This compilation of curation samples serves as a guide for the use of ontology terms, evidence codes, qualifiers, and free-text notes (for additional disease data). The examples will mainly cover gene ontology and disease curation.

The following material on evidence codes is based on material from the Gene Ontology website (http://www.geneontology.org/GO.evidence.shtml), except examples labeled "**RGD curation samples**".

1. Gene Ontology Curation

Guide to GO Evidence Codes

Introduction

A GO annotation consists of a GO term associated with a specific literature reference that describes the work or analysis upon which the association between a specific GO term and gene product is based. Each annotation must also include an evidence code to indicate how the annotation to a particular term is supported. Although evidence codes do reflect the type of work or analysis described in the cited reference which supports the GO term to gene product association, they are not necessarily a classification of types of experiments/analyses. Note that these evidence codes are intended for use in conjunction with GO terms, and should not be considered in isolation from the terms.

At RGD manually assigned evidence codes are primarily in two categories: experimental and curatorial statement.

Experimental evidence codes:

Use of an experimental evidence code in a GO annotation indicates that the cited paper displayed results from a physical characterization of a gene/gene product that has supported the association of a GO term. The *experimental evidence codes* used at RGD are:

- Inferred from Direct Assay (IDA)
- Inferred from Physical Interaction (IPI)
- Inferred from Mutant Phenotype (IMP)
- Inferred from Genetic Interaction (IGI)
- Inferred from Expression Pattern (IEP)

IDA: Inferred from Direct Assay

Last updated: 2017

- Enzyme assays
- In vitro reconstitution (e.g. transcription)
- Immunofluorescence (for cellular component)
- Cell fractionation (for cellular component)

- Physical interaction/binding assay (sometimes appropriate for cellular component or molecular function)

The IDA evidence code is used to indicate a direct assay was carried out to determine the function, process, or component indicated by the GO term. Curators therefore need to be careful, because an experiment considered as a direct assay for a term from one ontology may be a different kind of evidence for a term from another of the ontologies. In particular, there are more kinds of direct assays for cellular component than for function or process. For example, a fractionation experiment might provide "direct assay" evidence that a gene product is in the nucleus, but "protein interaction" (IPI) evidence for its function or process.

For transfection experiments or other experiments where a gene from one organism or tissue is put into a system that is not its normal environment, the annotator should use the author's intent and interpretation of the experiment as a guide as to whether IMP or IDA is appropriate. When the author is comparing differences between alleles, regardless of the simplicity or complexity of the assay, IMP is appropriate. When the author is using an expression system as a way to investigate the normal function of a gene product, IDA is appropriate.

Examples where the IDA evidence code should be used:

- Binding assays can provide direct assay evidence for annotating to the *xxx* binding molecular function terms. (**Note**: Use IDA only when no identifier can be placed in the with/from column; when there is an appropriate ID for the with/from column, use IPI).
- Assays describing the isolation of a complex by immunoprecipitation of a tagged subunit should use IDA, not IPI. Thus this type of assay can provide IDA for annotation to a component term for the specific complex because it is a direct assay for a complex.
- Transfections into a cell line, overexpression, or ectopic expression of a gene when the expression system used is considered to be an assay system to address basic, normal functions of gene product even if it would not normally be expressed in that cell type or location. If the experiments were conducted to assess the normal function of the gene and the assay system is believed to reproduce this function, i.e., the authors would consider their experiment to be a direct assay, and not a comparison between various alleles of a gene, then the IDA code should be used. This is in contrast with a situation where overexpression affects the function or expression of the gene and that <u>difference from normal is used to make an inference about the normal function</u>; in this case use the <u>IMP</u> evidence code.

Examples where the IDA evidence code should <u>not</u> be used:

- Binding assays where it is possible to put an ID corresponding to the specific binding partner that was shown to interact directly the gene product being annotated should be annotated with the IPI code, not with IDA.

- Transfection into a cell line, overexpression, or ectopic expression of a gene where the effects of various alleles of a gene are compared to each other or to wild-type. For this type of experiment, annotate using IMP.

RGD Curation Samples:

#1 Gene: Sult1e1, Curated GO term: "estrone sulfotransferase activity" (Molecular Function)**,** Evidence code: IDA, Reference - RGD:1299062, PMID:7857871

> **Abstract:** A new isoform of rat liver estrogen sulfotransferase (EST), rEST-6, which is distinct from the previously reported rat EST [Demyan et al., Molec. Endocrinol. 6 (1992) 589], has been cloned, expressed, purified and characterized. A PCR procedure using oligonucleotide primers synthesized to the 5'-nontranslated and 3'-nontranslated regions of the published rEST sequence was used to isolate rEST-6 cDNA. The cloned DNA is 1000 bp in length and encodes a protein of 295 amino acids with a calculated molecular mass of 35,300 Da. rEST-6 is selectively expressed in male rats, as confirmed by Northern blot and immunoblot analyses. Northern blot analysis of male and female rat liver RNA with the rEST-6 cDNA as a probe shows a band with male RNA but not with female RNA. Similarly, immunoblot analysis of male and female rat liver cytosols with an antibody to rat EST yields a strong immunoreactive band in rat liver cytosol from male rats but not from females. Subsequent to bacterial expression and purification of rEST-6, the enzyme was analyzed kinetically and shown to sulfate estrogens but not dehydroepiandrosterone, pregnenolone, cortisol or testosterone. Maximal sulfation activity towards both beta-estradiol and estrone occurred at a concentration of 1 microM with substrate inhibition at higher concentrations. These results indicate that multiple, closely related forms of EST are present in rat liver. Analysis of the activity and regulation of these different EST enzymes is important in understanding estrogen metabolism in rats.
>
> **All the information necessary for the annotation, except for positive identification of the gene, is found in the abstract. The full paper does not have a GenBank accession number, so the only way to positively identify "**rEST-6**" as "**Sult1e1**" is to check the GenBank refseq accession number from the RGD gene report for "**Sult1e1**". On the NCBI Nucleotide page for "**NM_012883**", the reference cited above (PMID:7857871) is given as a reference for the Sult1e1 mRNA sequence.** NOTE: rEST-6 has been added as a synonym for rat Sult1e1 in RGD.

#2 Gene: Pebp1, Curated GO term: "regulation of the force of heart contraction" (Biological Process)**,** Evidence code: IDA, Reference - RGD:2302820, PMID:16608915

Abstract: Hippocampal cholinergic neurostimulating peptide (HCNP), which derives from phosphatidylethanolamine-binding protein (also named Raf kinase inhibitor protein), enhances acetylcholine synthesis in the hippocampal medial septal nuclei. It is

present in the chromaffin secretory granules of the adrenal cells and under stress is cosecreted with peptide hormones and catecholamines. Using the isolated rat heart perfused according to Langendorff to reveal the cardiotropic action of HCNP on the mammalian heart, we showed that rat HCNP exerts, at concentrations of 5x10(-13) to 10(-6) M, a negative inotropism under basal conditions (left ventricular pressure variations ranging from -8.34+/-0.94% to -21+/-3.5%) and enhances the cholinergic-mediated negative inotropy through direct interaction with G-protein-coupled muscarinic receptor pathway. Under adrenergic stimulation (isoproterenol), the peptide exerts an antiadrenergic action. The analysis of the percentage of rate pressure product variations in terms of EC50 values of isoproterenol alone (-8.5+/-0.3; r2=0.90) and in the presence of rat HCNP at 0.01 nM (-6.9+/-0.36; r2=0.88) revealed a competitive type of antagonism of the peptide. HCNP does not affect either heart rate or coronary pressure. The evidence that HCNP in mammals may play a novel role as an inhibitory cardiac modulator throughout an involvement of the myocardial G-protein-coupled receptor pathway provides new insights regarding the neurohumoral control of heart function under normal and physiopathological conditions.

All the information necessary for the annotation is found in the abstract.

#3 Gene: Pam, Curated GO term: "extracellular space" (Cellular Component**),** Evidence code: IDA, MMO:0000429 (blood enzyme activity assay) in Notes text box, Reference - RGD:2302418, PMID:16448835

Title: Plasma peptidylglycine alpha-amidating monooxygenase (PAM) and ceruloplasmin are affected by age and copper status in rats and mice.

Abstract: In an attempt to identify a sensitive and improved marker of mammalian copper status during neonatal development experiments compared two plasma cuproenzymes, peptidylglycine alpha-amidating monooxygenase (PAM), an enzyme involved in peptide posttranslational activation, to ceruloplasmin (Cp), a ferroxidase involved in iron mobilization. Dietary Cu deficiency (Cu-) was studied in dams and offspring at postnatal age 3 (P3), P12, and P28. Rodent Cp activity rose during lactation whereas PAM activity fell. Reduction in Cp activity was more severe than reduction in PAM activity in Cu- offspring and dams. Cp activity was greater in rats than mice whereas PAM activity was similar in adults but greater in mouse than rat pups. Both cuproenzymes changed during neonatal development and when dietary copper was limiting. With proper controls, each enzyme can be used to assess copper status.

The abstract mentions the species, the protein name, and enzyme activity. The possible GO terms are the specific enzyme activity, lactation, aging, and the cellular location of that activity. The title and abstract indicate that the "cellular component" would be "extracellular space", because they state that Pam was in

the plasma. The MMO ID is provided for the display of expression data at the Alliance of Genome Resources (https://www.alliancegenome.org/). The other GO terms would need to be checked by viewing the full paper (see example #2 in the IEP section below).

IPI: Inferred from Physical Interaction

Last updated: 2017

- 2-hybrid interactions
- In vitro binding assay
- Co-purification
- Co-immunoprecipitation
- Ion/protein binding experiments

Covers physical interactions between the gene product of interest and another molecule (such as a protein, ion, or complex). IPI can be thought of as a type of IDA, where the actual binding partner or target can be specified, using the "With Info" field.

Examples where the IPI evidence code should be used:

- Binding assays where it is possible to put an ID corresponding to the specific binding partner that was shown to interact directly the gene product being annotated should be annotated with the IPI code, not with IDA.

Examples where the IPI evidence code should not be used:

- Assays describing the isolation of a complex by immunoprecipitation of a tagged subunit should use IDA, not IPI with the ID corresponding to the tagged subunit in the "With Info" field because not all subunits of the complex interact directly with the tagged subunit. Thus, this type of assay can provide IDA for annotation to a component term for the specific complex because it is a direct assay for a complex, but it would not necessarily be true to say that all members of the complex interact directly with the tagged subunit.
- Annotations to protein binding; GO:0005515 should not be used to describe an antibody binding to another protein. However, an effect of an antibody (used as a reagent) on an activity or process can support a function or process annotation, using the IMP code.

Usage of the "With Info" Field for IPI

Making an entry in the "With Info" field is mandatory when using this evidence code, and must include an identifier for the other protein or other macromolecule or other chemical involved in the interaction. When multiple entries are placed in the "With Info" field, they are separated by a pipe (ex:RGD:3567|RGD:16958) to mean "or", or a comma

(ex:RGD:3567, RGD:16958) to mean “and”. Use IDA when no identifier can be entered in the “With Info” field.

Two examples of how the “With Info” field is used with IPI are shown in the table below. Abcd3, a mouse gene, is annotated to protein binding; GO:0005515, based on Liu et al. 1999 (PMID:10551832). The “With Info” field has the UniProt protein ID of the protein Abcd3 binds to. Alb, a rat gene is annotated to drug binding based on Harada et al. 2002 (PMID:12458670). In this case the CHEBI ID (chemical ID) of the drug that Alb binds to is provided in the “With Info” field.

..	DB Object ID	DB Object Symbol	Qualifier	GO ID	DB:Reference	Evidence Code	With/From	..
..	MGI:1349216	Abcd3		GO:0005515	PMID:10551832	IPI	UniProt:P33897|UniProt:Q61285	..
..	RGD:RGD2085	Alb		GO:0008144	PMID:12458670	IPI	CHEBI:28939	..

Note: For an interacting protein, a protein ID is recommended in the “With Info” field for an IPI annotation, but a gene ID may be used if the database does not have identifiers for individual gene products (This is the case for RGD). A gene ID may also be used if the cited reference provides enough information to determine which gene ID should be used, but not enough to establish which protein ID is correct, for example, in cases where there is a one-to-many relationship between a gene and its protein products.

RGD Curation Samples:

#1 Gene: Xrcc1, Curated GO term: “protein binding” (Molecular Function)**,** Evidence code: IPI, With Info: RGD:3363, Reference - RGD:2302580, PMID:17412650

Abstract: Neuronal protection induced by ischemic preconditioning has an important role in the reduction of stroke volume and attenuation of neuronal cell death. Ischemic injury is associated with increased oxidative DNA damage, and failure to efficiently repair these oxidatively damaged lesions results in the accumulation of mutations and neuronal cell death. Although the effects of ischemic tolerance can have profound implications, the precise mechanisms mediating this phenomenon remain unclear. The base excision repair (BER) pathway has a major role in the repair of oxidative DNA base damage after ischemic injury. Using a rat model of ischemic preconditioning, we now report that the neuronal protection observed after induction of ischemic tolerance is associated with increased BER. In situ detection of single-strand breaks and apurinic/apyrimidinic sites reduced to baseline levels after reperfusion following ischemic preconditioning. By contrast, no change was seen in the quantity of in situ lesions after reperfusion in non-ischemic preconditioned brain. Induction of the BER

proteins XRCC1, DNA polymerase-beta, and DNA ligase III was seen after reperfusion in ischemically conditioned brain. Moreover, an increase in binding between XRCC1 and DNA polymerase-beta was seen under these conditions, as might be expected during formation of functional BER complexes. Using in vitro BER oligonucleotides, we directly demonstrated an increase in total BER capacity of nuclear extracts prepared from ischemic-conditioned brain after reperfusion compared with sham-operated brain. These findings provide direct evidence that increased BER is associated with the neuroprotection induced after ischemic preconditioning, and provides important new mechanistic insight into the important biologic pathways that protect neurons against irreversible ischemic injury.

The GO term "protein binding" might be applied to both Xrcc1 **and** Polb **(DNA polymerase-beta), based on immunoprecipitation experiments described in the full paper. However, because of a recent change in GOC (gene ontology consortium) policy, only direct evidence from in vitro binding of purified proteins or a two-hybrid screen is good enough for an IPI evidence code. If a term for binding to a specific type of protein was available for these two proteins, that term with IDA evidence code would be appropriate.**

Disease terms are also found in the abstract. Since "brain ischemia**" and "**reperfusion injury**" would fit neurological disease (a previously curated RGD disease portal), this reference should also be used for RDO (see IEP under the section of disease ontology curation samples).**

#2 Gene: Fabp1, Curated GO term: "fatty acid binding" (Molecular Function)**,** Evidence code: IPI, With Info: ChEBI:16196| ChEBI:15756, Reference - RGD:1582399, PMID:10666570

Title: Intestinal fatty acid binding protein may favor differential apical fatty acid binding in the intestine.

Abstract: The intestinal mucosa metabolizes fatty acids differently when presented to the lumenal or basolateral membrane. Expression of both liver and intestinal fatty acid binding proteins (L- and I-FABPs) uniquely in the enterocyte offers a possible explanation of this phenomenon. An organ explant system was used to analyze the relative binding of fatty acids to each protein. More fatty acid was bound to L-FABP than to I-FABPs (28% vs. 6% of cytosolic radioactivity), no matter on which side the fatty acid was added. However, a 2-3-fold increase in fatty acid binding to the intestinal paralog was noted after apical addition of palmitic or oleic acid in mucosa from chow fed rats. When oleic acid was added apically, a 1.4-fold increase in binding to I-FABP was observed in mucosa derived from chronically fat fed rats, consistent with the previously observed 50% increase in the content of that protein. Immunocytochemical localization of both FABPs in vivo demonstrated an apical cytoplasmic localization in the fasting state, and redistribution to the entire cytoplasm after fat feeding. These data are consistent with the hypothesis that I-FABP may contribute to the metabolic compartmentalization of apically presented fatty acids in the intestine

All the information necessary for the annotation is found in the abstract. To confirm the binding of palmitic and oleic acid to Fabp1, the full paper gives all the details. The ChEBI numbers for the two fatty acids can be gotten from the RGD term browser or the ChEBI database (http://www.ebi.ac.uk/chebi/).

IMP: Inferred from Mutant Phenotype

Last updated: 2017

- mutations, natural or introduced, that result in partial or complete impairment or alteration of the function of that gene
- polymorphism or allelic variation (including where no allele is designated wild-type or mutant)
- any procedure that disturbs the expression or function of the gene, including RNAi, anti-sense RNAs, antibody depletion, or the use of any molecule or experimental condition that may disturb or affect the normal functioning of the gene, including: inhibitors, blockers, modifiers, any type of antagonists, temperature jumps, changes in pH or ionic strength.
- overexpression or ectopic expression of wild-type or mutant gene that results in aberrant behavior of the system or aberrant expression where the resulting mutant phenotype is used to make a judgment about the normal activity of that gene product.

The IMP evidence code covers those cases when the function, process or cellular localization of a gene product is inferred based on differences in the function, process, or cellular localization between two different alleles of the corresponding gene. The IMP code is used for cases where one allele may be designated 'wild-type' and another as 'mutant'. It is also used in cases where allelic variation occurs naturally and no specific allele is designated as wild-type or mutant. Caution should be used when making annotations from gain-of-function mutations as it may be difficult to infer a gene's normal function from a gain of function mutation, although it is sometimes possible.

For transfection experiments or other experiments where a gene from one organism or tissue is put into a system that is not its normal environment, the annotator should use the author's intent and interpretation of the experiment as a guide as to whether IMP or IDA is appropriate. When the author is comparing differences between alleles, regardless of the simplicity or complexity of the assay, IMP is appropriate. When the author is using an expression system as a way to investigate the normal function of a gene product, IDA is appropriate.

Examples where the IMP code should be used

- use of an inhibitor of a gene product's activity in order to see the effect of absence, or significant depletion, of that gene product. For example, an experiment using baicalein to inhibit the activity of 12-LOX in a murine bladder cancer cell line inhibits cell proliferation in a concentration dependent manner (see PMID:15161019) results in an annotation to the GO term cell proliferation using the IMP evidence code for the 12-LOX gene.
- transfection into a cell line, overexpression, or ectopic expression of a gene where the effects of various alleles of a gene are compared to each other or to wild-type. For this type of experiment, annotate using IMP.

Examples where the IMP code should not be used

- mutation in gene B provides information about gene A being annotated. For this type of experiment, use the IGI code.
- complementation of a mutation in one organism by a gene from a different organism.
- Transfections into a cell line, overexpression, or ectopic expression of a gene when the expression system used is considered to be an assay system to address basic, normal functions of gene product even if it would not normally be expressed in that cell type or location. If the experiments were conducted to assess the normal function of the gene and the assay system is believed to reproduce this function, i.e., the authors would consider their experiment to be a direct assay, and not a comparison between various alleles of a gene, then the IDA code should be used. This is in contrast with a situation where overexpression affects the function or expression of the gene and that difference from normal is used to make an inference about the normal function; in this case use the IMP evidence code.

Usage of the “With Info” field for IMP

GOC recommends making an entry in the “With Info” field when using this evidence code to indicate the identifier for the allele in which the phenotype was observed. When multiple entries are placed in the “With Info” field, they are separated by pipes. Due to limited allele data, this will rarely be used in RGD.

Example for how the “With Info” field should be filled in

- The mouse gene product Actc1 (actin, alpha, cardiac ; MGI:87905), has a GO annotation to muscle thin filament assembly, GO:0030240; inferred from mutant phenotype, IMP of MGI:2180072 (symbol: Actc1tm1Jll; name: targeted mutation 1, James Lessard), from PMID:9114002. MGI:2180072 is entered in the with/from column for this annotation.

RGD Curation Samples:

#1 Gene: Rela, Curated GO terms: "positive regulation of cell proliferation", "positive regulation of chondrocyte differentiation" (Biological Process)**,** Evidence code: IMP, Reference - RGD:2298851, PMID:17884819

Abstract: NF-kappaB is a group of transcription factors involved in cell proliferation, differentiation, and apoptosis. Mice deficient in the NF-kappaB subunits p50 and p52 have retarded growth, suggesting that NF-kappaB is involved in bone growth. Yet, it is not clear whether the reduced bone growth of these mice depends on the lack of NF-kappaB activity in growth plate chondrocytes. Using cultured rat metatarsal bones and isolated growth plate chondrocytes, we studied the effects of two NF-kappaB inhibitors (pyrrolidine dithiocarbamate (PDTC) or BAY11-7082 (BAY)), p65 short interference RNA (siRNA), and of the overexpression of p65 on chondrocyte proliferation, differentiation, and apoptosis. To further define the underlying mechanisms, we studied the functional interaction between NF-kappaB p65 and BMP-2 in chondrocytes. PDTC and BAY suppressed metatarsal linear growth. Such growth inhibition resulted from decreased chondrocyte proliferation and differentiation and from increased chondrocyte apoptosis. In cultured chondrocytes, the inhibition of NF-kappaB p65 activation (by PDTC and BAY) and expression (by p65 siRNA) led to the same findings observed in cultured metatarsal bones. In contrast, overexpression of p65 in cultured chondrocytes induced chondrocyte proliferation and differentiation and prevented apoptosis. Although PDTC, BAY, and p65 siRNA reduced the expression of BMP-2 in cultured growth plate chondrocytes, the overexpression of p65 increased it. The addition of Noggin, a BMP-2 antagonist, neutralized the stimulatory effects of p65 on chondrocyte proliferation and differentiation, as well as its anti-apoptotic effect. In conclusion, our findings indicate that NF-kappaB p65 expressed in growth plate chondrocytes facilitates growth plate chondrogenesis and longitudinal bone growth by inducing BMP-2 expression and activity.

From the full paper:

Organ Culture—The second, third, and fourth metatarsal bone rudiments were isolated from Sprague-Dawley rat fetuses at 20 days post conception and cultured individually in 24-well plates (25, 26).

Chondrocyte Culture—The cartilaginous regions of the metatarsal rudiments were dissected, rinsed in phosphate-buffered saline, and then incubated in 0.2% trypsin for 1 h and 0.2% collagenase for 3 h.

Similarly, p65 siRNA-transfected chondrocytes exhibited reduced chondrocyte proliferation (Table 3) and differentiation (representative blot (Fig. 5*C* and supplemental Table 1)) when compared with chondrocytes transfected with control siRNA.

> **The abstract indicates that the bone culture material is rat, but the species of the chondrocyte culture is not specified. Since mouse is also mentioned in the abstract, the full paper must be examined to determine if the chondrocytes are rat in origin. The effect of the specific Rela inhibitor, p65 siRNA, is only mentioned in the abstract in regard to the chondrocytes. The full paper also confirms the role of Rela in proliferation and differentiation, based on inhibition by p65 siRNA.**
>
> In addition curators may find the evidence code decision tree (http://www.geneontology.org/page/evidence-code-decision-tree) a useful aid in selecting the correct evidence code for an annotation.

IGI: Inferred from Genetic Interaction

Last updated: 2017

- "Traditional" genetic interactions such as suppressors, synthetic lethals, etc.
- Functional complementation
- Rescue experiments
- Inference about one gene drawn from the phenotype of a mutation in a different gene

Includes any combination of alterations in the sequence (variant) or expression of more than one gene/gene product. This code can therefore cover any of the IMP experiments that are done in a non-wild-type background; the key is what the comparison is made against. If there is a single variant or difference between the two strains compared, use IMP. If there are multiple mutations or differences between the two strains compared, use IGI. When redundant copies of a gene must all be mutated to see an informative phenotype, use IGI. Caution should be used when making annotations from genetic combinations that include gain-of-function mutations as it may be difficult to infer a gene's normal function from a gain of function mutation, although it is sometimes possible. Note that some organisms, such as mouse, will have far more IGI than IMP annotations. Use IMP for "phenotypic similarity," as described above.

"Functional complementation" above refers to experiments in which a gene from one organism complements a deletion or other mutation in another species. For these annotations, the “With Info” field should list the identifier for the gene that is complemented by the introduced gene.

Examples where the IGI evidence code should be used:

- mutation in gene B provides information about gene A being annotated. For this type of experiment, use the IGI code.
- complementation of a mutation in one organism by a gene from a different organism. For this type of experiment, use the IGI code.

Examples where the IGI evidence code should not be used:

Usage of the “With Info” field for IGI

Making an entry in the “With Info” field is mandatory when using this evidence code. When multiple entries are placed in the “With Info” field, they are separated by pipes or commas.

Note that there has been some discrepancy between groups as to the use of the “With Info” field.

For example, if the annotation is based on a double mutation, one of which is the gene being annotated, the identifier for the gene being annotated would be placed in the DB_Object_ID field (column 2) while the identifier for the second mutated gene would be placed in the “With Info” field (column 8). If the annotation is based on a triple mutation, one of which is the gene being annotated, the identifiers for both of the two other mutated genes would be placed in the “With Info” field, separated by a pipe or comma. For use as an RGD example, disregard column numbers and pay attention to column headings and data found in those columns.

..	2. DB Object ID	3. DB Object Symbol	4. Qualifier	5. GO ID	6. DB:Reference	7. Evidence Code	8. With/From	..
..	FB:gene_A_ID	geneA		GO:0006796	PMID:11979277	IGI	FB:gene_B_ID	..
..	FB:gene_A_ID	geneA		GO:0006796	PMID:11979277	IGI	FB:gene_B_ID|FB:gene_C_ID	..
..	SGD:gene_A_ID	geneA		GO:0006796	PMID:11979277	IGI	PomBase:gene_B_ID	..

IEP: Inferred from Expression Pattern

Last updated: 2017

- Transcript levels or timing (e.g. Northerns, PCR)
- Protein levels (e.g. Western blots)

The IEP evidence code covers cases where the annotation is inferred from the timing or location of expression of a gene, particularly when comparing a gene that is not yet characterized with the timing or location of expression of genes known to be involved in a particular process. It may be difficult to determine whether the expression pattern really indicates that a gene plays a role in a given process, so the IEP evidence code is usually used in conjunction with high level GO terms in the biological process ontology. The minimal requirement is to have three time points of varying expression to annotate a gene product to a particular biological process.

Note that it is invalid to make annotations to terms from the cellular component or molecular function ontologies on the basis of expression pattern data. Use of this code is restricted to annotations to terms from the biological process ontology.

Examples where the IEP evidence code should be used:

- genes upregulated during a stress condition may be annotated to the process of stress response (for example, heat shock proteins)
- genes selectively expressed at specific developmental stages in specific organs may be annotated to *xxx* development

Examples where the IEP evidence code should not be used:

- Function and component annotations should not be made with IEP.
- Exogenous expression or overexpression of a gene should be not annotated using IEP; only the normal expression pattern should lead to an IEP annotation.
- Overexpression of a gene causing increased activity of an enzyme should be annotated to IDA or IMP (see IDA documentation above)
- Overexpression (wild type or mutated) of a gene causing an abnormal phenotype should be annotated to IMP
- Exogenous expression of a gene and assaying of its function should be annotated to IDA (like a transcription factor)

RGD Curation Samples:

#1 Gene: Pam, Curated GO terms: “lactation”, “response to copper ion” (Biological Process), Evidence code: IEP, Reference - RGD:2302418, PMID:16448835

Title: Plasma peptidylglycine alpha-amidating monooxygenase (PAM) and ceruloplasmin are affected by age and copper status in rats and mice.

Abstract: In an attempt to identify a sensitive and improved marker of mammalian copper status during neonatal development experiments compared two plasma cuproenzymes, peptidylglycine alpha-amidating monooxygenase (PAM), an enzyme involved in peptide posttranslational activation, to ceruloplasmin (Cp), a ferroxidase involved in iron mobilization. Dietary Cu deficiency (Cu-) was studied in dams and offspring at postnatal age 3 (P3), P12, and P28. Rodent Cp activity rose during lactation whereas PAM activity fell. Reduction in Cp activity was more severe than reduction in PAM activity in Cu- offspring and dams. Cp activity was greater in rats than mice whereas PAM activity was similar in adults but greater in mouse than rat pups. Both cuproenzymes changed during neonatal development and when dietary copper was limiting. With proper controls, each enzyme can be used to assess copper status.

The abstract mentions the species, the protein name, and enzyme activity. The possible GO terms are the specific enzyme activity, lactation, aging, and the cellular location of that activity. The title and abstract indicate that the “cellular component” would be “extracellular space”, because they state that Pam was in the plasma. “lactation” and “response to copper ion” can be associated with Pam by reading in the full paper (in this case only the figures and legends are freely accessed online) that expression of Pam was altered (implied by activity measured) during lactation and changes of copper in the diet.

Curatorial statement evidence code:

Use of the curatorial statement evidence code indicates an annotation made on the basis of a curatorial judgement that does not fit into one of the other evidence code classifications. The *curatorial statement codes* are:

- No biological Data available (ND) evidence code.

ND: No biological Data available

Last updated: 2017

Used for annotations when information about the molecular function, biological process, or cellular component of the gene or gene product being annotated is not available.

Use of the ND evidence code indicates that the curator found no information that allowed making an annotation to any term indicating specific knowledge from the ontology in question (molecular function, biological process, or cellular component) as of the date indicated. This code should be used only for annotations to the root terms, molecular function; GO:0003674, biological process; GO:0008150, or cellular component ; GO:0005575, which, when used in annotations, indicate that no knowledge is available about a gene product in that aspect of GO.

Annotations made with the ND evidence code should be accompanied by a reference (RGD:1598407) that explains that curators looked but found no information.

Note that use of the ND evidence code with an annotation to one of the root nodes to indicate lack of knowledge in that aspect makes a statement about the lack of knowledge **only** with respect to that particular aspect of the ontology. Use of the ND evidence code to indicate lack of knowledge in one particular aspect does **not** make any statement about the availability of knowledge or evidence in the other GO aspects.

Note: The ND evidence code, unlike other evidence codes, should be considered as a code that indicates curation status/progress than as method used to derive an annotation.

RGD Curation Samples:

#1 Gene: Stambp, Curated GO terms: “molecular_function” (Molecular Function), “biological_process” (Biological Process), “cellular_component” (Cellular Component), Evidence code: ND, Reference - RGD:1598407, free text note: “02/2009: no relevant rat data”.

Abstract: “Gene literature reviewed as part of ongoing gene curation. Every effort was made to find and review all existing literature for this gene as of the date of review.”

The ND evidence code is being used to indicate that the literature for this particular gene has been reviewed, but no relevant rat data was found as of the date in the free text note. The ND evidence code may be used for one or more of the three aspects of the Gene Ontology for any particular gene.

Computational analysis evidence codes:

Use of the computational analysis evidence codes indicates that the annotation is based on an *in silico* analysis of the gene sequence and/or other data as described in the cited reference. The evidence codes in this category also indicate a varying degree of curatorial input. The *computational analysis evidence codes* used at RGD are:

- Inferred from Sequence Orthology (ISO)

ISO: Inferred from Sequence Orthology

Last updated: 2017

Examples of when to use ISO:

RGD does not manually assign the ISO evidence code to Gene Ontology annotations. The only GO ISO annotations at RGD are imported from external sources. See the Disease Ontology and Pathway Ontology sections for examples of when to manually assign ISO.

2. Pathway Ontology Curation

The evidence codes used for pathway ontology are the same as for gene ontology. Examples of their use would be similar to those shown above in the previous section.

3. RGD Disease Ontology Curation

At RGD Disease Ontology curation uses a set of evidence codes that partially overlaps with those used for Gene Ontology curation.

Manual Evidence codes:

- Inferred by Association of Genotype from Phenotype (IAGP)
- Inferred from Experimental Data (IDA)
- Inferred from Expression Pattern (IEP)
- Inferred from Mutant Phenotype (IMP)
- Inferred from Sequence Orthology (ISO)

IAGP: Inferred by Association of Genotype from Phenotype

- polymorphism or segregation of genetic markers (SNPs, mutations, RFLPs, microsatellites)
- polymorphism or segregation of physical markers (FISH, centromeric, heterochromatic regions, chromosomal banding patterns)
- detection of polymorphisms in inbred stock

The IAGP evidence code is used to indicate a genetic variation of a particular gene is associated with a certain disease as indicated by the Disease Ontology term. This evidence code is most commonly used for polymorphisms or mutations.

RGD Curation Samples:

#1 Gene: Nmu, Curated RDO term: "Obesity", Evidence code: IAGP, Reference - RGD:1642094, PMID:16984985

Title: Association between neuromedin U gene variants and overweight and obesity.

Abstract: Neuromedin U (NMU) is an anorexic neuropeptide expressed in the hypothalamus. Mice lacking the NmU gene are hyperphagic and obese, whereas mice overexpressing Nmu are hypophagic and lean. OBJECTIVE: Our objective was to investigate whether variants in NMU are associated with human obesity. DESIGN: The coding region of NMU was analyzed for variants in obese Czech children and obese Danish adults. Identified missense variants were investigated for cosegregation with obesity in families or association with obesity in the general population. SETTING: The study was performed at Steno Diabetes Center, Denmark, and Department of Pediatrics, Charles University, Czech Republic. SUBJECTS AND METHODS: A total of 289 Czech children and adolescents with early-onset obesity and 84 Danish obese adults were analyzed for variants in NMU. A NMU Ala19Glu polymorphism was genotyped in 5851 Danish subjects of the Inter99 cohort, and a rare NMU Arg165Trp mutation was sequenced in the proband family and in 53 lean and unrelated Czech subjects. RESULTS: The rare NMU Arg165Trp variant cosegregated with childhood obesity in a Czech family. Homozygous carriers of the Glu allele of the NMU Ala19Glu polymorphism were more common in the overweight and obese subjects; the Glu/Glu frequency was 0.4 (95% confidence interval, 0.2-0.6) among 2586 lean subjects (BMI < 25 kg/m2) and 0.9 (95% confidence interval, 0.7-1.1) among 3265 overweight and obese subjects (body mass index >or= 25 kg/m2) [odds ratio, 2.5 (1.2-5.3); P = 0.01]. CONCLUSION: Amino acid variants in NMU associate with overweight and obesity, suggesting that NMU is involved in energy regulation in humans.

The title indicates the gene and a specific disease term. The abstract indicates that the coding region of the gene is being analyzed. The abstract indicates that a mutation is associated with childhood obesity in a family study and a different polymorphism is associated with overweight and obese subjects compared to lean subjects. The abstract has all the information needed for a basic annotation, plus enough information to make an optional free text note: "DNA:missense mutation, polymorphism:cds:p.R165W, p.A19E".

IDA: Inferred from Experimental Data

- Enzyme assays
- Physical interaction/binding assay

The IDA evidence code is used to indicate a direct assay was carried out to determine participation of a particular gene product in the molecular mechanism of a disease. It is the disease annotation equivalent of IDA. If the participation is based on experiments using specific inhibitors, drugs, or exogenous expression of a gene product use the IMP evidence code.

IEP: Inferred from Expression Pattern

- cases where the annotation is inferred from the timing or location of expression of a gene
- Transcript levels (e.g. Northerns, PCR)
- Protein levels (e.g. Western blots)

The IEP evidence code covers cases where the annotation is inferred from the timing or location of expression of a gene. Use this code to associate a gene product as a biomarker for a particular disease.

RGD Curation Samples:

#1 Gene: Mapk9, Curated RDO term: "Precancerous Conditions", Evidence code: IEP, Reference - RGD:2304231, PMID:18081878

Title: Development of glutathione S-transferase-P-negative foci accompanying nuclear factor-erythroid 2-related factor 2 expression during early stage of rat hepatocarcinogenesis.

Abstract: Glutathione S-transferase P (GST-P), a marker for rat hepatic preneoplastic lesions, is suggested to bind to Jun N-terminal kinase (JNK) to repress stress response, and GST-P gene expression is regulated by a transcription factor, nuclear factor-erythroid 2-related factor 2 (Nrf2). In this study, we examined by immunohistochemistry whether JNK2, p38 mitogen-activated protein kinase, and Nrf2 were expressed in GST-P-positive foci induced by the Solt-Farber protocol. At 2 weeks after partial hepatectomy, all GST-P-positive foci were negative for p38, and 86.4 +/- 5.6% and 64.7 +/- 6.3% of GST-P-positive foci were negative for JNK2 and Nrf2, respectively. Western blot analysis showed decreased p38 mitogen-activated protein kinase and JNK2 expression in livers treated with the protocol. In immunohistochemistry, besides GST-P-positive foci, GST-P-negative foci were detected as p38-negative foci in the surrounding tissues positive for p38. In contrast to GST-P-positive foci, most GST-P-negative foci showed enhanced Nrf2 expression. The number of GST-P-negative foci was 76 +/- 18/10 mm (2) of liver section at 2 weeks, but was undetectable at 1 week. The area of GST-P-negative foci was 0.09 +/- 0.05 mm (2), smaller than that of GST-P-positive ones (0.29 +/- 0.23). After treatment with carbon tetrachloride, small vacuoles due to liver injury were frequently observed inside GST-P-negative foci but less frequently in GST-P-positive foci. However, this treatment resulted in expression of JNK2, p38, and Nrf2 in both foci. These results showed development of GST-P-negative foci during the early stage of hepatocarcinogenesis and suggested that Nrf2 is not responsible for GST-P expression in rat hepatic preneoplastic foci.

The title indicates the species and a general disease term. The abstract indicates that preneoplastic lesions (Precancerous Conditions) are what is really being studied. The abstract gives the gene name(s), what is being measured (protein level), and that the tissue being studied is liver. The abstract has almost all the information needed for a basic annotation, plus enough information to make an optional free text note: "associated with Liver Neoplasms, Experimental;protein:decreased expression:liver". The full paper shows that the Mapk9 data is from a Western blot with no quantitation. Although it is preferable to have significance data (P < 0.05 or better), it is clear from the blot that scanning data would yield a significant difference.

IMP: Inferred from Mutant Phenotype

any procedure that disturbs the expression or function of the gene, including RNAi, anti-sense RNAs, antibody depletion, or the use of any molecule or experimental condition that may disturb or affect the normal functioning of the gene, including: inhibitors,

blockers, modifiers, any type of antagonists, temperature jumps, changes in pH or ionic strength.

- overexpression or ectopic expression of wild-type or mutant gene that results in aberrant behavior of the system or aberrant expression where the resulting mutant phenotype is used to make a judgment about the disease association of that gene product.

The IMP evidence code covers those cases when the function, process or cellular localization of a gene product is intentionally altered during experimental conditions. For transfection experiments or other experiments where a gene from one organism or tissue is put into a system that is not its normal environment, the curator should use the source of the gene or gene product to assign the annotation to a species.

RGD Curation Samples:

#1 Gene: Ccl22, Curated RDO term: "Anti-Glomerular Basement Membrane Disease", Evidence code: IMP, Qualifier: treatment, Reference - RGD:2306306, PMID: 12651599

Title: Mononuclear cell-infiltrate inhibition by blocking macrophage-derived chemokine results in attenuation of developing crescentic glomerulonephritis.

Abstract: Glomerular monocyte/macrophage (Mo/M phi) infiltrates play a role in many forms of glomerulonephritis (GN), and the intensity of Mo/M phi trafficking correlates with the loss of renal function and histological damage. We analyzed the functional role of macrophage-derived chemokine (MDC), a potent mononuclear cell chemoattractant, during the progression of anti-glomerular basement membrane (GBM) antibody (Ab) GN, a model of crescentic GN in the WKY rat, and whether the effects of MDC were dependent on its receptor CCR4. MDC mRNA and protein expression were markedly induced in nephritic glomeruli throughout the disease. Blocking the function of MDC did not affect the developing of the disease from days 2 to 7, but it dramatically blocked M omicron/M phi infiltration in the glomeruli, prevented crescent formation, and reversed renal function impairment during days 7 to 14 of the anti-GBM GN. In this study, we also found that MDC activity on M omicron/M phi in this GN was at least partly dependent on a new variant of CCR4. These results suggest that MDC is critically involved in the development of anti-GBM GN from acute glomerular injury to irreversible tissue damage. In addition, an antagonist to MDC may represent a prime drug target for therapeutic application to intervene in the progression of anti-GBM GN and in other M omicron/M phi-dominant GN.

The title indicates a curatable disease term. The abstract indicates the species, gene product involved, a more defined disease term, and two different types of experiments to indicate two different evidence codes for the annotation. The change in expression during the disease process suggests an IEP evidence code, but the

blocking of the gene product causing an improvement in the disease suggests the more preferred IMP evidence code and "treatment" as qualifier.

ISO: Inferred from Sequence Orthology

- sequence similarity
- recognized domains
- structural similarity

The ISO evidence code is used for the remaining two orthologs when making a disease annotation to the primary species of a reference. A disease annotation is always curated across all three species. No matter what evidence code is used for the annotation of the primary species, the other two species' orthologs will be curated with the ISO evidence code. This is currently done automatically by the curation tool.

#1 Genes: mouse Mapk9 and human MAPK9, Curated RDO term: "Precancerous Conditions", Evidence code: ISO, Reference - RGD:2304231, PMID:18081878

See example #1 for IEP

4. Phenotype Ontology

The evidence codes used for mammalian phenotype ontology and human phenotype ontology are the same as for disease ontology. Examples of their use would be similar to those shown above in the previous section.