NLM-GENE Corpus Annotation Guidelines

Rezarta Islamaj, Janice Ward, David Cissel, Keiko Sekiya, Oleg Rodionov, Nicholas Miliaras, and Olga Printseva

The purpose of the project is to create a collection of PubMed documents (corpus), annotated with the location and identity of interesting gene entities. The corpus will be used to improve the quality of the automated taggers distributed by the NCBI BioNLP group and will also be described and distributed as a standalone / independent product. The corpus and tagger improvements are intended to be useful for multiple tasks.

# How to annotate and what to mark:

To help create a corpus that identifies all gene mentions in text, we will use the TeamTat tool. All PubMed documents have been pre-annotated with GeneNorm+, and Species mentions have been highlighted to help annotators with their task. In TeamTat, annotators can:

1. Mark a document as “curatable”, to denote relevancy to the task
2. Review the pre-annotated genes in the document

Make corrections to the GeneID or highlighted text as needed

Delete incorrect annotations

3 Highlight Gene mentions not annotated by the tool and add the correct Entrez Gene ID

4 Assign the appropriate gene annotation type for all annotated genes

**GENERIF** – the annotated gene meets the criteria for creating a GeneRIF - the basic biology or

clinical significance of a gene/gene product is the primary point of the article.

**STARGENE** – the annotated gene is a main point of the article but does not meet the criteria for

creating a GeneRIF. (It is implied that a gene mention tagged as GENERIF is automatically the

STARGENE of the article)

**GENE –** the annotated gene is mentioned, but is not a main point.

**DOMAIN –** the word denotes a protein domain

**OTHER** – the gene mentioned is eithera gene product used as therapeutic or pharmacological agent, or a gene used as a tool (e.g. marker gene, gene used in techniques, etc.)

5. Mark a document as “done” to denote that you have finalized its curation

The tool allows quick access to:

1. The article’s PubMed and PubMed Central versions, in case the annotators need to consult the PubMed version or the full text of the article
2. The Entrez Gene Page for gene entities, and/or Taxonomy page for species entities, if the identifiers have been assigned

A notes field is provided for annotators to add comments/explanations regarding the annotation if needed.

Annotation Note:

* If multiple genes are listed as an overlapping ellipsis (or span), annotate the full phrase as a single mention, and assign GENE IDs to each gene in the elliptical string (or span) in order of appearance, separated by semicolons.
* If a text phrase needs to be annotated with multiple GENE IDs, highlight the full phrase as a single mention, and assign GENE IDs to each gene in the span separated by commas.

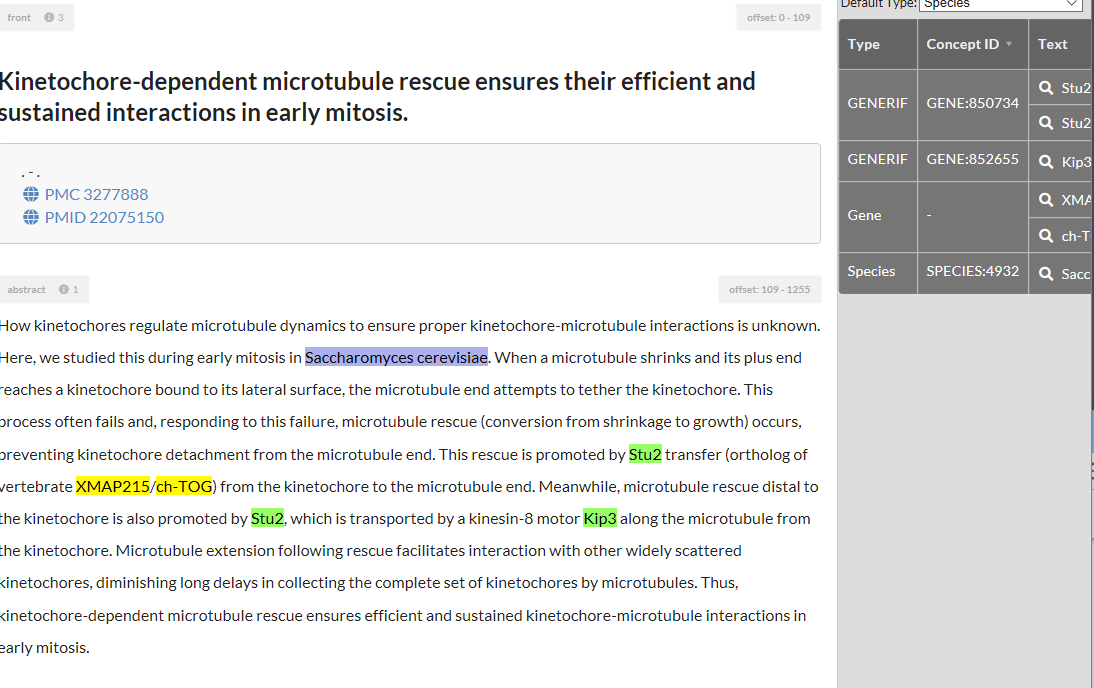


Figure 1 TeamTat tool for gene annotation

# The Annotators:

Six National Library of Medicine Indexers have volunteered to annotate PubMed documents for gene entities. The task is to be as comprehensive as possible covering every single gene entity, and grounding them to their correct identifiers. If needed the article’s full text will be consulted to identify the correct concept.

During our meetings, we aim to discuss issues to be consistent and agree with each other on annotating all gene mentions, abbreviated terms included, and selecting the same tags.

# The Gene Mentions:

Genes in scientific publications are mentioned by their scientific name, abbreviated form, or different variations.

1. If a text mention is a known name/shorthand notation of a gene, or can be traced to a known synonym of the gene, the mention will be annotated.

Examples:

Drosophila ovarian germline stem cells (GSCs) are maintained by Dpp [GENE: 33432] signaling and the Pumilio [GENE: 41094] (Pum [GENE: 41094]) and Nanos [GENE: 42297] (Nos [GENE: 42297]) translational repressors. (PMID: 21238926)

1. All genes in a list will be annotated separately. If a mention refers to more than one gene, all gene ids will be listed.

Examples:

* VEGF/ PDGF ligands [GENE: 32876, 33994, 33995] from the tubules attract hemocytes, which secrete components of the basement membrane to ensheath them. (PMID: 20708591)
* mRNA expression levels of genes involved in intestinal cholesterol transport and esterification were unchanged, but we observed downregulation of HMG-CoA reductase and synthase [GENE: 15357, 15360] and consequently less intestinal cholesterol biosynthesis. (PMID:22842588)

1. If the gene mention is not one contiguous string, highlight the whole mention, put the correct ID, and use the notes field to write the correct gene name.

Example:

* An Aγ-globin G->A gene polymorphism associated with β039 thalassemia globin [GENE: 3043, Mention: beta globin] gene and ... (PMID: 28851297)

More examples:

PMID: 24141718

* We show that polypyrimidine tract-binding protein [GENE: 5725] (PTB [GENE: 5725]) is central to one such complex that forms in apoptotic cells.
* Thus, during apoptosis initiated by TNF-related apoptosis inducing ligand [GENE: 8743] there is a change in the repertoire of RNA-binding proteins with which PTB [GENE: 5725] interacts.
* We show that altering the cellular levels of PTB [GENE: 5725] and its binding partners, either singly or in combination, is sufficient to directly change the rates of apoptosis with increased expression of PTB[GENE: 5725], YBX1 [GENE: 4904], PSF [GENE: 6421] and NONO[GENE: 4841]/p54(nrb) [GENE: 4841] accelerating this process.

# How to link to Entrez Gene:

The second part of annotation concerns the grounding of the concept to a known entry in Entrez Gene.

1. If a text mention is linked to a specific species and is a known name/shorthand notation, or can be traced to a known synonym of the gene, the mention will be annotated with the corresponding Entrez Gene ID.
2. Sometimes the same article may contrast genes from different organisms. While the mentions could be identical, the annotators will assign the correct ID, as specified by the context.

Examples:

* NMR structure of an acyl-carrier protein [GENE: 1195557] from Borrelia burgdorferi. A structure-homology search revealed that this protein is highly similar to the acyl-carrier protein [GENE: 1193326] from Aquifex aeolicus. (PMID:21904063)
* The cyclin-dependent kinase inhibitor p21[GENE: 12575] (waf1[GENE: 12575]/cip [GENE: 12575]) mediates the p53[GENE: 22059] -dependent G1/S checkpoint, which is generally considered to be a critical requirement to maintain genomic stability after DNA damage. (PMID:22162343)
* P21, waf1, cip are all mentions of the GENE: 12575. They should be annotated as separate mentions.

1. If a text mention is a protein, which is a product from two different genes, both shall be assigned as gene IDs.

Examples:

Hemocyte-secreted type IV collagen [GENE: 33726, 33727] enhances BMP [GENE: 33432] signaling to guide renal tubule morphogenesis in Drosophila. (PMID: 20708591)

1. How to annotate genes in mentions of experimental organisms:

In experimental studies, annotate genes for the organism that is the source of the gene (for example mouse gene transfected into human cells is annotated to record for mouse gene).

Examples:

* Here, we describe a key role of nitric oxide, an important signaling molecule in adult skeletal muscle, on satellite cells maintenance, studied ex vivo on isolated myofibers and in vivo using the a-sarcoglycan [GENE: 20391] null mouse model of dystrophy and a cardiotoxin-induced model of repetitive damage. (PMID:22084027)
* Here we report that when exposed to hypoxia/oxidative stress, a small fraction of hESCs, namely the SSEA3+/ABCG2+ [GENE: 9429] fraction undergoes a transient state of reprogramming to a low p53 [GENE: 7157] and high hypoxia inducible factor (HIF)-2a [GENE: 2034] state of transcriptional activity. (PMID:22689594)
* Cholesteryl ester accumulation and accelerated cholesterol absorption in intestine-specific hormone sensitive lipase-null mice. [GENE: 16890]
* The use of the ShcD knockout ESCs allowed the unmasking of this process as they presented deregulated Oct4 modulation and an enrichment in Oct4-negative Cdx2-positive cells with increased MAPK/extracellular-regulated kinases 1/2 activation, within the differentiating population.
* Intriguingly, although Bak(-/-) [GENE: 12018] mice have elevated platelet counts, Bak [GENE: 12018] (-/-)vavP-BCL-2 [GENE: 596] mice, like vavP-BCL-2 [GENE: 596] littermates, were thrombocytopaenic.
* However, in Rag1(-/-) [GENE: 19373] BCL-2tg [GENE: 596] mice, platelet levels were normal, implying that elevated lymphocytes are primarily responsible for BCL-2tg-induced [GENE: 596] thrombocytopaenia.

1. How to annotate mutations?

Annotate only the gene and do not include designations describing the mutation (position, etc.)

Examples:

* As a first step towards determining the structure of A(2A)R [GENE: 135] bound to an agonist, the receptor was thermostabilised by systematic mutagenesis in the presence of the bound agonist [(3)H]5'-N-ethylcarboxamidoadenosine (NECA). Four thermostabilising mutations were identified that when combined to give mutant A(2A)R-GL26 [GENE: 135], conferred a greater than 200-fold decrease in its rate of unfolding compared to the wild-type receptor. (PMID:21501622)
* A homozygous Arabidopsis T-DNA insertion mutant (alg10-1) [GENE: 831764] accumulated mainly lipid-linked Glc(2) Man(9) GlcNAc(2) and displayed a severe protein underglycosylation defect. (PMID:21707802)

Questions:

* **Do we annotate gene variant mentions that are not standard synonyms for the given gene?**

ANSWER: Do not annotate mentions of variants of a gene that are designated by the author which are not standard synonyms

Example: (PMID:21965601)

In this example, F57I and D67H are variants of the GENE 4069, whose mention is lysozyme. Lysozyme is annotated in this article.

*To achieve this objective, wild-type (WT) protein and the amyloidogenic variants F57I and D67H were expressed in Drosophila melanogaster using the UAS-gal4 system and both the ubiquitous and retinal expression drivers Act5C-gal4 and gmr-gal4.*

*We observed that expressing the destabilized F57I and D67H lysozymes [GENE: 4069] triggers UPR activation, resulting in degradation of these variants, whereas the WT lysozyme [GENE: 4069] is secreted into the fly hemolymph.*

* **Do we annotate specific MicroRNAs?**

MicroRNA-33b (miR-33b) is embedded in intron 16 of porcine SREBF1 [GENE: 397308] and is conserved among most mammals. (PMID: 24398549)

Answer:

If in system, annotate appropriately, if not, use NEW ENTRY code, if no orthologs exist, leave empty and write in Note field “No ortholog”

* **How to annotate general mentions of genes, that are not specific to a given organism? Often this occurs in an introductory sentence as background information.**

ANSWER: Annotate with a code. For example, [GENE: 397308-000], where 397308 refers to the gene for the organism mentioned in the article, and 000 signals that this is talking about the gene in general sense. If the organism is not mentioned in the article, then we can pick the human gene ID.

Examples:

Transforming growth factor (TGF)-β1 (GENE:7040-000) is a pleiotropic cytokine involved in MSC migration, differentiation, and immunomodulation.

* Selected human TGFB1 because human is the organism studied in this article; added code 000 to indicate a general mention of this gene
* **How to annotate GENE mentions that should be in Entrez GENE, but do not have an ID yet (NEW ENTRY)?**

ANSWER:

We should not annotate with a code that refers to a NEWENTRY. If not in the database, use another GENE ID (of an ortholog) with the code -111, for example [GENE: XXXX-111]. In the notes section write the correct organism, and its TAXONOMY ID.

If an ortholog is mentioned in the article, then pick that one to use with the code.

* **How to annotate gene families/groups/classes?**

Create a code: GENE: XXXX-222, where 222 denotes gene family, or class. Annotate the gene family mention with the specific GENE IDs of the members of that family listed in the abstract, adding the family code.

If no specific member is mentioned in the article, then pick a gene member of that family of the main Organism of the article and add code 333.

Examples:

* Mechanistically, ILK [GENE: 16202] promotes the recruitment of the F-actin binding protein [GENE: 29875-222] IQGAP1 [GENE: 29875] to the cell cortex, which, in turn, cooperates with its effector mDia1 [GENE: 13367] to locally stabilize MTs and to allow stable insertion of caveolae into the plasma membrane. (PMID: 20951348)
* ANSWER: Annotate as [GENE: 29875-222] where “222” is the code for “family/class”. In this annotation we are saying that IQGAP1 with ID: 29875 is a member of the “F-actin binding protein” family.
* The de novo methyltransferase [GENE: 13436-222, 13435-222, 54427-222] Dnmt3b [GENE: 13436] is required for methylation of both classes of CGI, whereas Dnmt3a [GENE: 13435] and Dnmt3L [GENE: 54427] are dispensable.
* ANSWER: following the family convention, “de novo methyltransferase” would be tagged with [GENE: 13436-222, 13435-222, 54427-222], comprising all elements of the family mentioned in the abstract.
* Our results assign an important role to the integrin [GENE: 16412-222] /ILK [GENE: 16202] complex for caveolar trafficking to the cell surface. (PMID: 20951348)
* ANSWER: following the family convention, “integrin” would be tagged by listing all members of its family mentioned in the abstract. Beta1 integrin is the only integrin mentioned in the abstract
* Nudix proteins play an important role in regulating the intracellular concentration of nucleotide cofactors and signaling molecules. (PMID:21904053)
* ANSWER: following the family convention, “nudix proteins” would be tagged by listing the other members mentioned in the other sentences of the abstract [GENE: 29621275-222, 29622055-222, 944824-222].
* **How to annotate gene families/groups/classes for which we do not have specific members in the current abstract?**

Create a code: GENE: XXXX-333, where 333 denotes a gene family (or class) that does not have specific members in the current abstract.

* Gartanin induces cell cycle arrest and autophagy and suppresses migration involving PI3K **[GENE:** 5290-333**]** /Akt **[GENE:** 207-333**]** /mTOR **[GENE:** 2475-333**]** and MAPK **[GENE:** 5594-333] signalling pathway in human glioma cells. (PMID: 27491646)
* ANSWER: pick the main organism of the abstract, then use the Gene ID of that organism that bel**o**ngs to this group that is ranked first in the GENE results, and use code 333, to denote the special case.
* Furthermore, it increased the number of Pax7 (+)[GENE: 18509] /Myf5 (-) [GENE: 17877] cells in a cGMP-independent pathway requiring enhanced expression of Vangl2 [GENE: 93840], a member of the planar cell polarity pathway involved in the Wnt [GENE: 22408-333] noncanonical pathway. (PMID:22084027)
* ANSWER: “Wnt” should be annotated with the family/group convention. (selected mouse wnt1 because it was first in the GENE results and added code 333)
* **How to annotate mentions of complexes (**includes multiprotein complexes and multi-subunit proteins

Create a code: GENE: XXXX-444, where 444 denotes a complex, that has at least a specific

member mentioned in the current abstract.

If no specific member is found in the current abstract, then similar to the family/class rule, pick the main organism of the article, and use gene member of that organism to annotate the complex with the code -555.

* We also show that RABL2**[**GENE:11158**]** interacts, in its GTP-bound state, with the intraflagellar transport (IFT)-B [GENE: 80173-444, 28981-444] complex via the IFT74 [GENE: 80173] -IFT81[GENE: 28981] heterodimer and that the interaction is disrupted by a mutation found in male infertile mice (Mot mice) with a sperm flagella motility defect. (PMID: 28428259)

Two genes that are part of the Intraflagellar transport-B complex are mentioned in the abstract, so the complex is annotated with the Gene IDs for the two genes with the code -444

* How to annotate mentions of domains?

We mark the mention and select type “protein domain”

* The immunoglobulin [ GENE:3514-333 ] domain of the sodium channel β3 subunit [Gene: 55800] contains a surface-localized disulfide bondthat is required for homophilic binding (PMID23118027)
* **How to annotate if multiple codes apply?**

More than one code may apply for a gene mention. If the gene mention is both general (code 000) and a family (code 222 or 333) or a complex (code 444 or 555) annotate only with the appropriate code for family or complex and ignore code 000.

If the gene mention is both a family (code 222 or 333) and a complex (code 444 or 555), annotate only with the family code and ignore the complex code.

Examples:

The CB1 cannabinoid receptor [GENE:12801-000], the main molecular target of endocannabinoids and cannabis active components, is the most abundant G protein-coupled receptor[GENE:12801-222] in the mammalian brain.

* For “G protein-coupled receptor” selected mouse gene because mouse is the organism studied in this article and added code 222 to the gene for mouse CB1 because G protein-coupled receptor is a group term.

The apoptotic rates resulting from 2DG-ABT treatment were higher in the cells treated with the PI3K (GENE: 5291-333) inhibitor, while the rates remained approximately the same in the cells treated with the ERK inhibitor. (PMID:27460078)

* PI3K is a family and also a complex composed of subunits so code 333 was used

What Not to Annotate

* Do not annotate gene products used as therapeutic or pharmacological agents -- Other

Example:

Genome-wide loss-of-function genetic screening identifies opioid receptor μ1 as a key regulator of L-asparaginase resistance in pediatric acute lymphoblastic leukemia (PMID28650467)

asparaginase is used therapeutically so will not be annotated.

* Do not annotate genes used as tools (e.g. marker genes, genes used in techniques) -- Other.

Examples:

In Wif1(lacZ/lacZ) mutant mice and cultured urorectum with exogenous Wif1, cloaca septation was defective with undescended urs and hypospadias-like phenotypes, and such septation defects were also observed in Shh(-/-) mutants and in endodermal β-catenin gain-of-function (GOF) mutants

lacZ is used as a tool to study Wif1. It will not be annotated. (PMID24632949)

To specifically delete the LIFR in the LE, we derived a line of mice in which Cre recombinase was inserted into the endogenous lactoferrin gene (Ltf-Cre) (PMID2836837)

Cre recombinase (Cre) is used as a genetic tool so it will not be annotated

Consistently, the surviving CD4+YFP+GFP+ T cell-derived cells were unresponsive and failed to proliferate during the early phase of secondary infection (PMID 27630165)

YFP (yellow fluorescent protein) and GFP (green fluorescent protein) are used to tag other

proteins so they will not be annotated.

* A discontinued record is marked as Other. Link to discontinued record.
* Do not annotate very general words like

Protein

Enzyme

Protein Isoforms

Isoenzymes

Organism Specific Proteins (e.g. Drosophila Proteins, Arabidopsis Proteins)

Location-Specific Proteins (e.g. Membrane Proteins, Nuclear Proteins, Blood Proteins)

Micro RNA

Non-coding RNA