IMEx Curation Rules

01/05/2015

These rules are not designed to replace the curation manuals of the member databases, but lay down the minimum requirements to which an IMEx record should be curated to.

Initiating Curation

Experimental evidence curation will only be curated when a complex or set of interacting molecules has at least been partially purified – whole cell, cytoplasmic or membrane extracts will not be regarded as a purification step.

Where the manual refers to high-throughput data, this should be taken as a single experiment describing >100 interactions (binary data) or interactors (n-ary data).

Publication

1. All papers released as an IMEx record should have been peer reviewed and journal published.

2. All publications must have a PMID cross-reference, or a DOI if the journal is not indexed in PubMed. Pre-publications will not be exchanged by IMEx.

3. All papers should be curated in full i.e. all protein interactions (protein-protein) described within the document which can be described in terms of a recognisable protein sequence should be curated and present in the record, even if some are outside of the area of interest of a topical database. Positive controls should be curated, but should be indicated as such. When datasets are annotated with author confidence values or judgements, only high confidence data will be exported to IMEx. Any data which the author indicates as contaminants or artefacts will also not be included in an IMEx dataset.

4. Databases have a duty to fully report the information in a paper or submitted to the database. When they are in any doubt as to the validity of that data, the doubt should be recorded as a ’Caution’ comment.

5. Databases may also have curated protein:nucleic acid and protein:small molecule records . these are currently outside of the remit of IMEx and should be removed from the record prior to export.

Publication annotation

Additional Annotation - this need not be added by all participating databases, but when added, should use the following PSI-MI CV terms

1. Author-list (MI:0636) - a list of all authors on the paper.

2. Journal (MI:0885) - Name and details of the journal from which paper has been taken.

3. Publication year (MI:0886) - year of publication of the paper

4. Author submitted (MI:0878) - indicates data was directly submitted prior to publication. IntAct format: "2004-09-10: JS Choudhary, The Wellcome Trust Sanger Institute, Hinxton, UK."

5. Contact e-mail (MI:0634) - E-mail address given on the publication as contact details of the author or organisation which has produced the data.

6. Curation request (MI:0873) - Annotation of a published paper which has been externally requested. "2004-09-10: JS Choudhary, The WellcomeTrust Sanger Institute, Hinxton, UK.

7. Caution (MI:0618) - this is used for warning about possible errors in experiments or for specifying grounds for confusion in experiments where the author has expressed misgivings about a technique while comparing with another described in the same or different paper.

8. Comment (MI:0612) - this annotation topic can be used to describe additional information which cannot fit under other annotation topics. It is desirable that the comments are restricted to as few as possible and are complete sentences.

9. Imex curation (MI:0959) - indicates that the publication has been curated to IMEx standards i.e. full coverage of all protein-protein interactions.

10. Copyright - Individual experiments or interactions might have specific copyright statements attached to them. A copyright statement on experiment level applies to all interactions which are part of the experiment. Copyright statements attached to individual interactions override the statements inherited from the experiment.

Experiment

1. Names and short names need not be added, databases may add these according to their own rules on import.

2. All experiments must have an Interaction Detection method, at least one Participant Detection Method and at least one participant with both an Experimental and Biological role. Host organism should also be added.

Host organism

This is the organism in which the interaction took place, for example a yeast two-hybrid experiment run using human proteins would have the host organism 'yeast'. It is not necessarily the same location as where the proteins were expressed. A GST-tagged protein may be expressed in vitro, bound to a column and then exposed to HeLa cell lysate. The host organism would be □'in vitro' as this is where the interaction occurs. When an interaction occurs extracellularly, the host organism is said to be in vitro. This does not apply to the periplasmic space, which is a recognised subcellular compartment. For example, a protein binding to an extracellular receptor should also be described as □'in vitro' even if the receptor is part of an intact cell to indicate that the interaction is extracellular. A crystal is always deemed to be formed in vitro, even when homodimeric.

Organism

1. The organism should be cross referenced to the appropriate NCBI taxID, with the addition of -1 to indicate 'in vitro' or -2 to indicate 'chemical synthesis'

2. If the organism is unknown, the interaction is not normally captured.

3. The UniProt organism identifier code provides a 5 letter code by which each organism may also be differentiated.

Tissue - two possibilities exist, the BRENDA tissue list and the UniProt tissue list (www.expasy.org/cgi-bin/lists?tisslist.txt). The UniProt list has accession numbers and is cross-referenced to eVOC but is itself non-hierarchical. BRENDA is hierarchical. The use of BRENDA is strongly recommended.

Cell lines - although a number of resources exist, none provide a hierarchy of tissue□¨cell type□¨derived cell lines. Suggested resources

1. CABRI (Common Access to Biological Resources and Information (www.cabri.org)

2. The Cell type onotology (cell.obo) maintained on the OBO website (obo.sourceforge.net/cgi-bin/table.cgi)

3. If the cell line cannot be traced in an appropriate public domain resource, an PMID to an appropriate reference should be added where possible.

Interaction Detection Method

The experimental method should be annotated, or mapped, to the most specific term available in the ’interaction detection method’ PSI-MI CV.Should a specific term not be available, the curator may annotate to the most appropriate parent, and request a new term on the tracker system.

Additional experimental detail, for example variations from standard protocols, should be added as an annotation note to the attribute list.

Where a sequential process is described, annotators should either use a common parent, or annotate the experimental method which includes a participant detection step.

When a multi-step purification procedure takes place, and a figure is shown for participant ID at each stage, the whole process should be curated in separate steps. If participants are only identified at the end of the process, the methodology should be collapsed to a common parent. Where a "double pulldown" is performed Protein A-His, Protein B-Flag e.g. anti-His column, followed by anti-Flag column, followed by participant determination should be entered twice bait/prey, prey/bait. If the sequential techniques cannot be collapsed into a common term in the hierarchy both/all stages of the purification should be entered, unless a particular technique is particularly significant to the purification. Cross-linking will almost always be taken as inferring any subsequent isolation technique, which will not be separately captured.

Ligand binding experiments are not normally accepted as evidence of an interaction, except in cases where the curator is satisfied that the author has proven that the ligand can only be binding to the protein(s) in question e.g. ligand shown not to bind when the receptor is not transfected in, western blots performed to show that the receptor has been successfully transfected in.

Functional assays e.g. patch-clamp, reporter gene assays are not accepted as evidence of an interaction.

Participant Detection Method

The experimental method by which the prey protein(s) have been identified should be annotated, or mapped, to the most specific term available in the ’interaction detection method’ PSI-MI CV. Should a specific term not be available, the curator may annotate to the most appropriate parent, and request a new term on the tracker system.

Use of ’Predetermined’

Predetermined (or child terms) will be used whenever a group has introduced a known protein (usually as a clone) into a system and then utilised the fact that they know to already be present in the system as a basis for identification - looking for a protein of the correct molecular weight would be a good example. If they have used a more specific method to confirm its presence, this should be the Participant Detection method. For example, two hybrid - if the experiment is a screen against a cDNA library, the participant detection method is always ’nucleotide sequence’. If it is a directed screen, using a matrix of known proteins, the participant detection method is ’predetermined’ unless the author has specifically stated that the molecules were subsequently resequenced. Similarly, if only one method can be used by the database, participant Identification of tagged-molecules should be by the method identifying their presence in the experiment e.g. ‘western blot’, not ‘predetermined’ even though the molecules have been cloned and sequenced as part of the tagging process. However, if it is specifically stated that the identity of the molecule has been reconfirmed, for example by western blot, the reconfirmation method should then be used in Participant Determination.

If a publication has participants identified by 'molecular weight estimation by staining' or a child term and a further confirmation of the participant is obtained by an alternative method such as MS or Western this data can be curated.

If in vitro purified proteins used and participants identification using 'molecular weight estimation by staining' or child terms this data can be curated.

However, if a PMID describes interactions based on participant identification by 'molecular weight estimation by staining' and child terms but no confirmation of the identity of the participant is available in this publication, this will not be curated. This includes purified complexes. A reference to a previous paper is not acceptable.

Secondary method for Identifying Participants

Some papers describe participant identification by a mixture of methods, e.g. one or two protein identified by Western Blot, remainder by mass spec. If a database cannot describe more than one participant detection method, the most commonly used, or most informative should be used, and other(s) should be described as an annotation note, until Participant becomes an annotated object in each database.

Experiment Annotation

Annotations on the Experiment relate to the experimental conditions only. Each topic is free text for additional information but associated with a PSI-MI CV term. This need not be added by all participating databases, but when added, should use the following PSI-MI CV terms

1. Antibodies (MI:0671) - this topic may be used to detail any key information about one or more antibodies which may be relevant to the experiment. It is not used to list manufacturer's details on standard commercially available antibodies.

2. Author-confidence (MI:0621) - if a confidence value is given on either the interaction or the participant, it must be accompanied by a statement using this topic, to explain how the value was derived.

3. Caution (MI:0618) - this is used for warning about possible errors in experiments or for specifying grounds for confusion in experiments where the author has expressed misgivings about a technique while comparing with another described in the same or different paper.

4. Comment (MI:0612) - this annotation topic can be used to describe additional information which cannot fit under other annotation topics. It is desirable that the comments are restricted to as few as possible and are complete sentences.

5. Data processing (MI:0633) - this annotation topic is used to describe the steps in processing of data to obtain the identifiers described in the entry.

In the case of large-scale data, (more than 100 binary interactions attached to a single experiment or more than 100 interactors in an interaction) information about the original number of interactions described by the authors in the paper and the corresponding number in the database, if different from the one published in the paper, should also be stored here.

6. Dataset (MI:0875) - this annotation topic is used to link various publications pertaining to a topic of interest

7. Exp-modification (MI:0627) - used to describe the experimental method used by the authors when the PSI-MI CV term definition does not fully or adequately describe the technique in use.

8. Library used (MI:0672) - added to experiments such as Y2H or phage display when a library is screened. The information about the library which was scanned to obtain the interacting protein clone is recorded under this annotation topic.

9. URL (MI:0614) - URL/Web address describing an experiment.

Interaction

1. Each separate interaction pertaining to an experiment should be curated to as much detail as possible.

2. A name or short name need not be given. These may be added by member databases on import, if required.

3. IMEx does not exchange negative interactions so these should be removed from the entry prior to export.

4. Interactions should be shown as binary or n-ary as per the raw experimental data, not as subsequently expanded sets even if this has been performed by the author. Submitters of expanded data should be requested to replace this with the experimental data. In the case of western blots, if it is not clear if the author has stripped and reprobed the same blot with multiple antibodies, or repeated the coimmunoprecipitation multiple times and checked each with single antibody, the experiment should be represented once as a one bait, many prey interaction.

5. Figure legends should be added to interactions.

Interaction Type

The interaction type should be annotated, or mapped, to the most specific term available in the einteraction detection methodf PSI-MI CV. Should a specific term not be available, the curator may annotate to the most appropriate parent, and request a new term on the tracker system.

Direct interaction

An experiment will only be deemed to show a 'Direct interaction' if the number of interactors equals 2 highly purified molecules and the interaction occurs in vitro, such that no host proteins may interfere.

Association

Any interaction with the Interaction Detection Method = affinity chromatography or one of its children (e.g. coimmunoprecipitation, pulldown, TAP) with 1 bait and >1 prey should be mapped to association.

If the number of prey is only equal to 1, it should be mapped to Physical Association.

Colocalisation

Colocalisations should have interaction type = colocalisation, not physical. This is used for most imaging techniques (depends on degree of resolution) and potentially also cosedimentation. Electron density methods (e.g. gold) will be captured when there is a clear distinction between the labelling particles, usually on the basis of size. When Interaction Detection Method is a child of cosedimentation, curators will have to use judgement as to which interaction type is appropriate. Colocalisations where one of the proteins is only used because it is an indicator of a particular subcellular location, are not captured.

Confidence

Essential text, if given in the paper. This is the author described measure of confidence, NOT any value subsequently assigned by the database. The author's description of how confidence was derived and then described should be added as a series of annotation notes describing

a. Unit - confidence value as given by the author or by database

b. Value - value as defined by author, for example High, Low, numerical.

Participants

1. All proteins should either be present in a recognisable database or have a sequence and species of origin (if the latter, curators are encouraged to submit to the UniProtKB so that it can be added to the protein sequence databases), or be internally created as protein entities internally by the source database. Chemically synthesised peptides can be described as 'in vitro'. Interactions involving natural chimeric proteins e.g. bcr-abl should be annotated. Databases may choose to additionally annotate artefactual chimeric molecules when sequence is available or can be reconstructed from the paper, but in most cases, this data will not be curated.
2. When a protein is known to have isoforms due to alternative splicing, initiation sites or promoter usage, but it is not clear from the article which of the isoforms is making the interaction, then annotation will be to the UniProtKB canonical sequence and this will be mapped by sequence match via PICR to the corresponding Ref-Seq sequence. When the Isoform can be identified, the annotation will be made to the correct UniProtKB isoform(s) and an exact match can be made to the corresponding Ref-Seq sequence. Proteins will normally be mapped to the well annotated UniProtKB/Swiss-Prot sequence, or to the longest UniProtKB/TrEMBL (and corresponding RefSeq) transcript available in the database when no UniProtKB/Swiss-Prot protein is available.

Where it is not immediately apparent from which species the protein has originated, the following steps can be taken

a. Reference chasing.

b. Curator writes to the author and receives additional information.

c. Curator takes sequence information from paper e.g. protein length, amino acid positions of mutants and unequivocally maps to a single entry in the protein sequence database.

d. Curator looks at previous work of author (or group donating clone) and only one species has previously been used.

In all cases, the decision tree used to assign a database accession number to the protein should be documented as an annotation note.

When interactors can be mapped to a specific strain, we will map to that strain. If this information is not available, or there is a discrepancy in the information available such as between publication and database submision, databases will make every attempt to resolve this in collaboration with the author. If the issue cannot be resolved, the strain cannot be traced, or the discrepancy cannot be resolved, IMEx members will map interactions to the UniProtKB reference strain for that species.

3. All databases create proteins when a sequence is not available in the protein sequence databases. The sequence should be given, when known and it is the originating databases responsibility to update the entry and notify other IMEx members should further information become available.

4. If peptides can be mapped to a single protein, this should be annotated as a binding site feature of that protein. When there is an ambiguity, the species shall be deemed to be the same as that of the protein with which it is interacting.

Where there is an ambiguity within a species i.e. the identical peptide is conserved across several proteins, the interactor is generally not captured and the absence of an interactor is documented.

5. Expression details should be added using, or mapping to, PSI-MI CV terms wherever possible.

6. Author-derived name should be added when possible.

7. Stoichiometry should be added, where measured or derived, for example in crystallography or density gradient centrifugation. In given with a decimal float, round up/down to nearest whole integer.

   If an interaction describes a homodimer (or homo-nmer) using a technique in which a single construct is used, the construct should be added once with the appropriate stoichiometry. If the stoichiometry is unknown, i.e. the molecule is a homo-oligomer, the molecule should be added twice, stoichiometry=0. Additionally, the GO cross-reference 'protein homooligomerization' GO:0051260 should be added at the interaction level.

   If 2 regions (each purified separately) of the same protein are interacting but it is unclear if this is an intra- or inter-molecular interaction, the molecule should be entered twice, with stoichiometry=0. The Interaction type should be self interaction/putative self-interaction, as appropriate. The corresponding exp role will be self/putative self. The interaction types self interaction/putative self-interaction should NOT be used for autocatalysis, when the additional biological role self/putative self will supply this information.

8. Strain/sub-species (virus, bacteria, parasites, plants....) - when the exact strain is given and the specific protein is available in UniProt/Ref-Seq interaction should be mapped to this protein. If the exact strain is not available, interactions should be mapped to a protein in the entry of the taxonomic parent and a Caution comment made. If the taxonomic parent is unavailable/unclear, the best match by BLAST may be used and a Caution added.

Participant Roles

Each molecule must have one and only one of both biological and experimental roles. In an experiment, such as affinity chromatography when a bait/prey relationship would be expected, but is not given by the author, the role should be 'unspecified'

Autocatalysis

Trans autocatalysis

2 molecules

Experimental role - neutral

Biological role - enzyme/enzyme target

Cis autocatalysis (when proven)

1 molecule (unless differentiated by tags, mutants etc.)

Experimental role - self

Biological role - self

Uncertain

1 molecule (unless differentiated by tags, mutants etc.)

Experimental role - putative self

Biological role - putative self

Experiments in which two participants must from a complex in participate in an interaction

Currently, we cannot sequentially annotate the formation of complex, and its subsequent participation in an interaction – in some cases the complex formation may only be inferred from the results rather than directly demonstrated. Example: cYFP-A + nYFP-B + CFP-C shown to be active in a FRET but the formation of A-B complex not directly demonstrated by BiFC. In this case the complex components should be shown as individual molecules but with the same role e.g. both may be 'fluorescence donor' or 'bait' as appropriate.

Features

1. To ensure compatibility across databases, it is recommended that that any shortlabel or full name be taken from InterPro and the alias field is used for internal database or author nomenclature. Use full name to describe required or possible.

2. It is strongly recommended that IMEx members supply an InterPro cross-reference, when possible, or to one or more of the member databases, for example Pfam, as this is required information for databases with a specific interest in domains.

3. Binding domains should be shown as linked, when this information is available.

Feature Range

FeatureRange: Location of the feature on the sequence of the interactor.

A feature may have more than one feature range, for example a domain formed by the 3-dimensional folding of a protein maybe formed from discontinuous regions of that sequence.

Authors often refer to a domain e.g. SH3, cytoplasmic without referring to actually residues on the sequence. These may be inferred by curators by using information supplied by InterPro members databases such as ProSite or Pfam, or by using programmes such as THMM or SignalP. In this case, the information must be tagged as such, by using the appropriate terms from the CV FeatureDetectionMethod.

Features should be matched to the given sequence of the primary protein AC number given in the entry, rather than that given by the author should there be a discrepancy. When location is unknown

leave start and end position empty and use ’undetermined sequence position’ (MI:0339) in startStatus and endStatus.

Deletion Mutants

1. Experimental detail should be collapsed to the shortest deletion mutant defining a binding site. It is not necessary to list each separate mutant used by an author.

2. The minimum length of a binding site is >3 amino acids. A deletion of one-three amino acids should be treated as a mutant. Deletion mutants modulating but not disrupting the binding can be treated as ‘mutation increasing/decreasing’ irrespective of length.

3. Entries may contain both a ‘sufficient to bind’ and a ‘necessary to bind’ if required to fully capture the data.

Mutants

1. Point mutations showing an effect should be described in full at the residue level. Mutations showing no effect need not be added.

2. When a series of point mutations have been made, if these have separately been shown to have an effect, these should be entered as separate features on the same interaction. If the effect has only been shown when mutations have all been made within the same instance of the molecule, this should be described as a single feature.

Mutated bait with differing effects of multiple prey

Bait A	w.t.	Mutant X	Mutant Y
Prey B	---	----	
Prey C	---		
Prey D	---	----	---
Prey E	---	----	

Wild Type

Bait A, Prey A, Prey B, Prey C, Prey D, Prey E

Mutant X (should be added to Bait A as ‘mutation’ (MI:0118)

Bait A, Prey D, Prey E

Mutant Y (should be added to Bait A as ‘mutation disrupting’ (MI:0573)

Bait A, Prey C

This exercise need not be applies to HT data (>100 interactors).

Post-translational Modifications

1. Where a PTM has been shown, or is believed to be, required for an interaction to occur, this should be added to the record at the residue level, where possible. These should be described as (required) or (possible)

Tagged molecule

1. Details of all tags should be added, at the residue level when possible.

2. If a tag is not in the PSI-MI list, the molecule should be marked as ’tagged (MI:0507)’ and details of tag given in text.

3. Complementation assay tags are implicit in the methodology and need not be marked as ’tagged’ or have the actual tag added unless the member database chooses to do so. Should the complementation method be non-standard however, the tags should be curated to the deepest level possible in the PSI-MI lists and further detail, if required, given as free text.

Compartment: The subcellular compartment in which the interaction has been shown to take place (experimentally verified not inferred). It is strongly recommended that all database use Gene Ontology Component CV, this will be made mandatory once all databases can comply and add this as a cross-rerefence on the interaction.. GO has terms for translocations which should be used rather than listing two locations. If no translocation has been demonstrated, then two (or more) terms may be listed. If a term is missing in GO, please request a new term rather than use an alias.

Protein Updates

After every UniProt release, it is recommended that the local database is updated to follow any changes in the protein database. That way, we are up to date with UniProt. The following flow explains the process step by step. Curators should follow a similar process when replacing or remapping proteins. Participants which have been identified by nucleotide sequencing, with no reference to the protein sequence (e.g. Y2H screening) may be remapped to new gene models for the same gene should an old one be withdrawn.