,633
	large-scale	1,348	20,120	6,722,723	8,857,060	8,870,892
	total[3]	1,606	20,139	6,739,357	8,892,505	8,906,525
Mus musculus	small-scale	846	2,503	8,687	10,537	10,589
	large-scale	711	21,263	4,048,895	363,228	380,417
	total	1,156	21,536	4,057,582	373,765	391,006
Rattus norvegicus	small-scale	6	6	8	179	196
	large-scale	51	13,525	81,221	0	166
	total	56	13,530	81,229	179	362
Danio rerio	small-scale	0	0	0	0	0
	large-scale	17	13,769	25,960	0	0
	total	17	13,769	25,960	0	0
Drosophila melanogaster	small-scale	173	281	699	3,262	6,026
	large-scale	462	18,757	367,930	10,948	11,237
	total	527	18,766	368,629	14,210	17,263
Caenorhabditis elegans	small-scale	18	93	109	116	143
	large-scale	282	16,499	315,909	9,746	9,905
	total	289	16,519	316,018	9,862	10,048
Saccharomyces cerevisiae	small-scale	191	2,166	5,349	5	5
	large-scale	317	6,549	232,365	0	0
	total	333	6,549	237,714	5	5

[1] TF: transcription factor
[2] In some cases the *number of binding sites* and the *number of binding sequences* are different because among the data we downloaded from JASPAR there are binding sequences with missing localization, for example when random sequences were investigated with the SELEX.
[3] The total number of transcription factors, target genes, etc. can be different from the sum of small- and large-scale data due to overlapping items.

Organism	Scale	Nr. of TFs[1]	Nr. of target genes	Nr. of interactions	Nr. of binding sites[2]	Nr. of binding sequences
Total	small-scale	2,073	9,729	31,486	49,544	52,592
	large-scale	3,188	110,482	11,795,003	9,240,982	9,272,617

2. Source Database: DoRothEA

Among TFLink interactions when the source of the data was the DoRothEA (1) database we indicate the original source as following:

- “FANTOM4_via_DoRothEA” ~ the FANTOM4 database (2),
- “IntAct_via_DoRothEA” ~ IntAct database (3),
- “KEGG_via_DoRothEA” ~ KEGG database (4),
- “TFactS_via_DoRothea” ~ TFactS database (5),
- “TFe_via_DoRothEA” ~ TFe database (6), and
- “TRRD_via_DoRothEA” ~ TRRD database (7).

3. Description of the TFLink website

3.1 The main page

url: https://tflink.net/[4]

There is a short description – containing links to the source databases – of the TFLink database and summary statistics about its content.

3.2 Entry pages example links

- Transcription factor entry page: https://tflink.net/protein/q9vhm6/
- Target gene protein page: https://tflink.net/protein/p52429/
- Transcription factor and target gene protein page: https://tflink.net/protein/p10242/

[4] The webserver is fully functional, accessible without login credentials, uses the *https* protocol, includes test data (see the FAQ section). The source code of the website is freely available https://github.com/korcsmarosgroup/TFlink. At the FAQ page we provide descriptions on how to use the web server and understand the content and the files available in the database. TFLink will be maintained for at least 5 years after the publication.

3.3 Downloadable file types

Here we list the type and content of all downloadable files can be found on TFLink entry pages and at the ‘Download’ section.

3.3.1 Interaction tables

Interaction table files are tab separated tables (*TSV*) of transcription factor - target gene interactions that contain either interactions validated by small-scale experiments or large-scale experiments or these two data altogether. Interaction tables contain the following data:

1. **UniprotID.TF** and/or[5] **UniprotID.Target:** *Uniprot IDs* of transcription factors and/or target genes
2. **NCBI.GeneID.TF** and/or **NCBI.GeneID.Target:** *NCBI Gene IDs* of transcription factors and/or target genes,
3. **Name.TF** and/or **Name.Target:** gene names of transcription factors and/or target genes,
4. **Detection.method:** names of the detection methods,
5. **PubmedID:** *Pubmed IDs* of the original publications (when available) and the publications of the databases,
6. **Organism:** scientific name of the organism,
7. **Source.database:** names of the original source databases, and
8. **Small-scale.evidence:** indication about if the data were confirmed by small-scale evidence (with “Yes” or “No”),
9. **TF.TFLink.ortho:** *Uniprot IDs* of ortholog transcription factors that are available at the TFLink gateway,[6]
10. **TF.nonTFLink.ortho:** *Uniprot IDs* of ortholog transcription factors that are not available at the TFLink gateway,[6]
11. **Target.TFLink.ortho:** *Uniprot IDs* of ortholog target genes that are available at the TFLink gateway,[6]
12. **Target.nonTFLink.ortho:** *Uniprot IDs* of ortholog target genes that are not available at the TFLink gateway.[6]

---

[5] If the interaction table was downloaded at the ‘Download’ section the file contains both TF and *Target IDs* and names. If it was downloaded from an entry page, it contains names and IDs of the transcription factor (when downloading “Transcription factors of …” or names and IDs of the target genes (when downloading “Targets of …”).

[6] Each entry consists of a shortened name of the organism (Hs: *Homo sapiens*, Mm: *Mus musculus*, Rn: *Rattus norvegicus*, Dr: *Danio rerio*, Dm: *Drosophila melanogaster*, Ce: *Caenorhabditis elegans*, Sc: *Saccharomyces cerevisiae*) and the *UniProt ID* separated by a colon (*e.g.* Mm:Q3UPW2).

3.3.2 Interaction *MITAB* tables

Interaction *MITAB* files (8) contain transcription factor - target gene interactions in *HUPO-PSI MITAB 2.8* format. *MITAB 2.8* (as defined by the Human Proteome Organization - Proteomics Standards Initiative, HUPO-PSI (9)) is a standardised format – including standardised vocabulary also – used to describe molecular interactions. While other databases may refer to the detection methods by multiple different names, databases that use the *MITAB* format (*e.g.* TFLink, MINT (10) or IntAct (3)) use the same *code* for a given technique. For example, the *electrophoretic mobility shift assay* technique could be identified by both its full name, and the shortened name EMSA, but in databases utilising the *MITAB* format it's always referred to by the *psi-mi:"MI:0413"* code. This makes the identification of interaction properties more efficient and helps avoid potential misunderstandings. The *MITAB* files are tab-delimited tables containing 46 columns and no header. If needed a header file for the *MITAB* tables is available at the TFLink 'FAQ' page. The interaction tables and *MITAB* files can be used as input data for the Cytoscape software (11) to perform systems and network biology studies.

3.3.3 Interaction *GMT* files

Interaction *GMT* (Gene Matrix Transposed) is a tab delimited file format that describes gene sets – target genes of a transcription factor – in each row. The first and second column contains information about the transcription factors (various IDs and gene names). The first cell in each row is always unique. From the third to the last column the target genes of the transcription factor are listed. The number of target genes can vary from transcription factor to transcription factor, therefore the number of cells can be different in every row. The user can choose between *GMT* files with *Uniprot IDs*, *NCBI Gene IDs*, and *gene names*. The *GMT* files are useful for enrichment and gene overrepresentation analyses and can be an input file for the MulEA (https://github.com/koralgooll/MulEA) and GSEA software (12).

3.3.4 Binding site tables

Binding site table files are tab separated tables (*TSV*) of binding site annotations that contain:

1. **TFLinkID:** unique *TFLink IDs* of the binding sites,
2. **UniprotID.TF:** *Uniprot IDs* of the transcription factors,
3. **Name.TF:** gene names of the transcription factors,

4. **Organism:** scientific name of the organism,
5. **Assembly:** version of the genome assembly,
6. **Chromosome:** name of the chromosome,
7. **Start:** start coordinates of the binding sites,
8. **End:** end coordinates of the binding sites,
9. **Strand:** coding strand ("+" indicates the forward strand, and "-" the reverse strand),
10. **Genome.browser:** a hyperlink to the particular genomic location at the UCSC genome browser website (https://genome.ucsc.edu/),
11. **Detection.method:** names of the detection methods,
12. **PubmedID:** *Pubmed IDs* of the original publications (when available) and the publications of the databases,
13. **Source.database:** names of the original source databases,
14. **Small-scale.evidence:** indication about if the data were confirmed by small-scale evidence (with "Yes" or "No"),
15. **Number.of.TFBS.overlaps:** the number of overlapping binding sites of the same transcription factor, and
16. **TFBS.overlaps:** list of *TFLink IDs* of overlapping binding sites of the same transcription factor.

3.3.5 Binding site annotation files

Binding site annotation files contain:

1. **##sequence-region …:** sequence regions with the name, start and end site of chromosomes,
2. **seqid:** names of the chromosomes,
3. **source:** starting with "TFLink_from_" and then the names of the source databases,
4. **type:** "TF_binding_site"
5. **start:** the start coordinates of the binding sites,
6. **end:** the end coordinates of the binding sites,
7. **score:** "."
8. **strand:** the coding strand ("+" indicates the forward strand, and "-" the reverse strand),
9. **phase:** "."
10. **attributes:**
    a. **ID:** unique *TFLink IDs* of the binding sites,
    b. **Name:** names of the transcription factors, and

c. **Note:** *Uniprot IDs* of the transcription factors

in *GFF3* format. For a detailed description of the format, please visit: https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md.

3.3.6 Binding site sequences

Binding site sequence files are *FASTA* files containing the DNA sequences of the transcription factor binding sites. The header of each sequence contains:

1. unique internal *TFLink ID* of the binding site,
2. *Uniprot ID* and gene name of the transcription factor,
3. version of the genome assembly,
4. name of the chromosome, and
5. start and end coordinates of the sites.

3.4 Frequently asked questions

url: https://tflink.net/faq/

At the FAQ part of the TFLink website we tried to answer all possible upcoming questions and help the user to navigate and understand the content and the files available in the database.

References

1. Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., et al. (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.*, **29**, 1363–1375.
2. Kawaji, H., Severin, J., Lizio, M., et al. (2011) Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res.*, **39**, D856–D860.
3. Orchard, S., Ammari, M., Aranda, B., et al. (2014) The MIntAct project: IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
4. Kanehisa, M., Goto, S., Sato, Y., et al. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
5. Essaghir, A., Toffalini, F., Knoops, L., et al. (2010) Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene

expression data. *Nucleic Acids Res.*, **38**, e120.

6. Yusuf, D., Butland, S.L., Swanson, M.I., et al. (2012) The Transcription Factor Encyclopedia. *Genome Biol.*, **13**, R24.
7. Kolchanov, N.A., Ignatieva, E.V., Ananko, E.A., et al. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res.*, **30**, 312–317.
8. Perfetto, L., Acencio, M.L., Bradley, G., et al. (2019) CausalTAB: the PSI-MITAB 2.8 updated format for signalling data representation and dissemination. *Bioinformatics*, **35**, 3779–3785.
9. Sivade Dumousseau, M., Alonso-López, D., Ammari, M., et al. (2018) Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics*, **19**, 134.
10. Licata, L., Briganti, L., Peluso, D., et al. (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
11. Shannon, P., Markiel, A., Ozier, O., et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
12. Subramanian, A., Tamayo, P., Mootha, V.K., et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.