**TFCheckpoint 2.0 guidelines**

***Initial considerations***

The TFCheckpoint database aims to keep a collection of human, mouse and rat orthologs of candidate sequence-specific DNA binding transcription factors (DbTFs) provided by multiple databases. TFCheckpoint is just a list of lists without any confidence or quality score showing how likely a protein is a DbTF. The reference/anchor point is a list of proteins classified as human, mouse or rat DbTFs by at least one database. These candidate DbTFs are mainly identified by their official HGNC symbols when human orthologs are available, or by their official MGI (for mouse) or RGD (for rat) gene symbols when human orthologs are not available.

How to deal with GO terms for protein entries:

* If we add them through Miguel’s scripts they become a liability, so better to link from protein name out.
* But which list to take? This should be the Logie list.
* In addition, there could be a list of names that are lacking in Logie, but are supported by some criteria that still make them interesting:
  + Available but incomplete data that may indicate their role in gene regulation - from GO terms?
  + Their presence on n lists, but skipped by Logie
  + Our personal notes - cases where we ‘caved’

Possible contacts for TFclass: biobase since 2014 part of Quiagen

Mathias Krull <[mkl@biobase.de](mailto:mkl@biobase.de)>

Volker Matys <[Volker.Matys@biobase-international.com](mailto:Volker.Matys@biobase-international.com)>

Torsten Schoeps - [torsten.schoeps@bioinf.med.uni-goettingen.de](mailto:torsten.schoeps@bioinf.med.uni-goettingen.de) database person?

***Brief description of columns and possible issues***

In this section you will find a brief description of each column and possible issues related to that column. **The major issues regarding the construction of TFCheckpoint is (1) finding human, mouse and rat orthologs and (2) mapping among official symbols, Uniprot IDs, Ensembl IDs and EntrezGene IDs. The extraction of DbTFs from sources is, in general, a smooth process.**

1. Human\_mouse\_rat\_ortholog: this is the reference/anchor point of TFCheckpoint as described above. **Issues: some entries are not represented by gene official symbols; instead, these entries are represented by Uniprot IDs (for example, lines 20-63, 65-98, 3743-3848) or gene synonyms (e.g., line 2641) or Genbank IDs (eg, lines 99-183)**.

2. Entrez\_human: Entrez GeneID for the human ortholog. **Issues: some entries are missing.**

3. Entrez\_rat: Entrez GeneID for the rat ortholog. **Issues: some entries are missing.**

4. Entrez\_mouse: Entrez GeneID for the mouse ortholog. **Issues: some entries are missing.**

5. Uniprot\_human: Uniprot ID for the human ortholog. **Issues: some entries are missing and other entries show more than one Uniprot ID; we expect only one Uniprot ID with a reviewed status per entry. If only unreviewed sequences are available, we expect that the largest sequence will be the reference one.**

6. Uniprot\_rat: Uniprot ID for the rat ortholog. **Issues: some entries are missing and other entries show more than one Uniprot ID; we expect only one Uniprot ID with a reviewed status per entry. If only unreviewed sequences are available, we expect that the largest sequence is the reference one.**

7. Uniprot\_mouse: Uniprot ID for the mouse ortholog. **Issues: some entries are missing and other entries show more than one Uniprot ID; we expect only one Uniprot ID with a reviewed status per entry. If only unreviewed sequences are available, we expect that the largest sequence is the reference one.**

8. Sequence-specific DNA binding activity? (GO:0043565 and children): if at least one of the orthologs is annotated with the term GO:0043565 or any of its children, then the value is 1; otherwise, is 0. Such information is extracted from QuickGO and the main identifier for mapping purpose here is **Uniprot ID**.. **Issue: outdated as this information was gathered on November last year.**

9. Annotation information (Evidence type, Pubmed ID, Taxon): contains the triple “evidence type, pubmed ID, taxon ID” imported from QuickGO along with the annotation used to complete the previous column. **Issues: This column can have multiple triples and, for now, these triples are separated by pipes. The main question is: is this a suitable way to show these annotation-associated pieces of information? Will both user and computaters be happy with this? Wouldn’t it be better to simply provide a link to the annotation instead of importing it from QuickGO?**

10. Transcription regulation activity? (GO: 0006357): if at least one of the orthologs is annotated with the term GO:0006357 or any of its children, then the value is 1; otherwise, is 0. Such information is extracted from QuickGO and the main identifier for mapping purpose here is **Uniprot ID**.. **Issue: outdated as this information was gathered on November last year.**

11. Annotation information (Evidence type, Pubmed ID, Taxon): contains the triple “evidence type, pubmed ID, taxon ID” imported from QuickGO along with the annotation used to complete the previous column. **Issues: This column can have multiple triples and, for now, these triples are separated by pipes. The main question is: is this a suitable way to show these annotation-associated pieces of information? Will both user and computers be happy with this? Wouldn’t it be better to simply provide a link to the annotation instead of importing it from QuickGO?**

12. DNA-binding transcription factor activity? (GO:0000981): if at least one of the orthologs is annotated with the term GO:0000981 or any of its children, then the value is 1; otherwise, is 0. Such information is extracted from QuickGO () and the main identifier for mapping purpose is **Uniprot ID**. **Issue: empty for now, information should be retrieved from QuickGO.**

13. Annotation information (Evidence type, Pubmed ID, Taxon): contains the triple “evidence type, pubmed ID, taxon ID” imported from QuickGO along with the annotation used to complete the previous column. **Issues: This column can have multiple triples and, for now, these triples are separated by pipes. The main question is: is this a suitable way to show these annotation-associated pieces of information? Will both user and computers be happy with this? Wouldn’t it be better to simply provide a link to the annotation instead of importing it from QuickGO?**

14. Lambert et al, 2018: compilation of human proteins classified as DbTFs by Lambert et al (2018) (<http://humantfs.ccbr.utoronto.ca>). The values are “Lambert\_yes”, “Lambert\_no” or “-”. The main identifier for mapping purpose here is **Ensembl ID**. HGNC symbols are also provided, but some of them are outdated; so, it is advisable to disregard them here. No detected issues here.

15. Vaquerizas et al: compilation of human proteins classified as DbTFs by Vaquerizas et al (2009) (list in this [Supplementary File](https://static-content-springer-com.proxy.bnl.lu/esm/art%3A10.1038%2Fnrg2538/MediaObjects/41576_2009_BFnrg2538_MOESM6_ESM.txt)). The values are “Vaquerizas\_a”, “Vaquerizas\_b”, “Vaquerizas\_c”, “Vaquerizas\_other”, “Vaquerizas\_x” (“a”, “b”, “c”,”other” and “x” mean different confidence levels as explained [here](https://static-content-springer-com.proxy.bnl.lu/esm/art%3A10.1038%2Fnrg2538/MediaObjects/41576_2009_BFnrg2538_MOESM4_ESM.pdf)) or “-” . The main identifier for mapping purpose here is **Ensembl ID.** HGNC symbols are also provided, but some of them are outdated; so, it is advisable to disregard them here. No detected issue here. No detected issue here

16. TF\_Class\_human: compilation of human proteins classified as DbTFs by the TFClass database. The source of information is a turtle file (a [syntax](https://en.wikipedia.org/wiki/Syntax_(programming_languages)) and [file format](https://en.wikipedia.org/wiki/File_format) for expressing data in the [Resource Description Framework](https://en.wikipedia.org/wiki/Resource_Description_Framework) [RDF] data model) found [here](http://tfclass.bioinf.med.uni-goettingen.de/suppl/tfclass.ttl.gz). The main identifier for mapping purpose here is **Uniprot ID**. Many other identifiers are offered, but they are messy.

17. TF\_Class\_mouse: compilation of mouse proteins classified as DbTFs by the TFClass database. The source of information is a turtle file (a [syntax](https://en.wikipedia.org/wiki/Syntax_(programming_languages)) and [file format](https://en.wikipedia.org/wiki/File_format) for expressing data in the [Resource Description Framework](https://en.wikipedia.org/wiki/Resource_Description_Framework) [RDF] data model) found [here](http://tfclass.bioinf.med.uni-goettingen.de/suppl/tfclass.ttl.gz). The main identifier for mapping purpose here is **Uniprot ID**. Many other identifiers are offered, but they are messy. **Issues: still empty. The mouse Uniprot IDs have been already extracted, but using these IDs to find their human orthologs has not been trivial. This is why this column is empty.**

18. TF\_Class\_rat: compilation of rat proteins classified as DbTFs by the TFClass database. The source of information is a turtle file (a [syntax](https://en.wikipedia.org/wiki/Syntax_(programming_languages)) and [file format](https://en.wikipedia.org/wiki/File_format) for expressing data in the [Resource Description Framework](https://en.wikipedia.org/wiki/Resource_Description_Framework) [RDF] data model) found [here](http://tfclass.bioinf.med.uni-goettingen.de/suppl/tfclass.ttl.gz). The main identifier for mapping purpose here is **Uniprot ID**. Many other identifiers are offered, but they are messy. **Issues: still empty. The rat Uniprot IDs have been already extracted, but using these IDs to find their human orthologs has not been trivial. This is why this column is empty.**

19. AnimalTFDB:

20. DBD:

21. JASPAR:

22. ORFeome:

23. Ravasi et al:

24. TcoF-DB\_DbTF: Collection of human proteins classified as DbTFs according to the TcoF database criteria

25. TcoF-DB\_coTF\_human: Collection of human proteins classified as coTFs according to the TcoF database criteria. **Issue: as TFCheckpoint aims to hold a collection of proteins considered DbTFs, so there is no sense in keeping this column. The content of this column could be simply moved to the previous one.**

26. TcoF-DB\_coTF\_mouse: Collection of mouse proteins classified as coTFs according to the TcoF database criteria. **Issue: as TFCheckpoint aims to hold a collection of proteins considered DbTFs, so there is no sense in keeping this column.**

27. TFCat:

***Some questions (this is a non-comprehesive list)***

1. Should we add the following source, namely, [Saeed et al 2014](https://www.ncbi.nlm.nih.gov/pubmed/25258085) (list prepared by Colin), found in Ruth’s list? This is not the current Colin’s list; this is the original one from 2014.

2.

***Constructing TFCheckpoint 2.0 from scratch***

Find below a tentative pipeline for constructing TFCheckpoint 2.0 from scratch if needed.

1. Download source files from the databases of interest and QuickGO

2. From databases: extract only human, mouse and rat proteins classified as DbTFs

3. XX

4. XX

Martin’s summary of TFcheckpoint 2.0:

Mapping table:

HGNC approved gene symbol

HGNC gene name

entrez\_human

entrez\_rat

entrez\_mouse

Uniprot\_Human

Uniprot\_Rat

Uniprot\_mouse

Lambert et al, 2018.<https://www.ncbi.nlm.nih.gov/pubmed/29425488>

Saeed, 2014.<https://www.ncbi.nlm.nih.gov/pubmed/25258085>

Schmeier, 2016.<https://www.ncbi.nlm.nih.gov/pubmed/29425488>

Vaquerizas et al, 2009. Census<https://www.ncbi.nlm.nih.gov/pubmed/19274049>

Wingender et al, 2018.<https://www.ncbi.nlm.nih.gov/pubmed/29087517>

TFClass\_human

TFClass\_mouse

TFClass\_rat

AnimalTFDB 3.0.<https://www.ncbi.nlm.nih.gov/pubmed/30204897>

DBD.<https://www.ncbi.nlm.nih.gov/pubmed/18073188>,<http://www.transcriptionfactor.org/index.cgi?Home>

JASPAR.<http://jaspar.genereg.net/>,<https://www.ncbi.nlm.nih.gov/pubmed/31701148>

ORFeome.<https://www.ncbi.nlm.nih.gov/pubmed/15489324>

Ravasi.et.al.<https://www.ncbi.nlm.nih.gov/pubmed/20211142>

Schmeier, 2017.<https://www.ncbi.nlm.nih.gov/pubmed/27789689>

TcoF-DB v2x

TcoF.DB\_DbTF

TcoF.DB\_coTF\_human

TcoF.DB\_coTF\_mouse

TFCat

Logie et al, being built, linked through Uniprot IDs

GO terms:

Sequence-specific DNA binding activity? (GO:0043565 and children)

Annotation information (Evidence type, Pubmed ID, Taxon)

Transcription regulation activity? (GO: 0006357)

Annotation information (Evidence type, Pubmed ID, Taxon)

DNA-binding transcription factor activity? (GO:0000981)

Annotation information (Evidence type, Pubmed ID, Taxon)