This document is replaced by the publication:
https://www.sciencedirect.com/science/article/pii/S1874939921000705

Gene Ontology guidelines for transcription factor annotation

Authors: Pascale Gaudet, Colin Logie, Ruth Lovering
Date last updated: 2023-10-24

GO has refactored the MF branch representing the activities of proteins involved in transcription. In addition to RNA polymerase, we defined three different types of activities involved in transcription and its regulation:

> **I. GTFs:** General transcription factors, the molecular machine that assembles with the RNA polymerase at the promoter to form the pre-initiation complex (PIC).
> **II. dbTFs:** Specific DNA binding transcription factors that provide genomic addresses and specify the cell types and the conditions under which specific genes are expressed. Central to dbTF function is their binding to specific DNA sequences (often named transcription factor binding sites (TFBS), and
> **III. coTFs:** Transcription coregulators (also known as transcription cofactors) serve multiple functions, such as bridging the GTF and the dbTFs, specifying the regulatory effect of DbTFs, modifying the chromatin structure to render it more or less accessible for transcription. coTFs normally exert their function independent of high affinity binding to specific DNA sequences.

The present guidelines aim to help curators apply the revised transcription terms. Please use more specific child terms whenever possible.

GTFs annotations should include, depending on the evidence available:

- MF
  - GO:0140223 general transcription initiation factor activity
- BP
  - GO:0006351 transcription, DNA-templated
- CC
  - is active in GO:0000785 chromatin

    - part of GO:0097550 transcriptional preinitiation complex

dbTFs annotations may include:

- MF
    - GO:0003700 DNA binding transcription factor activity
        - GO:0001228 DNA-binding transcription activator activity, RNA polymerase II-specific
        - GO:0001227 DNA-binding transcription repressor activity, RNA polymerase II-specific
    - GO:0000987 cis-regulatory region sequence-specific DNA binding
        - GO:0000978 RNA polymerase II cis-regulatory region sequence-specific DNA binding
- BP
    - GO:0006355 regulation of transcription, DNA-templated,
        - GO:0045893 positive regulation of transcription, DNA-templated children
        - GO:0045892 negative regulation of transcription, DNA-templated children
- CC
    - is active in GO:0000785 chromatin
    - part of GO:0005667 transcription factor complex

coTFs annotations may include:

Note that coTFs perform a wide range of functions, the common functions are listed below, however, this list is not exhaustive:

- MF
    - GO:0003712 transcription coregulator activity
        - GO:0003713 transcription coactivator activity
        - GO:0003714 transcription corepressor activity
    - GO:0140097 catalytic activity, acting on DNA
        - GO:0009008 DNA-methyltransferase activity
        - etc
    - GO:0140096 catalytic activity, acting on a protein
        - GO:0033558 protein deacetylase activity
            - GO:0004407 histone deacetylase activity
    - GO:0030234 enzyme regulator activity
        - GO:0035034 histone acetyltransferase regulator activity
        - GO:0035033 histone deacetylase regulator activity
    - GO:0140297 DNA-binding transcription factor binding

- BP
    - GO:0006355 regulation of transcription, DNA-templated

        - GO:0031507 heterochromatin assembly
            - GO:0140719 constitutive heterochromatin assembly
            - GO:0140718 facultative heterochromatin assembly
            - GO:0071514 genomic imprinting
            - GO:0033696 heterochromatin boundary formation
        - GO:0016570 histone modification
            - GO:0031056 regulation of histone modification
        - GO:0006304 DNA modification
            - GO:0006306 DNA methylation
                - GO:0044030 regulation of DNA methylation
- CC
    - is active in GO:0000785 chromatin
    - part of GO:0005667 transcription factor complex

Other transcription regulator activities

There are also proteins that inhibit dbTFs, for example by sequestering them in the cytoplasm or nucleus ('dbTF-inhibitors'). The difference between coTFs and dbTF regulators is that the latter do not act *at* the genomic location of target regulated gene, whereas coTFs are associated with the transcriptional regulatory complex. Therefore, the input of the coTF is the dbTF, not the target gene.

- MF
    - GO:0140416 DNA-binding transcription factor inhibitor activity
    - looking for example for positive regulators
- BP:
    - GO:0006355 regulation of transcription, DNA-templated
- CC
    - as appropriate

Annotation extensions

- MF and BP: target gene of a dbTF or a coTF: has_input [dbTF]
- MF localization: is_active in [GO:cellular component, cell, tissue….]
- BP localization: occurs_in [GO:cellular component, cell, tissue….]

Annotating a transcription regulator from experimental data

The following annotation approach follows the strategy that we currently recommend, to ensure that curators use all information available, and do not restrict to annotating papers individually and out of the more general context. Note that the information does not necessarily need to be extracted from a single paper; reviewing a wide range of papers is recommended to ensure

annotations are as accurate as possible, so that annotations are based on multiple observations from different articles and independent research groups.

The following four questions provide a checklist to assess whether a gene can be annotated as a transcriptional regulator:

1. **What is the starting hypothesis: are the authors characterizing a transcription regulator?** Scientific models are built by adding new data to the existing corpus of evidence. New data can either support or contradict existing models. The introduction section of research articles can be used to understand what prior knowledge the article builds on, and what aspect of the existing model or what new model the authors are assessing. The intent of the authors is essential to understand what GO term should be chosen, with the caveat that inconsistent terminology has been used in transcription research articles and therefore may not always be consistent with the GO term labels. Curators must look carefully at the GO term definitions and the placement of the term in the ontology to ensure that the meaning of the GO term corresponds to the concept being described in the article.

2. **Does knowledge from specific protein domains or characterized orthologs support the hypothesis?**
   The presence of specific domains and the existence of well characterized orthologs can provide useful support for interpreting experimental data. Note that domain information and sequence homology data should be used very carefully: not all domains have a single function; and only closely related orthologs whose function have been unambiguously characterized can be used to support the association of a gene with a GO term, *if those are consistent with the experimental data presented in the article*.

   a. **GTFs:** GTFs have been characterized in several organisms, from bacteria, to yeast, to mammalian cells (PMID:25693126), and therefore orthology should provide strong support for the decision to associate these proteins with a child specific for RNA polymerase I, II or III of the MF term "GO:0140223 general transcription initiation factor activity". In addition, the naming of GTFs is well established across human and model organism nomenclature groups and can be used to help guide these decisions. Thus, for human GTFs the HUGO Gene Nomenclature Committee (HGNC, www.genenames.org) provide the gene symbol TAF#, for TATA-box binding protein associated factors, and GTF2#s and GTF3#s, for general transcription factor II and III subunits respectively, although a few GTFs have more specific names such as BTAF1: B-TFIID TATA-box binding protein associated factor 1.

   b. **dbTFs:** Gene products associated with the GO term "GO:0003700 DNA-binding transcription factor activity" **should have experimental evidence to confirm their ability to bind DNA and that this binding regulates the expression of a**

limited set of target genes. In these cases the direct target gene(s) can also be included in the annotation using the "has input relation". Proteins that belong to families of well characterized transcription factors, such as those that contain GATA and homeobox domains and proteins with a one-to-one ortholog already demonstrated to be a dbTF, weaker evidence of DNA binding, such as ChIP experiments is sufficient. For proteins with domains known to be associated with functions other than DNA binding (such as zinc fingers) or proteins with enzymatic activity, strong evidence of DNA binding is required.

In addition, in some cases, neither the protein nor a member of the protein's family will have been previously associated with the dbTF activity term. In these cases, clear experimental evidence of sequence-specific DNA binding and gene transcription regulation via cognate DNA motifs located in gene-associated *cis*-regulatory modules will be required for the protein to be classified as a dbTFs.

c. **coTFs**: A coTF is defined as a protein that interacts specifically with a dbTF or a coTF at a *cis*-regulatory region (GO:0003712). This interaction either activates or represses the transcription of specific genes, often acting by altering chromatin structure and modifications. There are many roles that coregulators can play: for example, one class of transcription coregulators modifies chromatin structure through covalent modification of histones. A second class of coregulators modifies the conformation of chromatin in an ATP-dependent manner. A third class modulates interactions of dbTFs with other coTFs.

   Many coTFs have enzymatic activity and do not bind DNA. For coTFs that do bind DNA, many recognize very short, common DNA sequences, not sufficiently unique to enable the coTF to regulate the expression of a limited set of genes in a discrete environmental or developmental stage (for example AT-hook coTFs).

   The distinction between a dbTF and a coTF can be difficult to make, so a more exhaustive review of the literature, including looking at the characterized orthologs, is highly recommended before annotating a coTF.

3. **Are there other GO annotations or published experimental results consistent with the hypothesis ?**
   Keeping in mind the gene-by-gene/pathway-by-pathway annotation approach, it can be valuable to take into account results from other research articles, to make sure results and annotations are consistent. Curators should avoid creating annotations that are inconsistent with existing annotations, by either choosing a different term for annotation, or by reviewing and eventually disputing annotations that appear to be incorrect (see next section).

4. **Are the experimental results consistent with the hypothesis ?**
   The curator should carefully look at the results presented in the paper and, if those are consistent with the hypothesis that this is a transcription regulator, then appropriate GO annotations can be made.

Note that DNA-binding transcription factors as well as co-regulators often act as activators or repressors in different promoters or dependent on the context, so annotation to both activator and repressor is not considered inconsistent. This may be further resolved through additional context details, e.g. cell type, environmental conditions, etc.

Reviewing existing annotations

If time allows, it is very useful for the research community and for future annotation that other annotations associated with the gene be reviewed. If there are conflicting annotations, then the supporting data should be reviewed to see whether the annotations are inconsistent with the data, in which case annotations should be fixed.

In cases where the primary data is conflicting across different papers (for example a protein is sometimes described as a transcription factor, and sometimes as a coregulator), then the literature should be reviewed carefully to decide whether:

i. the annotation is incorrect (bad choice of term, wrong protein annotated)
ii. knowledge has evolved. If necessary some (usually older) papers should be marked as ‘do not curate’ and the associated annotations should be removed.
iii. the protein plays multiple roles under different conditions (ie, acts as a DNA-binding transcription factor in certain contexts and as a cofactor on others), as these two molecular functions are not mutually exclusive.
iv. no clear activity has been established yet, in which case: either annotations to both DNA-binding transcription factor and cofactor may be made (if the evidence supports it), or if the data is not sufficient, do not annotate.

GO-CAM representation of transcription

[GO-CAM Modeling Guidelines: DNA binding transcription factor activity]

Potential pitfalls

The annotation of dbTFs can be challenging for multiple reasons:

- Some activities are hard to distinguish from each other, and adding to this difficulty, transcription regulators form complexes that are more or less fluid, so that it can be difficult to detect which protein in a complex is responsible for a specific activity.
- It can be difficult to distinguish certain activities, and some proteins do have multiple functions.
- Researchers use "transcription factor" loosely. It can mean a GTF, a dbTF, or a coTF.
- The terms "cofactor", "coactivator", and "corepressor" are also used for different activities: either as described in this document, and sometimes (especially in older papers), they are used to describe a dbTF that acts as a dimer.
- The complexity of the transcription process means that no single experiment is usually sufficient to define the function of a protein: interpretation of experimental results that investigate dbTFs must rely on existing knowledge.
- Many proteins presumed to function as dbTFs have never been experimentally demonstrated to bind DNA, but their role is indirectly inferred by the presence of specific domains and in some cases, evidence of an effect on the transcription of putative direct target genes.
- The presence of a DNA binding domain in a protein does not always imply the protein functions as a dbTF.

_____ADDITIONAL INFORMATION_____

Additional information

Ontology structure

Molecular Function

MF: GO:0140110 transcription regulator activity and children

The transcription regulator activity (GO:0140110) MF branch of the GO describes the activities of transcription regulators: GTFs, dbTF, and coTFs (Figure 1).

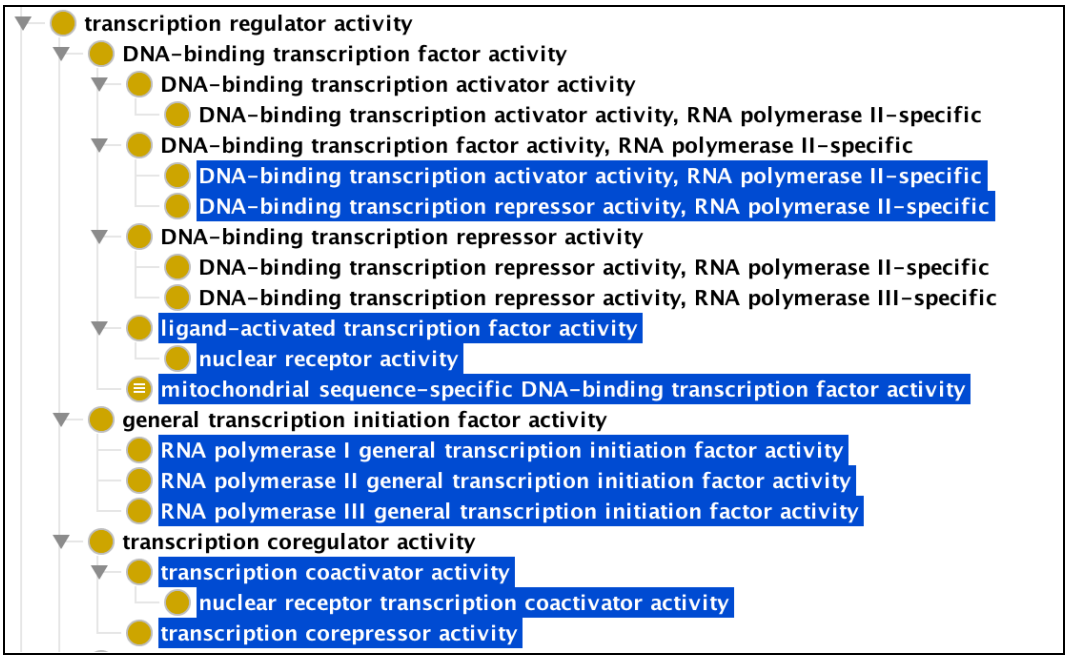

Figure 1. Transcription regulator activity branch of the Gene Ontology. This part of the Molecular Function (MF) ontology describes the activities of transcription regulators: GTFs, dbTF, and coTFs. Highlighted in blue are the most specific GO terms available to describe these activities in eukaryotic cells (prokaryotes use a single polymerase).

MF: GO:0000987 cis-regulatory region sequence-specific DNA binding

The transcription regulatory region sequence-specific DNA-binding sub-tree of GO includes terms describing specific regulatory regions, such as the core promoter (including the TATA box and the transcription start site), cis-regulatory regions (bound by dbTFs), and specific types of cis-regulatory motifs (such as E-box and N-box). Overview of the GO structure for DNA binding activities is shown in Figure 1.

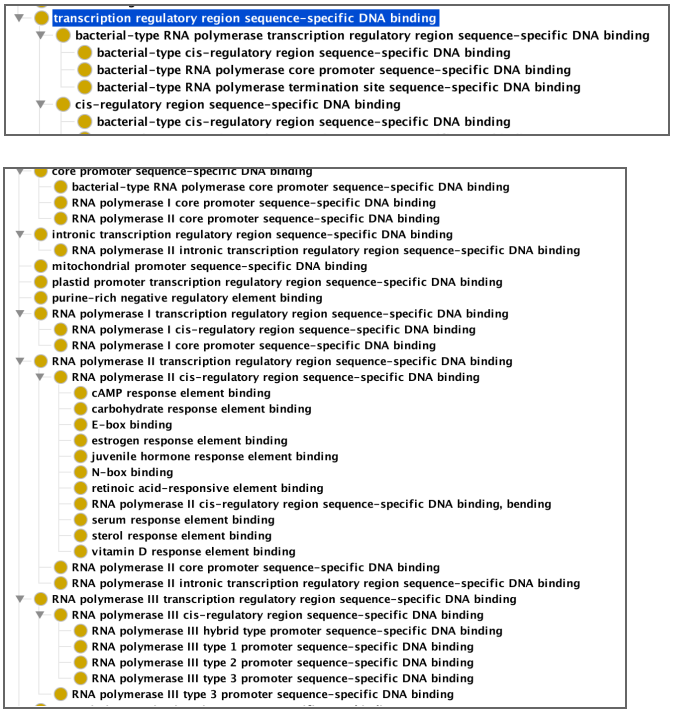

Figure 2. DNA binding branch of the Gene Ontology. This part of the Molecular Function (MF) ontology describes DNA binding. The key DNA binding terms that should be associated with dbTF are highlighted in blue, more specific GO terms are available to provide information about the DNA motifs bound by the dbTF.

Biological process

NOTE THAT THIS PART OF THE ONTOLOGY IS STILL BEING REVIEWED
There are currently two axes of classification: by product (Figure 3) and by the RNA polymerase generating the transcript (Figure 4). Ideally both terms should be used; the product generated by the transcription event is the most meaningful biologically. **We plan to remove the RNA polymerase-specific branch of the BP ontology.** Having a single branch will ensure that consistent and efficient annotation can be achieved.

For the 'regulation of transcription' branch ("GO:0006355 regulation of transcription, DNA-templated", and children; Figure 5), the 'product' axis does not currently exist. We will rename the polymerase-centric terms to the product-specific equivalent as appropriate. If there are two polymerases responsible for the same transcript, two terms will be instantiated (for

example, GO:1905380 regulation of snRNA transcription by RNA polymerase II; NEW term regulation of snRNA transcription by RNA polymerase III will both remain).

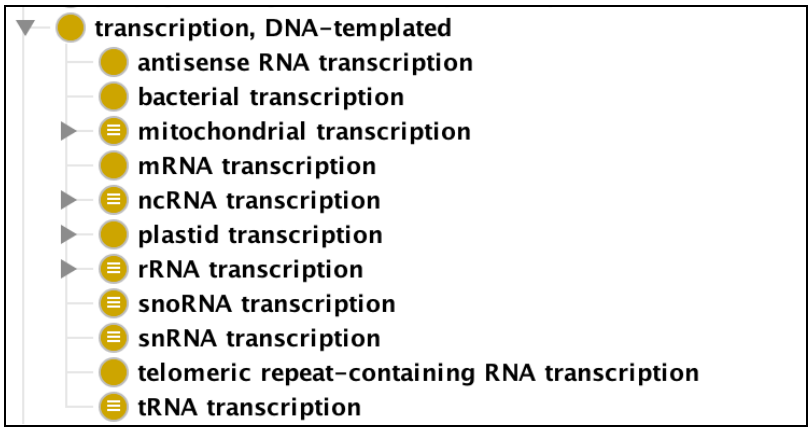

Figure 3. Transcription branch of the Gene Ontology. This part of the BP ontology provides terms to describe the role of GTFs in transcription of a specific type of transcript.

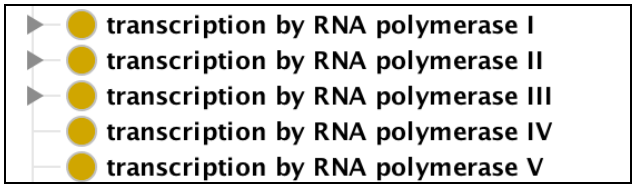

Figure 4. Transcription branch of the Gene Ontology. This part of the BP terms available to describe the role of GTFs in transcription mediated by a specific RNA polymerase.

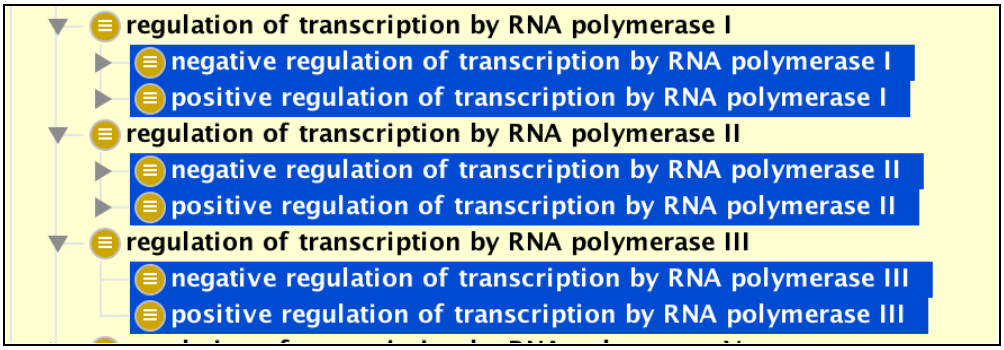

Figure 5. Regulation of transcription branch of the Gene Ontology. This part of the BP terms available to describe the role of dbTFs and coTFs in regulating transcription. Highlighted in blue are the key GO terms available to describe regulation of transcription in eukaryotic cells (prokaryotes use a single polymerase). The more specific child terms for the highlighted terms describe the role of transcription in other processes, eg "GO:1900464 negative regulation of cellular hyperosmotic salinity response by negative regulation of transcription from RNA

polymerase II promoter", these terms are being phased out. Instead, aside from its biochemical activity in the process of transcription, a transcription factor can be annotated to the biological process it receives input from, as a molecular signaling pathway or biomolecular sensor endpoint, and provides transcriptional output for through its target genes.

Cellular components: Complexes and cellular locations

NOTE THAT THIS PART OF THE ONTOLOGY IS STILL BEING REVIEWED

Annotation examples from research articles

Below we describe annotation work flows from a workshop held at EBI in April 2020.

This is a concept, perhaps other format is needed, a landscape page may be better suited

hypothesis-example/steps	**GTF**	**dbTF**	**coTF**
Example article	PMID:10924514	PMID:26314965	PMID:22783022
Step 1: Hypothesis	GTF2H2 is a subunit of the TFIIH GTF.	NKX6.3 is a transcription regulator.	SIN3A (Q96ST3) is a transcription co-repressor.
Step 2: Database mining for protein domains and orthologs	GTF2H2 contains a TFIIH subunit Ssl1/p44 domain (IPR012170), described in InterPro as a component of the transcription factor TFIIH core. UniProt describes GTF2H2 as a "Component of the TFIID-containing RNA polymerase II pre-initiation complex" (https://www.uniprot.org/uniprot/Q13888#function).	NKX6.3 contains a DNA binding homeobox domain (IPR020479). UniProt describes NKX6.3 as a "Putative transcription factor, which may be involved in patterning of central nervous system and pancreas." (https://www.uniprot.org/uniprot/A6NJ46#function).	SIN3A contains a SIN3A domain (IPR037969), among others. IPR037969 is associated with the GO Molecular Function GO:0003714 transcription corepressor activity (https://www.ebi.ac.uk/interpro/entry/InterPro/IPR037969/) UniProt describes SIN3A as "Acts as a transcriptional repressor. Corepressor for

			REST." (https://www.uniprot.org/uniprot/Q96ST3#function).
Step 3: GO annotation and literature mining	Existing annotations were consistent with the hypothesis that the gene encodes a GTF: there are IDA annotations to GO:0016251 RNA polymerase II general transcription initiation factor activity and GO:0008353 RNA polymerase II CTD heptapeptide repeat kinase activity, both functions of GTFs subunits.	Existing annotations were consistent with the hypothesis that the gene encodes a dbTF. - GO:0003700 (and children) DNA-binding transcription factor activity IDA, IBA, ISA, ISM - GO:0000978 RNA polymerase II cis-regulatory region sequence-specific DNA binding IBA - GO:0043565 sequence-specific DNA binding IDA, IBA	Existing annotations were conflicting. After review of the curated articles, several annotations were removed, because many of the activities were those of proteins SIN3A interacts with, not its function. Searching PubMed for SIN3A, and filtering for reviews, since there is a lot of literature describing SIN3A, result in many articles mentioning epigenetics, further supporting the coTF repressor hypothesis (https://pubmed.ncbi.nlm.nih.gov/?term=Sin3a&filter=pubt.review).
Step 4: New data extraction from publication	Figure 4, compares wild type GTF2H2 and a mutant in abortive initiation and promoter escape assays. The results show that the mutant GTF2H2 is unable to initiate transcription. This data supports the annotation GO:0016251 RNA polymerase II general	Figure 5 shows increased expression of target genes containing predicted binding motif sequences (TAAT), which was abolished by decreasing NKX6.3 expression with an antisense transcript. Moreover, ChIP data confirmed that	Figure 4 shows by ChIP assay that the SIN3A is found in the vicinity of the SOCS3 gene, a gene regulated by STAT3. The same figure also shows that STAT3 does not interact with the SOCS3 promoter in the presence of SIN3A, indicating a repressor effect of

	transcription initiation factor activity	NKX6.3 is bound to regulatory regions of the target genes (see notes above on the use of ChIP data to support an annotation). The dbTF activity annotation for NKX6.3 can therefore include the direct target genes information using the relation "has input". This supports an annotation of NKX6.3 to: - GO:0000978 RNA polymerase II cis-regulatory region sequence-specific DNA binding - GO:0001228 DNA-binding transcription activator activity, RNA polymerase II-specific - GO:0001227 DNA-binding transcription repressor activity, RNA polymerase II-specific - GO:0045944 positive regulation of transcription by RNA polymerase II GO:0000122 negative regulation	SIN3A on STAT3's transcription activator function. This supports an annotation of SIN3A to: - GO:0003714 transcription corepressor activity - GO:0000122 negative regulation of transcription by RNA polymerase II

		of transcription by RNA polymerase II	

Evidence for a GTF

PMID:10924514 Studies of the function of the GTF2H2 subunit of the TFIIH GTF

1. **The hypothesis** is that GTF2H2 is a subunit of the TFIIH GTF.
2. **Protein domains or characterized orthologs:** GTF2H2 contains a TFIIH subunit Ssl1/p44 domain (IPR012170), described in InterPro as a component of the transcription factor TFIIH core.
   UniProt describes GTF2H2 as a "Component of the TFIID-containing RNA polymerase II pre-initiation complex" (https://www.uniprot.org/uniprot/Q13888#function).
3. **Are there other GO annotations or published experimental results consistent with the hypothesis ?** Yes, existing annotations were consistent with the hypothesis that the gene encodes a GTF: there are IDA annotations to GO:0016251 RNA polymerase II general transcription initiation factor activity and GO:0008353 RNA polymerase II CTD heptapeptide repeat kinase activity, both functions of GTFs subunits.
4. **Are the experimental results consistent with the hypothesis ?** Yes, GTF GO:0016251 RNA polymerase II general transcription initiation factor activity from Figure 4, which compares wild type GTF2H2 and a mutant in abortive initiation and promoter escape assays. The results show that the mutant GTF2H2 is unable to initiate transcription.

Evidence for a dbTF

PMID:26314965 studies the activity of the NKX6.3 (UniProt:A6NJ46) transcription factor. Going through the checklist shows the following:

1. **The hypothesis** stated by the authors is that NKX6.3 is a transcription regulator.
2. **Protein domains or characterized orthologs:** NKX6.3 contains a DNA binding homeobox domain (IPR020479). UniProt describes NKX6.3 as a "Putative transcription factor, which may be involved in patterning of central nervous system and pancreas." (https://www.uniprot.org/uniprot/A6NJ46#function).
3. **Are there other GO annotations or published experimental results consistent with the hypothesis ?**
   All already existing annotations were consistent with the hypothesis that the gene encodes a dbTF.

GO ID	Term label	Evidence(s)	Consistent with hypothesis?

GO:0003700 (and children)	DNA-binding transcription factor activity	IDA, IBA, ISA, ISM	Yes
GO:0000978	RNA polymerase II cis-regulatory region sequence-specific DNA binding	IBA	Yes
GO:0043565	sequence-specific DNA binding	IDA, IBA	Yes

4. **Are the experimental results consistent with the hypothesis ?** Yes, Figure 5 shows increased expression of target genes containing predicted binding motif sequences (TAAT), which was abolished by decreasing NKX6.3 expression with an antisense transcript. Moreover, ChIP data confirmed that NKX6.3 is bound to regulatory regions of the target genes (see notes above on the use of ChIP data to support an annotation). The dbTF activity annotation for NKX6.3 can therefore include the direct target genes information using the relation "has input".
   This supports an annotation of NKX6.3 to:
   - GO:0000978 RNA polymerase II cis-regulatory region sequence-specific DNA binding
   - GO:0001228 DNA-binding transcription activator activity, RNA polymerase II-specific
   - GO:0001227 DNA-binding transcription repressor activity, RNA polymerase II-specific
   - GO:0045944 positive regulation of transcription by RNA polymerase II
   - GO:0000122 negative regulation of transcription by RNA polymerase II

Evidence for a coTF

PMID:22783022 studies the activity of the SIN3A transcription co-activator. This paper shows that the transcription factor STAT3 is a target of regulation by SIN3A. Going through the checklist shows the following:

1. **The starting hypothesis** stated by the authors is that SIN3A (Q96ST3) is a transcription co-repressor.
2. **Protein domains or characterized orthologs:**
   SIN3A contains a SIN3A domain (IPR037969), among others. IPR037969 is associated with the GO Molecular Function GO:0003714 transcription corepressor activity (https://www.ebi.ac.uk/interpro/entry/InterPro/IPR037969/)
   UniProt describes SIN3A as "Acts as a transcriptional repressor. Corepressor for REST." (https://www.uniprot.org/uniprot/Q96ST3#function).
3. **Are there other GO annotations or published experimental results consistent with the hypothesis ?**
   a. Existing annotations are conflicting. After review of the original articles, several annotations were removed, because many of the activities were those of proteins SIN3A interacts with, not its function.

b. Searching PubMed for SIN3A, and filtering for reviews, since there is a lot of literature describing SIN3A, result in many articles mentioning epigenetics, further supporting the coTF repressor hypothesis (https://pubmed.ncbi.nlm.nih.gov/?term=Sin3a&filter=pubt.review).

GO ID	Term label	Evidence(s)	Consistent with hypothesis? /Action if not
GO:0003714	transcription corepressor activity	IBA	Yes
GO:0003700	DNA-binding transcription factor activity	IEA (Ensembl)	No / Removed
GO:0000976	transcription regulatory region sequence-specific DNA binding	ISS	No / Removed
GO:0033558	protein deacetylase activity	IMP	No / Removed
GO:0004407	histone deacetylase activity	IBA	No / Removed
GO:0000976	transcription regulatory region sequence-specific DNA binding	ISS	No / Removed

4. **Are the experimental results consistent with the hypothesis ?**
   Yes, Figure 4 shows by ChIP assay that the SIN3A is found in the vicinity of the SOCS3 gene, a gene regulated by STAT3. The same figure also shows that STAT3 does not interact with the SOCS3 promoter in the presence of SIN3A, indicating a repressor effect of SIN3A on STAT3's transcription activator function. This supports an annotation of SIN3A to:
   - GO:0003714 transcription corepressor activity
   - GO:0000122 negative regulation of transcription by RNA polymerase II

Guidance on specific small scale experiments

Experiments providing evidence for DNA binding transcription factor activity can be found in Tables 3 and 4 of Tripathi 2013 (PMID:27270715).

Guidance on specific high-throughput experiments

The strategy outlined above for the annotation of transcription regulators should be applied when curating high-throughput data. The guidelines below are mostly relevant to dbTFs, if the experimental data can be used to capture information about coTFs this will be indicated.

High-throughput data can be used to capture the ‘direct’ target of a dbTF or a coTF. The direct target may be the genomic coordinates and/or the target gene.

HT-SELEX experiments (dbTF data)

High throughput selection experiments such as the HT-SELEX protocol (PMID:1697402, PMID: 2200121) yield data that provides strong experimental evidence of the DNA sequence recognised by a protein. This is then presented as an optimal DNA motif that represents the consensus of many binding observations. Crucially, sequence-specific DNA binding does not necessarily imply transcription regulation function. For instance, PRMT14 has a specific DNA binding activity, but it is involved in meiotic recombination site determination. For annotation to the dbTF Mf term, it is still necessary to have evidence that at least one target gene is regulated. However, often that evidence was already obtained in the course of other published studies. The HT-SELEX experiments, then, provide support for annotation to:

- GO:0000987 cis-regulatory region sequence-specific DNA binding or the descendent GO:0000978 RNA polymerase II cis-regulatory region sequence-specific DNA binding

However, if there is also evidence of regulation of transcription of a gene associated with the regulator region (either by transfection and measuring mRNA or by a reporter gene assay, etc), then HT-SELEX data provides support for the following annotations:

- GO:0003700 DNA-binding transcription factor activity or a child
- GO:0006357 regulation of transcription by RNA polymerase II (or GONEW: regulation of mRNA transcription)

for the TF IDA (HT-SELEX) is one of the strongest experimental evidence types to assign as dbTF. Still, further strengthens this conclusion. Eg: PRMT14 (is a meiosis factor not a TF). Cumulating evidence is important (Oriol Formes email of 18 February).

ChIP-seq experiments (dbTF or coTF data)

As ChIP-seq experiments include cross-linking of proteins to other proteins and/or DNA, on their own ChIP-seq experiments are inconclusive. However, if (a) the protein is known to bear a DNA binding domain and (b) there is evidence of regulation of transcription (either by transfection and measuring mRNA or by a reporter gene assay, etc), then the CHIP-seq data provides support for the following annotations:

- GO:0000987 cis-regulatory region sequence-specific DNA binding or the descendent GO:0000978 RNA polymerase II cis-regulatory region sequence-specific DNA binding
- GO:0003700 DNA-binding transcription factor activity or a child
- GO:0006357 regulation of transcription by RNA polymerase II (or GONEW: regulation of mRNA transcription)

If (c) the protein is known to be a coTF and (b) there is evidence of regulation of transcription (either by transfection and measuring mRNA or by a reporter gene assay, etc), then the CHIP-seq data provides support for the following annotations:

- GO:0003712 transcription coregulator activity or a child

- GO:0006357 regulation of transcription by RNA polymerase II (or GONEW: regulation of mRNA transcription)

Bacterial 1-hybrid experiments

The bacterial one-hybrid (B1H) system is a method for identifying the sequence-specific target site of a DNA-binding domain. In this system, a given transcription factor (TF) is expressed as a fusion to a subunit of RNA polymerase. In parallel, a library of randomized oligonucleotides representing potential TF target sequences are cloned into a separate vector containing the selectable genes HIS3 and URA3. If the DNA-binding domain (bait) binds a potential DNA target site (prey) in vivo, it will recruit RNA polymerase to the promoter and activate transcription of the reporter genes in that clone (https://en.wikipedia.org/wiki/Bacterial_one-hybrid_system; PMC1435991).