Contents

Annotation Guidelines for the *Physcomitrella patens* genome version 1.2/1.6

Last major change
2010-10-05

What has changed?

The PHYPA1 genome portal at JGI will continue to be online, but annotation has been transferred to cosmoss.org, effective July 2009. All user annotations up to June 2009 have been transferred, except for changes to gene models (see below, structural annotation).

Important links

- Cosmoss genome portal: http://www.cosmoss.org/
- Cosmoss genome browser: https://www.cosmoss.org/mgb2/gbrowse/physcome/
- Genonaut annotation interface: https://www.cosmoss.org/annotation/genonaut/
- How to use the annotation interface: https://www.cosmoss.orghttps://www.cosmoss.org/physcome_project/wiki/Genonaut
- Genome project wiki: https://www.cosmoss.orghttps://www.cosmoss.org/physcome_project/wiki/Main_Page
- Cosmoss accounts: https://www.cosmoss.orghttps://www.cosmoss.org/physcome_project/wiki/Cosmoss_accounts
- Account generation: send e-mail to helpdesk-cosmoss@uhura.biologie.uni-freiburg.de

Annotators

In order to be able to annotate *Physcomitrella patens* gene models, you need to be a registered cosmoss user. With this account you can access the genonaut annotation interface and non-public information in the wiki. It also enables you to use more cosmoss.org features, like an additional set of databases. The list of annotators is available to registered users.

Gene-specific updates

Keep up to date with your gene of interest - all annotators for a given gene model will be notified if new annotation is added to any term of the gene.

Annotation privilege

If a user is a constant source of mischief or adds inappropriate material (spam) using the interface, the site administrators can revoke the ?annotator? privilege.

Structural annotation

It is not yet possible to apply changes to gene structures, however, an appropriate interface will be made available in the future. If you have applied structural changes/promoted alternative gene models using the JGI browser please tell us, so we can promote the model on cosmoss.org. Also, please feel free to name alternative models to us that shall replace the selected V1.2/1.6 model in the next release.

Versions and future releases

In agreement with the NCBI and JGI (phytozome.org), future genome annotations will be unified. The published annotation (Rensing et al. 2008 Science 319:64) is V1.1, which is based on the V1 assembly. Based on the same assembly, the V1.2 annotation release (devoid of further contaminations and gene models overlapping with transposons) has been generated and is available on cosmoss.org. It has also been made available to NCBI and JGI and will be made available there. The V1.6 is still based on the V1 assembly, but features improved gene models. The V1.6 is planned to be released later in 2009. The Physcomitrella genome consortium is collaborating with the JGI to generate the v2 assembly and subsequently the V2.0 annotation.

This release is not to be expected before 2010.

Conflict management

The user who initially provides an annotation, is considered the owner of the annotation. Ownership only refers to a specific annotation and not the entire gene. If you alter an annotation that has previously been created by someone else, a conflict is raised and the term in question is blocked for further modification until the conflict is resolved. For this the original annotator is notified via e-mail notifying him/her of the change and he/she has the opportunity to accept or deny the change within four weeks. If the change is accepted, the ownership of the annotation is changed to the new annotator. If he/she does not respond, the change will be accepted automatically and the new annotator is the owner. There are no irresolvable conflicts. If you are not satisfied with the original annotators judgment on your modification, a curator can be called in to moderate the discussion. Curators can finalize annotations to resolve annotation conflicts.

Curators

Currently there are four curators from the cosmoss.org team (Andreas Zimmer, Daniel Lang, Karol Buchta and Stefan Rensing), who are also the site administrators. Curators can force changes, oversee use annotations, and work on future annotation releases. If you are interested to become a curator, please tell us.

Gene identifiers and accession numbers

On cosmoss genes and gene products can be accessed via a *primary identifier* and an *accession number*. These are used as *unique, non-redundant database identifiers* describing the *technical concept of gene models* to access sequence and annotation records describing their encoded gene products.

This is in contrast to the *biological concept of gene and protein names and description lines*, which can change as our knowledge about a particular gene increases.

Note

As an annotator, you cannot assign gene ids.

Cosmoss gene (locus) IDs (CGIs)

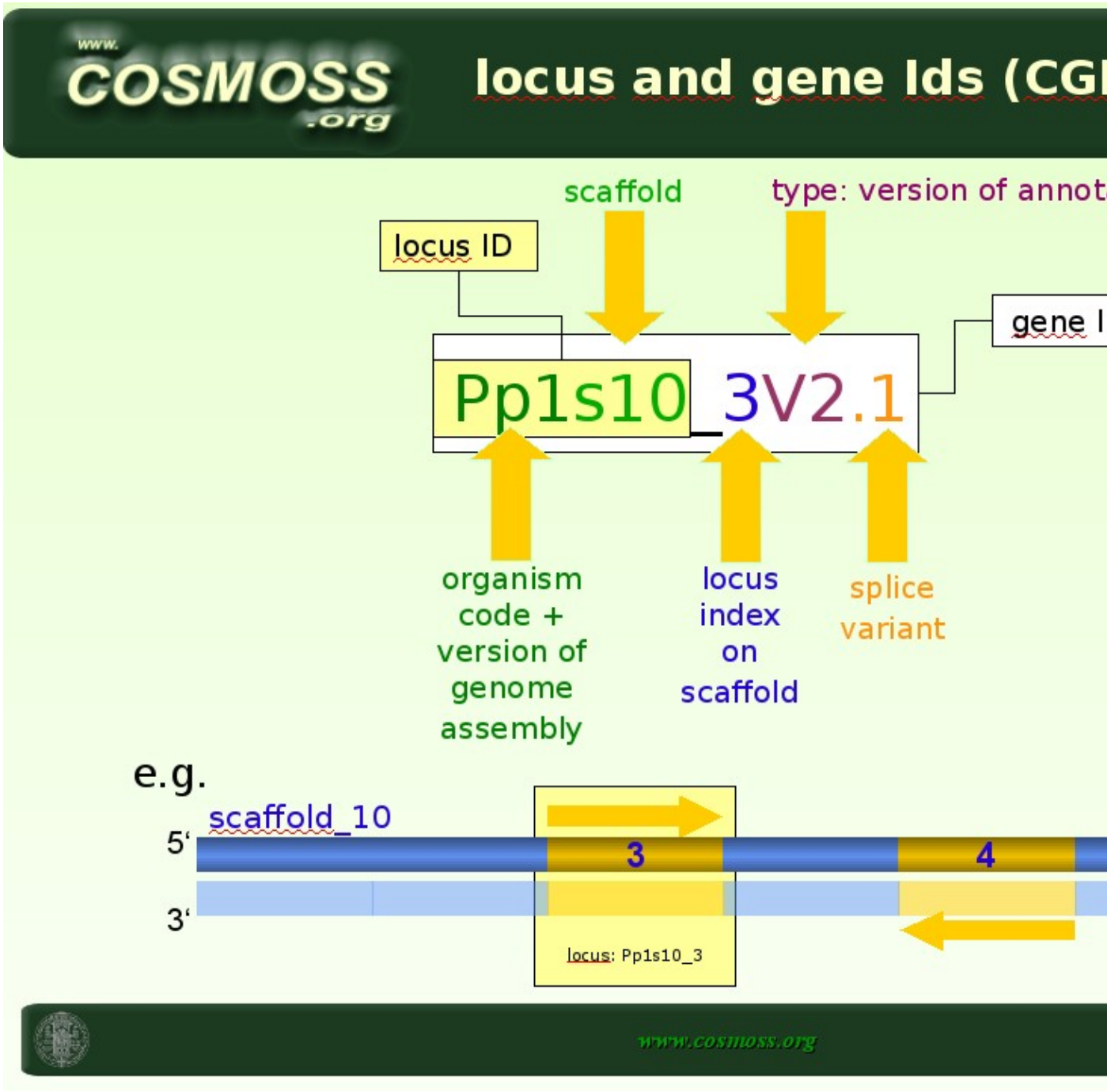
On cosmoss each gene model has a primary identifier (NCBI locus_tag), the unique gene (locus) ID (cosmoss gene id; CGI). The CGI provides a unique address to a gene (model) for a given genome (assembly/annotation) release. Using a clustering procedure, all overlapping gene models are grouped into a unique locus. For each locus a unique number is assigned, which is specific for a given assembly. All CGIs include the number of the scaffold and the number of the locus they belong to as well as information on the version of the assembly and annotation or the gene predictor they are derived from.

CGI Syntax

```
OrganismCode+AssemblyVersion+ScaffoldNumber+_+LocusNumber+Type+. +
SpliceVariant
```

e.g

`Pp1s275_2V2.1` (= Phypa_196781)



Phypa_ID

The Moss genome community decided at Moss 2006 to give a unique 6-digit numerical accession number (e.g. Phypa_123001) to each gene model, which serves as its permanent accession number. The Phypa_IDs are identical with the protein_ids assigned by the JGI.

With each new release, new models receive their own Phypa_IDs. If the model is unchanged (start and stop coordinates unchanged), the Phypa_ID remains unchanged as well. Accession numbers of discarded models will not be used again.

Which IDs to cite in you paper

Please use cosmoss gene ids (CGI;Pp1sX_XVX.X) only and use the wiki page (e.g. Genome_Annotation/V1.6) to refer to the release.

To avoid confusion and since the nomenclature of CGIs always carries information about the gene locus, it a good practice to cite only the model that you've used in your analysis, independent of whether its a release model or not.

If you've found one of the other models (V1.2 or models from the all models track) better models, use the respective CGI in the method section of you paper. In case you've altered the model using Apollo, put in the CGI of the user model you've created (Pp1sX_XU2.X_username).

In the case an existing model from another track than the current release is already the "best" description of the gene, please send us the CGIs.

Manual annotation ? naming conventions

When completing the name, alias and description fields, bear in mind that your annotation will eventually be translated into a Genbank entry. For an example of such an entry, see: http://www.ncbi.nlm.nih.gov/nucleotide/168053776

Names and descriptions should convey some meaning as to the function of the gene product. Names based upon a quantifiable feature such as biochemical assay, protein-protein or genetic interaction, or mutant phenotype are preferred to names based upon sequence similarity alone. Regardless of the derivation, it is not likely that a name or a description can convey all that is known about a particular gene and names have changed to reflect new knowledge.

(Short) gene and protein name

Short acronyms or symbols that represent this locus based on annotation or previous name. Don?t assign a name if function is unknown (the automatically assigned unique identifier will be sufficient).

Although, names and synonyms can be redundant, we strongly discourage you to do so. Before assigning a name check whether it has ambiguous meaning e.g. by cross-checking other databases like the Arabidopsis Gene Symbol Registry.

Gene name

Gene names are written in small letters and protein names in CAPITAL letters. At least the first letter of a gene name has to be small.

Protein name

Usually, the protein name will be the gene name in CAPITAL letters. There is no formatting (italics, bold face, underline) whatsoever available.

An organism letter code like **Pp** must not be included in the gene name annotation. It is only relevant in the comparative sense, i.e. when comparing the individual members of a gene family in multiple organisms. Organism specific prefixes that are appended for clarity in publications should not be part of the gene name.

Please keep in mind that the primary identifier of a gene or protein is the cosmoss gene id and the accession number Phypa_ID, which provide unique and unambiguous references for database retrieval and should always be included in publications (at least in the Methods section).

Gene indexing

To identify paralogous members of gene families, you can use indexes, such as:

```
hsp70_2 gene for putative heat shock 70 class protein HSP70_2
mads4 gene for putative MADS-box family protein MADS4
rl20_1 gene for putative ribosomal large subunit protein RL20_1
```

If you have not carried out a phylogenetic analysis and/or can reference to a publication describing this nomenclature, please refrain from assigning gene indexes. In particular, please do not assign gene indexes based on the order of BLAST hits or in an arbitrary fashion. Any numbering has to be supported with a cosmoss reference describing the phylogenetic analysis. If the analysis is not published in a peer reviewed article then you can provide your evidence as a custom reference. Corresponding phylogenetic trees can be made available as well via a special cosmoss interface e.g. https://www.cosmoss.org/bm/supplementary_trees/Rensing_et_al_2008 or by cross-referencing a published phylogenetic study in TreeBASE .

Using subfamily indices

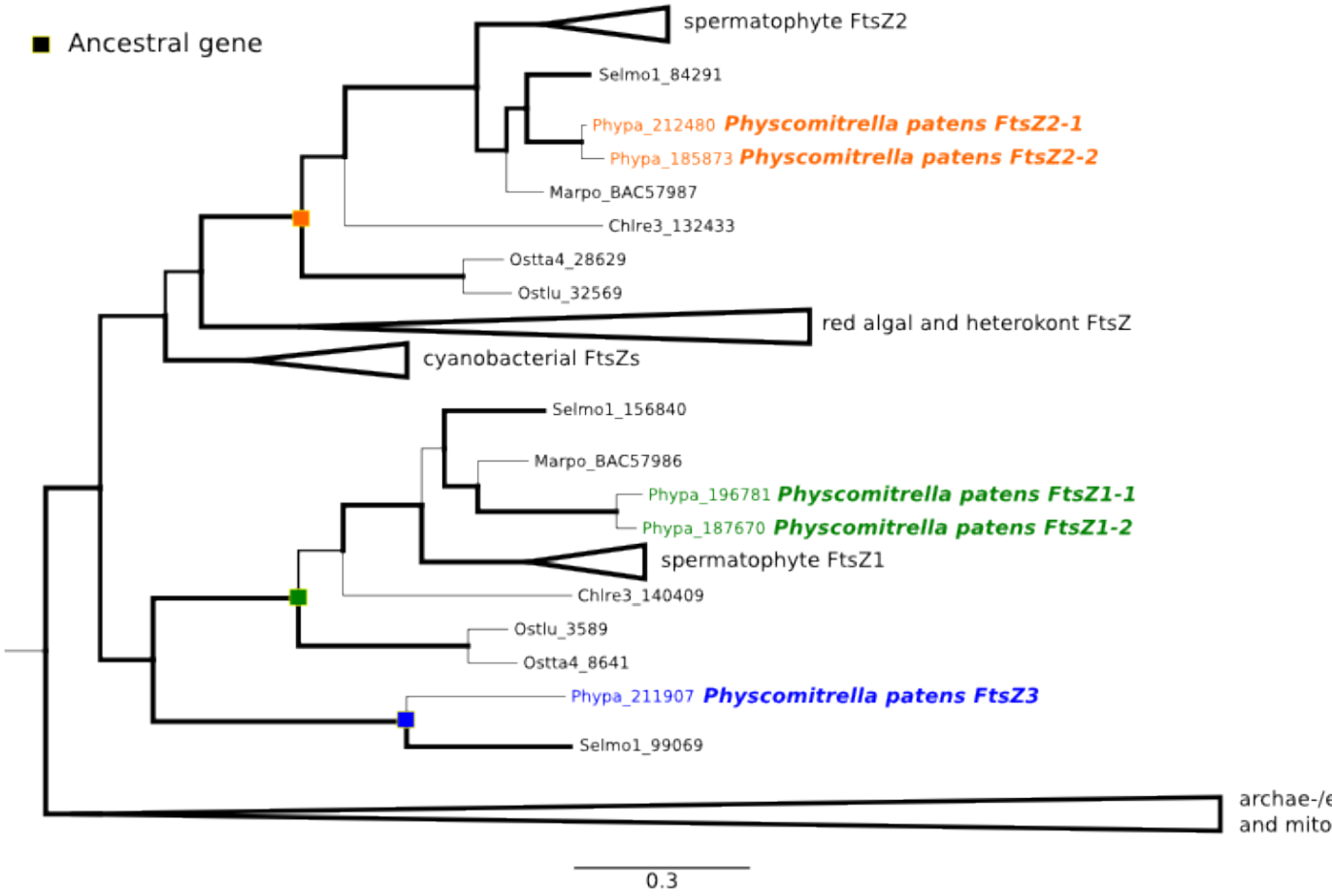
Naming of lineage-specific inparalogs or paleologs i.e. paralogs derived from the whole genome duplication event. Subfamily indices should by assigned in the following fashion:

```
hsp70_2-1 gene for putative heat shock 70 class protein HSP70_2-1
i.e. 2 denoting subfamily 2 and 1 gene #1
```

Index separators

We do not enforce the use of specific index separators ("-",";","_","."). If the name is based on an existing gene family name which contains a number (e.g. hsp70), an index separator should be used to provide gene indices identifying the members of the gene family in *Physcomitrella* (e.g. "hsp70_2" or "hsp70-2" or "hsp70.2"). The only requirement is that this scheme has to be applied consistently throughout the entire gene family.

Example: FtsZ gene family



- renaming of FtsZ gene family based on improved phylogenetic resolution (Martin et al 2009:Plant Biology 11, 744-750)
- renaming of a gene: old name: *ftsZ1-2* changed to new name: *ftsZ3* due to phylogenetic evidence indicating a distinct ftsZ3 subfamily absent in seed plants
- The *Physcomitrella* ftsZ gene family is comprised of three 3 subfamilies: *ftsZ3* (blue) represents different ftsZ subfamily than *ftsZ1-1* / *ftsZ1-2* (green) and *ftsZ2-1* / *ftsZ2-2* (orange)

Splice variants

If there are multiple splice variants for a gene, all variants carry the same gene name. If the alternative transcripts encode for different proteins this can be indicated by adding indexing using alphabetical indices [a-z]:

```
glp7 gene for germin-like protein variant GLP7a
glp7 gene for germin-like protein variant GLP7b
```

Aliases/synonyms

Several aliases can be given to each gene. Aliases represent alternative gene/protein names.

If a previous gene name is replaced, genonaut will ask you whether you want to keep the old name as a synonym. We advice you strongly to do so.

Full name and description (line)

This field will be part of the field *Name* or *product* at NCBI, and will accompany the sequence as the description line in Fasta format, right after the identifier. It should be a short (<85 characters) precise description of the gene and gene product and if possible its main function(s). It should include the **full standard name, ideally explaining the acronyms used as short gene and protein names**.

General description rules

- If it exists, use the approved nomenclature.
- Use a concise description, not a comment or phrase.
- Ideally the description should be unique and attributed to all orthologs.
- Avoid the use of molecular weights in protein names; "unicornase subunit A" is preferred to "unicornase 52 kDa subunit"
- Do not use the term "homolog" if it has not been determined via a phylogenetic analysis.
- Where possible, avoid the use of commas
- Don't add species, gene identifier or accession number (e.g. *Physcomitrella patens* Phypa_10204)
- Don't include accession numbers of foreign databases (e.g. PFAM, EC numbers, Arabidopsis gene ids...)
- Use lowercase letters, except when uppercase are required (for example, in acronyms such as DNA or ATP).
- Wherever appropriate, the name should use American spelling conventions.
- Do not build molecular weights into abbreviations
- Greek letters must be written in full e.g. "alpha", and written entirely in lower case with the exception of "Delta" in the context of steroid/fatty acid metabolism nomenclature. *Additionally the Greek letters that are followed by a number should be preceded or followed by a dash "-" e.g. "unicornase alpha-1".
- Do not use diacritics, such as accents, umlauts. Many computer systems can only understand ASCII characters.

--Lang 09:05, 5 October 2010 (UTC)adapted from the NCBI nomenclature for naming proteins

Genes encoding proteins with experimental evidence

```
xyz gene for some function protein XYZ
```

--> i.e. use no qualifier

reference to your experimental evidence

Annotations of this category have to be provided with a cosmoss reference entry (PubMed or custom reference) describing the evidence.

Function deduced by similarity

```
xyz gene for putative some function protein XYZ
```

e.g.

```
hsp70 gene for putative heat shock 70 class protein HSP70

mads gene for putative MADS-box family protein MADS

rl20 gene for putative ribosomal large subunit protein RL20
```

--> i.e. use the qualifier *putative* if the function is deduced by similarity (stringent BLAST filtering or phylogenetic analysis indicates that genes are homologous, see also Cosmoss_workshop_2009#BLAST.2C_homology_and_hit_filtering)

Genes with unsure similarity deduction

Used if there is only weak evidence for sequence homology or homology is only confined to a domain that is found in multiple gene families.

based on weak sequence homology

```
gene for XYZ-like protein

e.g. cytochrome B-like protein
```

containing a (promiscuous) domain

```
gene for <domain|repeat>-containing protein

e.g. ankyrin-repeat containing protein
```

--Lang 08:20, 5 October 2010 (UTC)the previously suggested *similar to* is deprecated to be conform to NCBI gene product naming guidelines.

Genes encoding proteins with unknown function

```
gene for hypothetical protein
```

--> i.e. use the qualifier ?hypothetical?

Pseudogenes

```
xyz pseudogene
xyz putative pseudogene
```

--> e.g. for truncated gene models, models with premature stop-codons, models within a transposable element context.

Additional qualifiers

Note
: these are optional! If not annotated manually, they will be assigned automatically prior to release.

tentative
: For models that are not supported by transcript data

```
tentative xyz gene for putative some function protein XYZ
```

expressed
: If there is expression evidence (RT, QPCR, Microarray, EST, SAGE...)

```
expressed xyz gene for hypothetical protein
```

Notes (comments)

The entry in this field will be transferred to the NCBI Gene page in the field ?Comment?. It can be as long as you want, provided the information is accurate and useful to researchers not familiar with this type of protein. Include information about the functions of the protein, its domains, splicing variants, interactions or subcellular location, comments about its phylogenetic origin, relationship to paralogs and orthologs, clustering with genes of related function, or overlap with neighboring genes etc...

Additional info

Annotation terminology and conventions

In genonaut, (functional) annotation for a gene can be provided in the form of values for a given term. Currently we support the terms:

- name
- alias
- description
- Gene Ontology terms:
    - ♦ cellular component (GO:CC)
    - ♦ biological process (GO:BP)
    - ♦ molecular function (GO:MF)

- note

For each specific value/annotation (e.g. the name ftsZ1-1) you can provide a reference that contains the basis/evidence for your annotation (see References for details). In addition, for Gene Ontology annotations you have to provide an evidence code, a three letter acronym to indicate the reliability of the annotation (see http://www.geneontology.org/GO.evidence.shtml for details). Commonly used evidence codes are

- IEA Inferred from Electronic Annotation
- IDA Inferred from Direct Assay
- ISS Inferred from Sequence or Structural Similarity
- IMP Inferred from Mutant Phenotype
- RCA Inferred from Reviewed Computational Analysis
- TAS Traceable Author Statement

Note

you can get info on when to use which evidence code in the mouse-over information within the genonaut manual annotation interface.

References

To allow you to either cite an already published paper that is indexed via PubMed or a custom reference to a peer reviewed article in a journal not indexed by PubMed or an unpublished evidence, we have introduced **cosmoss references**:

- every reference submitted by an annotator gets a unique cosmoss_ref ID and a stable URL:
- e.g. https://www.cosmoss.org/annotation/references?cosmoss_ref=1
- this cosmoss reference ID can be used for stable cross-reference between different databases
- references can have two different scopes:
    - ♦ references can be specific for a certain feature e.g. the name or a GO term
    - ♦ general references for a gene

There are four ways to cite in genonaut:

- reuse an existing cosmoss reference ID
- PubMed ID?s: provide a valid PubMed ID (PMID) and the interface collects all information automatically from PubMed
- custom references: in case the journal isn?t listed in PubMed or the article is not yet published
- comments: a plain text field and/or a URL to refer to other databases and sources.