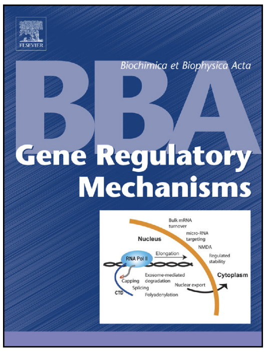
Journal Pre-proof



Gene Ontology representation for transcription factor functions

Pascale Gaudet, Colin Logie, Ruth C. Lovering, Martin Kuiper, Astrid Lægreid, Paul D. Thomas

PII: S1874-9399(21)00070-5

DOI: https://doi.org/10.1016/j.bbagrm.2021.194752

Reference: BBAGRM 194752

To appear in: *BBA - Gene Regulatory Mechanisms*

Received date: 6 January 2021

Revised date: 24 August 2021

Accepted date: 25 August 2021

Please cite this article as: P. Gaudet, C. Logie, R.C. Lovering, et al., Gene Ontology representation for transcription factor functions, *BBA - Gene Regulatory Mechanisms* (2021), https://doi.org/10.1016/j.bbagrm.2021.194752

Gene Ontology representation for transcription factor functions

Pascale Gaudet[1], Colin Logie[2], Ruth C. Lovering[3], Martin Kuiper[4], Astrid Lægreid[5], Paul D. Thomas[6]

Author affiliations:

1 Swiss-Prot group, SIB Swiss Institute of Bioinformatics, 1 Rue Michel-Servet, 1211 Genève, Switzerland

2 Molecular Biology Department, Faculty of Science, Radboud University, PO box 9101, 6500HB Nijmegen, The Netherlands

3 Functional Gene Annotation, Preclinical and Fundamental Science, UCL Institute of Cardiovascular Science, University College London, London, UK

4 Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

5 Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

6 Division of Bioinformatics, Department of Preventive Medicine, University of Southern California, Los Angeles CA, USA

Abstract

Transcription plays a central role in defining the identity and functionalities of cells, as well as in their responses to changes in the cellular environment. The Gene Ontology (GO) provides a rigorously defined set of concepts that describe the functions of gene products. A GO annotation is a statement about the function of a particular gene product, represented as an association between a gene product and the biological concept a GO term defines. Critically, each GO annotation is based on traceable scientific evidence. Here, we describe the different GO terms that are associated with proteins involved in transcription and its regulation, focusing on the standard of evidence required to support these associations. This article is intended to help users of GO annotations understand how to interpret the annotations and can contribute to the consistency of GO annotations. We distinguish between three classes of activities involved in transcription or directly regulating it - general transcription factors, DNA-binding transcription factors, and transcription co-regulators.

Introduction

The Gene Ontology (GO) develops a computational model of biological systems, ranging from the molecular to the organism level, across all species in the tree of life. GO aims to provide a comprehensive representation of the current scientific knowledge about the functions of gene products, namely, proteins and non-coding RNA molecules (1)(2). GO is organized in three aspects. GO Molecular Functions (**MF**) describe activities that occur at the molecular level, such as “DNA binding transcription factor activity” or “histone deacetylase activity”. Biological Processes (**BP**) represent the larger processes or ‘biological programs’ accomplished by multiple molecular activities. Examples of broad biological process terms are "transcription" or "signal transduction". Cellular Components (**CC**) are the cellular structures in which a gene product performs a function, either cellular compartments (e.g., "nucleus" or "chromatin"), or stable macromolecular complexes of which they are parts (e.g., "RNA polymerase II"). Together, annotations of a gene to terms from each of those aspects describe what specific function a gene product plays in a process and where this activity occurs in the cell. Ideally every gene product should have an annotation from each of the three aspects of GO.

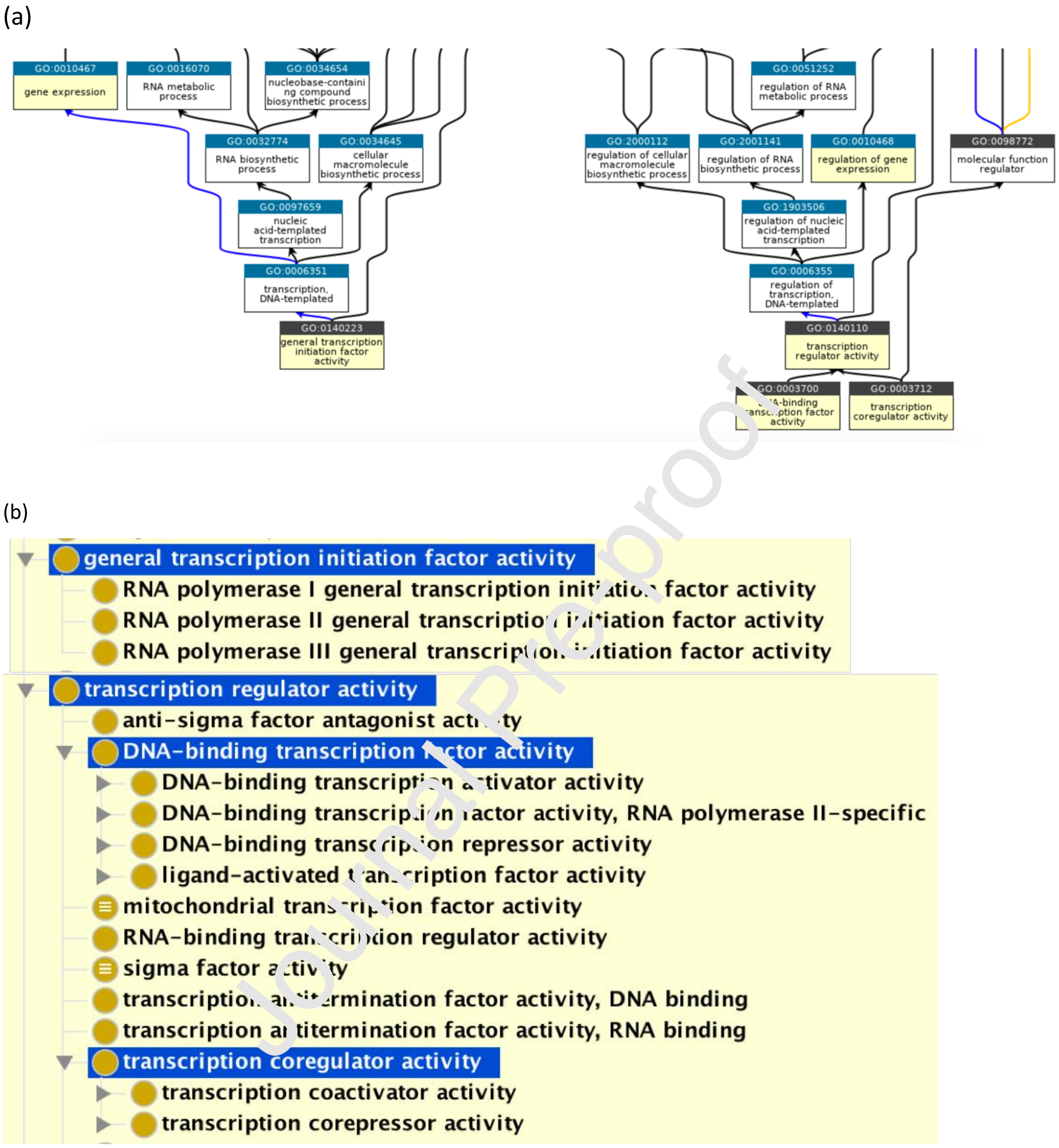
The specific genes expressed in a given cell define the identity and functionalities of that cell. Regulation of transcription is highly complex and leads to differential gene expression in specific cells or under specific conditions. In human cells, it has been estimated that several thousand proteins participate in gene expression and its regulation, directly or indirectly (3)(Velthuijs, *in preparation*). This includes the general transcription machinery, the factors that make the chromatin more or less accessible, specific DNA-binding transcription factors, and the signaling molecules that regulate the activity of all those proteins. This complexity is difficult to accurately represent in ontological form. Tripathi *et al.* (4) redesigned that part of the ontology in 2013 to define precise molecular functions for the various proteins involved in transcription and its regulation. Nearly 10 years after its implementation, we had to acknowledge that this framework was too complex and difficult to navigate, leading to inconsistent annotations and thus poorly serving the user community. The work described here was also motivated by the GREEKC consortium, whose goals include curation tools development, reengineering of ontologies, development of curation guidelines and text mining tools, developing platforms to analyze and render the molecular logic of transcription regulatory networks for which a robust infrastructure is needed. Therefore, we thoroughly reviewed the Gene Ontology representation of molecular activities relevant to transcription, with a simpler and more pragmatic approach, more aligned with available experimental data.

We have revised the GO MF terms representing the activities of proteins involved in transcription, with the input from domain experts. In addition to RNA polymerase, we defined three different types of activities that take place on the DNA to mediate or regulate transcription: general transcription factors (GTFs), DNA-binding transcription factors (dbTFs), and transcription coregulators (coTFs).

Here we present the annotation approach recommended by the GO consortium (5), applied to the recent refactoring of the transcription domain of GO. This approach aims to 1) help biocurators – annotation producers - interpret published data and correctly assign the MFs terms for GTF, dbTF, or coTF to a protein, and 2) help users understand how the data is generated and how to interpret them. The annotation of factors involved in transcription and its regulation is challenging for multiple reasons. Contrary to other molecular functions, for example enzymes, where one protein or a well-defined complex catalyses a precise reaction, the measurable output of transcription activities is the result of multiple nearly simultaneous activities of GTF, dbTF, coTF, as well as RNA polymerase, hence, individual activities can be hard to distinguish experimentally. Moreover, these factors often form large complexes, such that the level of resolution of the experimental setup is essential to determine the precise activity of any given protein. Older experimental methods often did not provide enough details, leading to inaccurate classifications of certain proteins. In addition, researchers use "transcription factor" loosely, at times meaning GTF, dbTF, or coTF. This complicates the annotation process and necessitates solid expertise for correct interpretation of the data. The experimental data itself is difficult to parse for unambiguous assignment of a function to a protein: typically, a single experiment is insufficient for accurately determining the function of these proteins, thus, interpretation of experimental results that investigate dbTFs must rely on pre-existing knowledge. Also, 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 known specific DNA-binding domains and in some cases, evidence of an effect on the transcription of putative direct target genes. To add to the complexity, the presence of a DNA-binding domain in a protein does not always imply that the protein functions as a dbTF (6).

GO description of molecular functions relevant for transcription

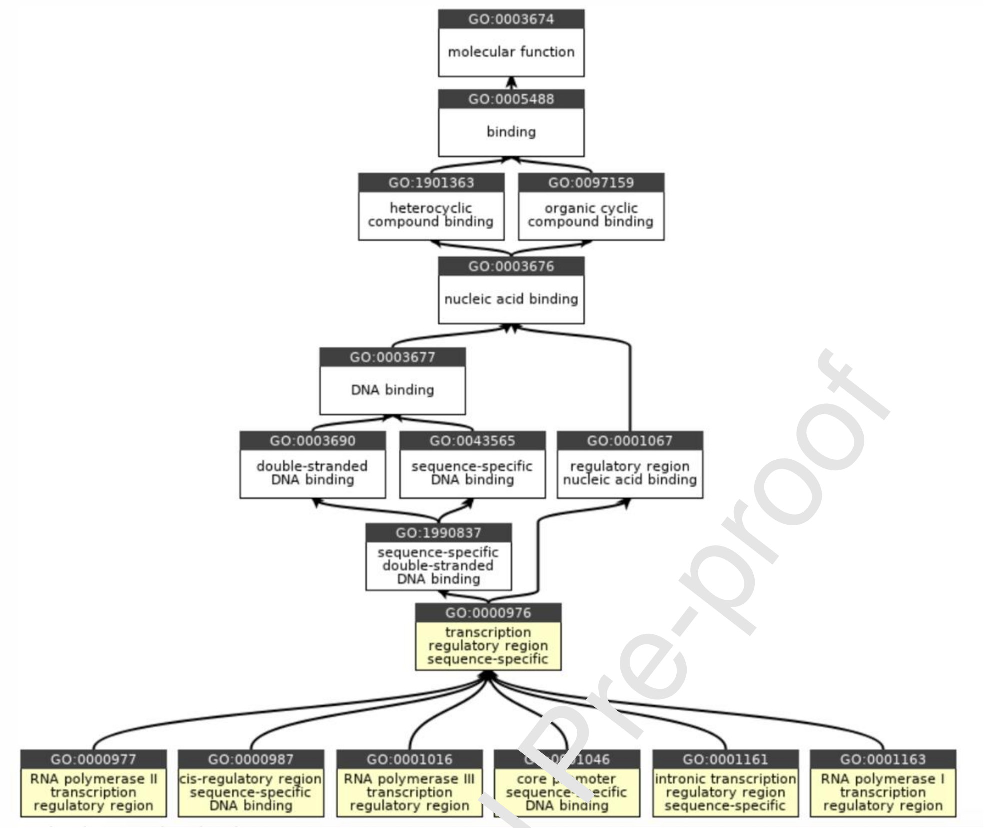
We distinguish between three types of activities involved in transcription or directly regulating it: general transcription factors (GO:0140223), DNA-binding transcription factors (GO:0003700), and transcription co-regulators (GO:0003712). The **general transcription initiation factor activity** term and its descendants describe the activities of general transcription initiation factors for RNA polymerase I, II and III, which play a direct role in the biological process of transcription at the core promoter (Sant *et al. in preparation*). In contrast, the **GO:0140110 transcription regulator activity** branch describes the activities of transcription regulators: dbTF and coTFs, that act at any type of cis-regulatory module (Figure 1). DNA-binding transcription factors are adaptors that bind chromatin at specific genomic addresses to coordinately regulate the expression of genes sets. This is encoded in the ontology via links between the DNA-binding transcription factor activity term and its descendants and to their counterpart branch of the MF ontology describing DNA binding. The GO:0000976 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). An overview of the GO structure for DNA binding activities is shown in Figure 2. The definitions and placement of GO terms in the ontology can be viewed in the AmiGO (7)(8); http://amigo.geneontology.org/amigo, and QuickGO (9); https://www.ebi.ac.uk/QuickGO/ browsers.


(a)
GO:0010467
gene expression
GO:0016070
RNA metabolic process
GO:0034654
nucleobase-containing compound biosynthetic process
GO:0032774
RNA biosynthetic process
GO:0034645
cellular macromolecule biosynthetic process
GO:0097659
nucleic acid-templated transcription
GO:0006351
transcription, DNA-templated
GO:0140223
general transcription initiation factor activity
GO:0051252
regulation of RNA metabolic process
GO:2000112
regulation of cellular macromolecule biosynthetic process
GO:2001141
regulation of RNA biosynthetic process
GO:0010468
regulation of gene expression
GO:0098772
molecular function regulator
GO:1903506
regulation of nucleic acid-templated transcription
GO:0006355
regulation of transcription, DNA-templated
GO:0140110
transcription regulator activity
GO:0003700
DNA-binding transcription factor activity
GO:0003712
transcription coregulator activity
(b)

- general transcription initiation factor activity
  - RNA polymerase I general transcription initiation factor activity
  - RNA polymerase II general transcription initiation factor activity
  - RNA polymerase III general transcription initiation factor activity
- transcription regulator activity
  - anti-sigma factor antagonist activity
  - DNA-binding transcription factor activity
    - DNA-binding transcription activator activity
    - DNA-binding transcription factor activity, RNA polymerase II-specific
    - DNA-binding transcription repressor activity
    - ligand-activated transcription factor activity
  - mitochondrial transcription factor activity
  - RNA-binding transcription regulator activity
  - sigma factor activity
  - transcription antitermination factor activity, DNA binding
  - transcription antitermination factor activity, RNA binding
  - transcription coregulator activity
    - transcription coactivator activity
    - transcription corepressor activity

Figure 1. Transcription regulator activity branches of the Gene Ontology. (a) Graphical representation of the placement of the parent terms for transcription regulator molecular functions. Black headers correspond to MF and cyan headers to BP terms. (b) Transcription regulators are dbTF and coTFs. The general transcription initiation factors play a direct role in transcription. Top-level terms of each branch are highlighted in blue.

(a)


GO:0003674
molecular function
GO:0005488
binding
GO:1901363
heterocyclic compound binding
GO:0097159
organic cyclic compound binding
GO:0003676
nucleic acid binding
GO:0003677
DNA binding
GO:0003690
double-stranded DNA binding
GO:0043565
sequence-specific DNA binding
GO:0001067
regulatory region nucleic acid binding
GO:1990837
sequence-specific double-stranded DNA binding
GO:0000976
transcription regulatory region sequence-specific
GO:0000977
RNA polymerase II transcription regulatory region
GO:0000987
cis-regulatory region sequence-specific DNA binding
GO:0001016
RNA polymerase III transcription regulatory region
GO:0001046
core promoter sequence-specific DNA binding
GO:0001161
intronic transcription regulatory region sequence-specific
GO:0001163
RNA polymerase I transcription regulatory region

(b)

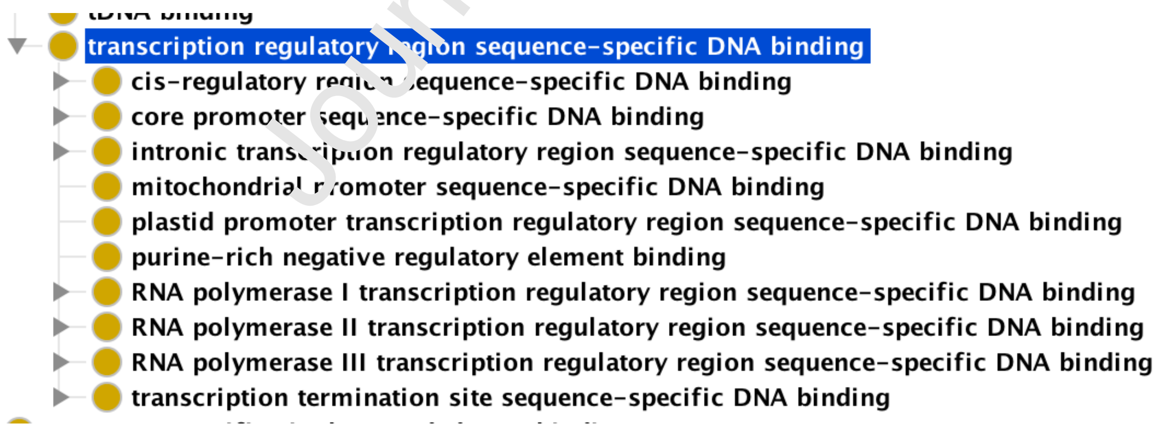

transcription regulatory region sequence-specific DNA binding
cis-regulatory region sequence-specific DNA binding
core promoter sequence-specific DNA binding
intronic transcription regulatory region sequence-specific DNA binding
mitochondrial promoter sequence-specific DNA binding
plastid promoter transcription regulatory region sequence-specific DNA binding
purine-rich negative regulatory element binding
RNA polymerase I transcription regulatory region sequence-specific DNA binding
RNA polymerase II transcription regulatory region sequence-specific DNA binding
RNA polymerase III transcription regulatory region sequence-specific DNA binding
transcription termination site sequence-specific DNA binding

Figure 2. DNA binding branch of the Gene Ontology. This part of the Molecular Function (MF) ontology describes DNA binding. (a) Graphical representation of the placement of the terms describing sequence-specific promoter binding. (b) Hierarchical view of the sequence-specific transcription regulatory region binding terms.

Strategy for annotating transcription-associated activities

GO terms are associated with gene products based on two general approaches: from experimental data and from sequence inferences (10). The GO database has a total of 8 million annotations, about 7% of which are to human gene products. For human, there are > 915,000 annotations derived from experimental data (GO release 2020-10-10 obtained from http://amigo.geneontology.org). Sequence inference methods provide more than 106,000 annotations for human proteins based on phylogenetic relationships (65,000 annotations) (11); protein domains (6,730 annotations) (12); and Ensembl orthology predictions (35,000 annotations) (13). The next sections describe the annotation of the different types of proteins involved in transcription and its regulation.

Transcription activity annotations supported by experimental data

The following annotation approach follows the recommendations of the GO consortium. First and foremost, it is necessary to use as much information as possible, rather than annotating articles individually and out of the wider context. When extracting information, a gene-by-gene or pathway-by-pathway approach is considered best practice (5). Reviewing a range of articles ensures that the annotations closely reflect the current state of knowledge. Ideally, the corpus of annotations for a gene product should be based on multiple observations from different articles by independent research groups. Five steps used to determine whether a gene can be annotated as a transcriptional regulator are outlined in Figure 3. Appendix 1 provides examples of each of those different activities.

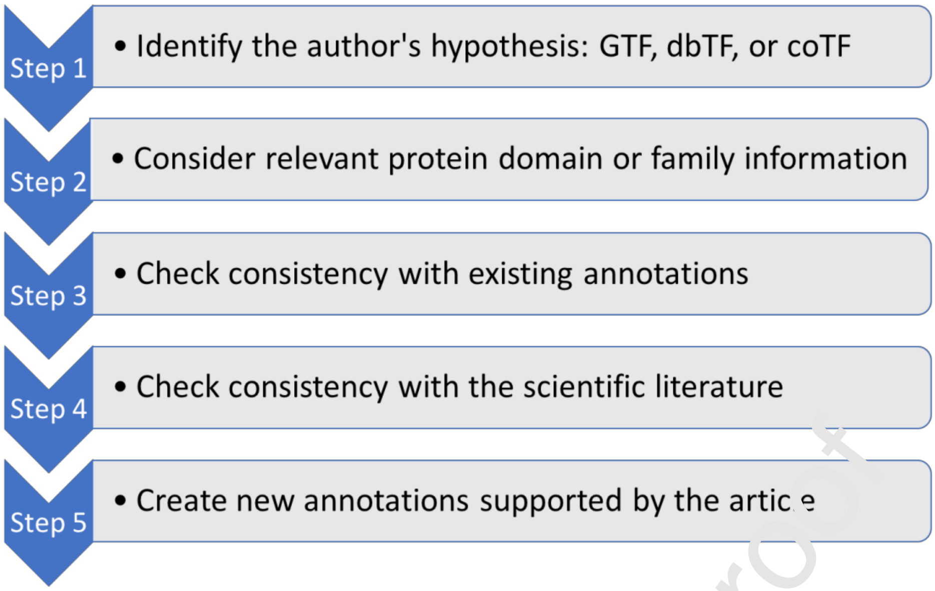

Step 1
• Identify the author's hypothesis: GTF, dbTF, or coTF
Step 2
• Consider relevant protein domain or family information
Step 3
• Check consistency with existing annotations
Step 4
• Check consistency with the scientific literature
Step 5
• Create new annotations supported by the article

Figure 3. Five steps to transcription activity annotation. The five key steps to associating a transcription MF term with a protein starts with identifying the starting hypothesis, to confirm that the authors are characterizing a GTF, dbTF or coTF. Secondly, considering whether the knowledge from specific protein domains or characterized orthologs support the hypothesis. Thirdly, checking whether existing annotations from GO, UniProt and Model Organism databases are consistent with the hypothesis. Fourthly, reviewing other published experimental data to ensure no contradictory findings have been reported. Finally, creating new GO annotations, if the experimental results are consistent with the identified hypothesis.

1. **Identify 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 which aspect of the existing model or what new model the authors are assessing. The hypothesis tested by the authors is essential to choose a GO term, with the caveat that inconsistent terminology has been used in transcription research articles and therefore may not always be aligned with the GO term categories.

2. **Determine whether 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 this data should be used with caution. For instance, ARID-, AT hook-, and some HMG-, GATA-, zinc finger domain-containing proteins and proteins binding structural features such as the DNA minor groove rarely bind DNA in a sequence-specific manner; some of them merely function to increase the avidity or stability of a transcription factor complex and its associated co-factors and do not - in their own capacity - provide the specific genomic address to guide transcription to specified target genes. Such proteins are not considered dbTFs in GO.

To support the association of a gene with a GO term from homologous sequences from other species, only closely related orthologs whose function have been unambiguously characterized can be used *if those are consistent with the experimental data presented in the article*.

- **GTFs** function as the molecular machine that assembles with the RNA polymerase at the promoter to form the pre-initiation complex (PIC). GTFs have been characterized in several organisms, from archaea to yeast and mammalian cells (14)(15), 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.

- **dbTFs** are specific double-stranded DNA-binding transcription factors that provide genomic addresses and respond to the conditions under which specific genes are expressed. Central to dbTF function is their binding to specific double-stranded DNA sequences that are often named transcription factor binding sites (TFBS). Gene products associated with the GO term "GO:0003700 DNA-binding transcription factor activity" have the ability to bind DNA *and* this binding regulates the expression of a specific set of target genes. The direct target gene(s) can also be included in the annotation using the "*has input* relation". A human dbTF catalog developed by the GREEKC project ((6); also accessible from https://www.ebi.ac.uk/QuickGO/targetset/dbTF) may be consulted to check whether a specific human protein is annotated to dbTF function with experimental or phylogenetic evidence. When considering proteins that belong to families of well characterized transcription factors, such as those that contain bHLH, bZIP, homeobox, ETS, Forkhead, etc. domains and proteins with a one-to-one ortholog already demonstrated to be a dbTF, then weaker evidence of DNA binding, such as ChIP experiments is sufficient. In contrast, special care must be taken to annotate proteins bearing domains that are not exclusively found in transcription factors, such as RING, MYND and PhD zinc fingers. Similarly, for proteins with enzymatic activity: while there are rare cases of dbTFs with enzymatic activities, such as ENO1, dbTF and enzymatic activity are usually mutually exclusive. For proteins not in the dbTF catalog, clear experimental or phylogenetic evidence of sequence-specific DNA binding and gene transcription regulation via cognate DNA motifs located in gene-associated *cis*-regulatory modules is required for the protein to be classified with high confidence as a dbTF.

- **coTFs**: Transcription coregulators (also known as transcription cofactors; GO:0003712) represent a group of different functions that take place at *cis*-regulatory regions to make transcription of specific gene sets either more (coactivators) or less (corepressors) efficient. Coregulators can modify chromatin structure through covalent modification of histones, ATP-dependent chromatin remodelling, and modulate dbTF interactions with other transcription coregulators. We classify the Mediator Complex, which bridges dbTFs and the RNA polymerase, as a transcription coactivator (16)(17)(18). Many coTFs have enzymatic activity and normally exert their function independent of high affinity binding to specific DNA sequences. CoTFs that do bind DNA typically recognize very short DNA sequences that are not sufficiently unique in the genome to enable regulation of a limited set of genes in a discrete environmental or developmental stage. One example of this is CPF1, that binds the CpG dinucleotide and helps most CpG islands gain epigenomic marking (19)(20)(21).

It is important to keep in mind that DNA binding proteins that regulate transcription are not necessarily dbTFs. Key points that help distinguish between the three activities discussed above are that (i) dbTFs bind DNA in a sequence-specific manner, and regulate precise sets of genes; (ii) coTFs usually do not directly bind DNA, and when they do they don't exhibit strong sequence-specificity (iii) coTFs often have catalytic activities (such as histone methyltransferase, protein kinase, or ubiquitin ligase), which is highly unusual in dbTFs; (iv) GTFs are required for core promoter activity and are considered to act at each promoter to promote transcription initiation (14)(22), although the exact subunit composition at individual promoters may vary.

3. **Confirm that existing annotations are consistent with the hypothesis.** New annotations need to be consistent with existing annotations, unless the existing annotations are believed to be wrong or out of date. Annotations made to a term as well as a more specific descendant reflect differences in granularity of annotation, and are not generally considered inconsistent. When the new annotation uses a term in a different branch than existing annotations, a review of the evidence supporting the existing annotations is undertaken and, if necessary, annotations that appear to be incorrect are disputed (see section "Ensuring a coherent set of annotations").

4. **Check that other published experimental results do not contradict the hypothesis.** The application of the gene-by-gene or pathway-by-pathway annotation approach ensures that results from other research articles are taken into account and that all annotations are in line with the current state of knowledge. Again, if inconsistencies are

noticed, great care is taken to confirm correct interpretation of the data, this is particularly important if there is evidence for multiple, distinct transcription activity functions.

5. **Validate that the experimental results are consistent with the hypothesis.** If the results presented in the curated article are consistent with the hypothesis presented by the authors, then the appropriate transcription activity GO term(s) are associated with the gene product.

Proteins that are involved in transcription and its regulation have historically been studied through small-scale, focused experimental approaches. For some examples of the small-scale experiments that do provide evidence for DNA binding transcription factor activity the biocurator can use Tables 3 and 4 of Tripathi *et al.*, 2013, (4) and in Santos-Zavaleta et al., 2019, (23). Recent advances in high-throughput methodologies now provide robust data that, when interpreted with sufficient care, support the assignment of a function role to many proteins, including transcription regulators. This includes HT-SELEX (24)(25), Protein Binding Microarrays (26), ChIP (27), one- and two-hybrid experiments (28)(29). For these experiments, the data quality and the false positive rate must be evaluated before annotations are created. For example, human HT-SELEX data will have more false positives if native dbTFs are assayed in nuclear extracts or over-expressed in eukaryotic cells, compared with heterologous proteins purified from prokaryotic cells, as the latter reduces the probability of indirect interactions with endogenous factors. For high-throughput transcription data, only articles with low rates of false positives, are curated. Those various techniques provide multiple independent lines of evidence, strengthening the confidence in the annotation when they converge on a single motif or molecular function. The GO recommendations on curation of high-throughput experimental data should be applied when such data is annotated (30).

Annotations based on non-experimental evidence

There are only about 500 human dbTFs for which there is experimental evidence satisfying the criteria presented here. Across all areas of biology several reliable methods infer protein function from available experimental data. Indeed, there are approximately 1,000 human proteins annotated as dbTFs by non-experimental methods (Lovering et al. same BBA issue, prepublication available at (6)). Phylogenetic annotations are assigned by a group of biocurators with expertise in evolutionary biology, and require experimental evidence for at least one member of a clade of evolutionarily related proteins (11). The GO knowledgebase also contains GO terms assigned by automated pipelines based on protein domain (InterPro2GO) and orthology (Ensembl). InterPro2GO (12) is based primarily on local (partial) homology: protein domains are mapped to specific GO terms, and any protein with one of these domains will be annotated to the appropriate GO term(s). Ensembl Compara (13) generates groups of one-to-one orthologs among closely related species and propagates all experimental annotations to each members of the group. While manual annotations based on these methods are allowed, the GO consortium recommends using the automated pipelines that are maintained centrally and ensure a consistent annotation corpus across all annotated species.

Ensuring a coherent set of annotations

During the process of annotation, other relevant annotations associated with the gene are reviewed. If there are conflicting annotations, the supporting data should be reassessed to determine whether the annotations are inconsistent with the data, in which case the annotations must be fixed (5).

In cases where the primary data is conflicting across different articles (for example a protein is sometimes described as a transcription factor, and sometimes as a coregulator), then the literature will be reviewed carefully to decide whether the annotation is incorrect (bad choice of term, wrong protein annotated), whether the knowledge has evolved, if the protein plays multiple roles under different conditions (i.e., acts as a DNA-binding transcription factor in certain contexts and as a cofactor in others). If no activity has yet been established, no MF annotation will be made.

Note that individual DNA-binding transcription factors can act as both activators or repressors dependent on the context, hence association of both activator and repressor terms with a single protein is not considered inconsistent. The specific conditions under which this happens, such as relevant signaling pathways, cell type, as well as specific target genes, *etc*., may be further specified through additional context details ((31); see an example in Figure 4).

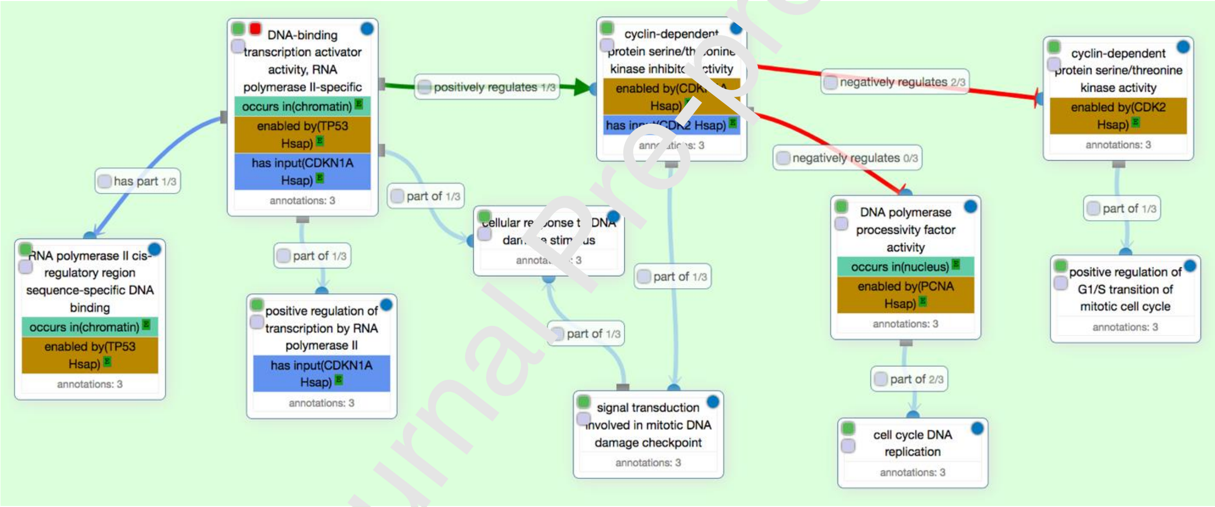

DNA-binding transcription activator activity, RNA polymerase II-specific
occurs in(chromatin)
enabled by(TP53 Hsap)
has input(CDKN1A Hsap)
annotations: 3
positively regulates 1/3
cyclin-dependent protein serine/threonine kinase inhibitor activity
enabled by(CDKN1A Hsap)
has input(CDK2 Hsap)
annotations: 3
negatively regulates 2/3
cyclin-dependent protein serine/threonine kinase activity
enabled by(CDK2 Hsap)
annotations: 3
negatively regulates 0/3
has part 1/3
part of 1/3
cellular response to DNA damage stimulus
annotations: 3
part of 1/3
DNA polymerase processivity factor activity
occurs in(nucleus)
enabled by(PCNA Hsap)
annotations: 3
part of 1/3
RNA polymerase II cis-regulatory region sequence-specific DNA binding
occurs in(chromatin)
enabled by(TP53 Hsap)
annotations: 3
part of 1/3
positive regulation of transcription by RNA polymerase II
has input(CDKN1A Hsap)
annotations: 3
part of 1/3
part of 1/3
positive regulation of G1/S transition of mitotic cell cycle
annotations: 3
signal transduction involved in mitotic DNA damage checkpoint
annotations: 3
part of 2/3
cell cycle DNA replication
annotations: 3

Figure 4. Representation of biological context of dbTF activity. The level of cyclin-dependent kinase inhibitor p21 (CDKN1A) is regulated by the transcription factor p53 (TP53) upon DNA damage, signaling cell cycle arrest to the cell (http://noctua.berkeleybop.org/editor/graph/gomodel:5fa76ad400000000).

Pitfalls in annotating transcription regulators

During the review of dbTF GO annotations (6), in which over 3,000 GO annotations were reviewed, a variety of common errors in data interpretation were identified. One of the most common errors was caused by the difficulty in distinguishing a dbTF from a coTF, as the evidence for those two functions can be quite similar. To prevent this error, biocurators ensure that the protein has a sequence-specific double-stranded DNA-binding domain and conduct an exhaustive review of the literature, including articles associated with the protein's close orthologs. Furthermore, the literature supporting the dbTF activity of a protein that also has evidence for another function, in particular, RNA binding, will be carefully checked before assigning a dbTF activity. The work on the human dbTF catalogue added a GO 'DNA-binding transcription factor activity' annotation to 583 proteins, and removed erronous assignments for 256 proteins.

Transcription regulators most often act as members of complexes, some of which also contain proteins with other activities. In some cases, only some subunits of a complex interact with DNA: for instance, while the RFX complex contains three members: RFX5, RFXAP and RFXANK, only RFX5 binds DNA directly. But the DNA-binding ability of the complex is facilitated by all three subunits so RFXAP and RFXANK are not coTFs (32). In this case, RFXAP and RFXANK are annotated using the "*contributes to*" qualifier, to indicate that they participate in, but are not directly responsible for the activity.

Another activity that can easily be confused for a coTF is a dbTF inhibitor. These proteins interact with a dbTF, but not at the DNA, to prevent the dbTF from reaching its target genes. Well characterized examples are the I-SMADs, SMAD6 and SMAD7 (33), that act by competing with active SMADs at receptors, thus blocking further intracellular signalling, and should be annotated to "GO:0140416 transcription regulator inhibitor activity".

It must be noted that these approaches to avoid errors in dbTF activity assignment are not unequivocal, as some proteins do have multiple functions. For example, the glucocorticoid receptor (NR3C1), which is a canonical dbTF, has recently been shown to bind double-stranded RNA motifs (34); ATF2 (activating transcription factor 2) and CLOCK are dbTFs that have been reported to also exhibit histone acetyltransferase activity (35)(36)(37)(38); some dbTFs, such as NFIB (nuclear factor I B), also function as dbTF inhibitors (39). Finally, general and sequence-specific effects can be difficult to separate, as has been established for the MYC dbTF (40).

Conclusion

The annotation approach presented here is designed to help biocurators annotate factors involved in transcription and its regulation, as well as for users of GO annotations to understand their meaning and the evidence behind them. This work complements the redesign of this part of the GO to significantly simplify the ontology structure. The new ontology structure and the present standards were applied to the review of human proteins associated with GO terms describing dbTF activity (6). We anticipate that adoption of this annotation approach by all groups who produce GO associations will increase annotation consistency across all species, for transcription and also more widely across all areas represented by GO.

Acknowledgements

We thank many GREEKC and GO consortium members for useful discussions that led to the development of these guidelines, in particular Marcio L. Acencio, Helen Attrill, and Valerie Wood.

Funding sources

The GO Consortium is funded by the National Human Genome Research Institute (US National Institutes of Health), grant number HG002273. RCL has been supported by Alzheimer's Research UK grant (ARUK-NAS2017A-1) and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. GREEKC is supported by the COST Action grant CA15205.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–9.

2. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 08;47(D1):D330–8.

3. Tupler R, Perini G, Green MR. Expressing the human genome. Nature. 2001 Feb 15;409(6822):832–3.

4. Tripathi S, Christie KR, Balakrishnan R, Huntley R, Hill DP, Thommesen L, et al. Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. Database J Biol Databases Curation. 2013;2013:bat062.

5. Poux S, Gaudet P. Best Practices in Manual Annotation with the Gene Ontology. Methods Mol Biol Clifton NJ. 2017;1446:41–54.

6. Lovering RC, Gaudet P, Acencio ML, Ignatchenko A, Jolma A, Fornes O, et al. A GO catalogue of human DNA-binding transcription factors. bioRxiv. 2020 Jan 1;2020.10.28.359232.

7. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015 Jan;43(Database issue):D1049-1056.

8. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinforma Oxf Engl. 2009 Jan 15;25(2):288–9.

9. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. Bioinforma Oxf Engl. 2009 Nov 15;25(22):3045–6.

10. Gaudet P, Škunca N, Hu JC, Dessimoz C. Primer on the Gene Ontology. Methods Mol Biol Clifton NJ. 2017;1446:25–37.

11. Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinform. 2011 Sep;12(5):449–62.

12. Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 2015 Jan;43(Database issue):D213-221.

13. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res. 2019 08;47(D1):D745–51.

14. Sainsbury S, Bernecky C, Cramer P. Structural basis of transcription initiation by RNA polymerase II. Nat Rev Mol Cell Biol. 2015 Mar;16(3):129–43.

15. Koster MJE, Snel B, Timmers HTM. Genesis of chromatin and transcription dynamics in the origin of species. Cell. 2015 May 7;161(4):724–36.

16. André KM, Sipos EH, Soutourina J. Mediator Roles Going Beyond Transcription. Trends Genet TIG. 2020 Sep 10;

17. Eychenne T, Werner M, Soutourina J. Toward understanding of the mechanisms of Mediator function in vivo: Focus on the preinitiation complex assembly. Transcription. 2017;8(5):328–42.

18. Yin J, Wang G. The Mediator complex: a master coordinator of transcription and cell lineage development. Dev Camb Engl. 2014 Mar;141(5):977–87.

19. Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature. 2010 Apr 15;464(7291):1082–6.

20. Lipski J, Zhang X, Kruszewska B, Kanjhan R. Morphological study of long axonal projections of ventral medullary inspiratory neurons in the rat. Brain Res. 1994 Mar 21;640(1–2):171–84.

21. Long HK, Blackledge NP, Klose RJ. ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. Biochem Soc Trans. 2013 Jun;41(3):727–40.

22. Cramer P. Organization and regulation of gene transcription. Nature. 2019;573(7772):45–54.

23. Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, Ledezma-Tejeida D, et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. Nucleic Acids Res. 2019 08;47(D1):D212–20.

24. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. Nature. 1990 Aug 30;346(6287):818–22.

25. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 1990 Aug 3;249(4968):505–10.

26. Andrilenas KK, Penvose A, Siggers T. Using protein-binding microarrays to study transcription factor specificity: homologs, isoforms and complexes. Brief Funct Genomics. 2015 Jan;14(1):17–29.

27. Kim TH, Dekker J. ChIP-seq. Cold Spring Harb Protoc. 2018 01;2018(5).

28. Sewell JA, Fuxman Bass JI. Options and Considerations When Using a Yeast One-Hybrid System. Methods Mol Biol Clifton NJ. 2018;1794:119–30.

29. Paiano A, Margiotta A, De Luca M, Bucci C. Yeast Two-Hybrid Assay to Identify Interacting Proteins. Curr Protoc Protein Sci. 2019;95(1):e70.

30. Attrill H, Gaudet P, Huntley RP, Lovering RC, Engel SR, Poux S, et al. Annotation of gene product function from high-throughput studies using the Gene Ontology. Database J Biol Databases Curation. 2019 01;2019.

31. Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Van Auken K, Carbon S, et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet. 2019;51(10):1429–33.

32. Masternak K, Barras E, Zufferey M, Conrad B, Corthals G, Aebersold R, et al. A gene encoding a novel RFX-associated transactivator is mutated in the majority of MHC class II deficiency patients. Nat Genet. 1998 Nov;20(3):273–7.

33. Miyazawa K, Miyazono K. Regulation of TGF-β Family Signaling by Inhibitory Smads. Cold Spring Harb Perspect Biol. 2017 Mar 1;9(3).

34. Parsonnet NV, Lammer NC, Holmes ZE, Batey RT, Wuttke DS. The glucocorticoid receptor DNA-binding domain recognizes RNA hairpin structures with high affinity. Nucleic Acids Res. 2019 05;47(15):8180–92.

35. Kawasaki H, Schiltz L, Chiu R, Itakura K, Taira K, Nakatani Y, et al. ATF-2 has intrinsic histone acetyltransferase activity which is modulated by phosphorylation. Nature. 2000 May 11;405(6783):195–200.

36. Hirayama J, Sahar S, Grimaldi B, Tamaru T, Takamatsu K, Nakahata Y, et al. CLOCK-mediated acetylation of BMAL1 controls circadian function. Nature. 2007 Dec 13;450(7172):1086–90.

37. Grimaldi B, Nakahata Y, Sahar S, Kaluzova M, Gauthier D, Pham K, et al. Chromatin remodeling and circadian control: master regulator CLOCK is an enzyme. Cold Spring Harb Symp Quant Biol. 2007;72:105–12.

38. Wang Z, Wu Y, Li L, Su X-D. Intermolecular recognition revealed by the complex structure of human CLOCK-BMAL1 basic helix-loop-helix domains with E-box DNA. Cell Res. 2013 Feb;23(2):213–24.

39. Liu Y, Bernard HU, Apt D. NFI-B3, a novel transcriptional repressor of the nuclear factor I family, is generated by alternative RNA processing. J Biol Chem. 1997 Apr 18;272(16):10739–45.

40. Nie Z, Guo C, Das SK, Chow CC, Batchelor E, Simons SS, et al. Dissecting transcriptional amplification by MYC. eLife. 2020 27;9.

All authors have read and approved the manuscript.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Highlights

- The Gene Ontology (GO) provides a rigorously defined set of concepts that describe the functions of gene products
- GO annotations link a gene product and a GO concept, and are supported by scientific evidence
- Transcription, which plays a central role in defining the identity and functionalities of cells, is mediated by large complexes, and delineating the function of individual gene product can be challenging
- To improve consistency of the GO data, we present recommendations elaborated by the GO consortium and GREEKC members