Annotation Rules

Annotation Rules (May 2024)

1. IntAct **7**
   1.1 IntAct data model 7
   1.2 Curation Depth 8
      Brief Summary of IMEx vs MIMIx Curation 8
   1.3 Controlled Vocabulary terms 9
   1.4 Accession numbers 9
   1.5 Short Label 10
   1.6 Full Name 10
   1.7 Annotation 10
      1.7.1 comment 10
      1.7.2 remark-internal 10
      1.7.3 caution 11

2. The Editor and publication life-cycle **13**
   2.1 The publication life cycle 13

3. Initiating curation – Publication Level **14**
      This is an example of a completed Publication level page 14
   3.1 Duplication of Effort Avoidance 14
   3.2 Publication annotations 15
      3.2.1 data-processing 15
      3.2.2 contact-email 15
      3.2.3 contact-comment 15
      3.2.4 experimental modification 15
      3.2.5 curation depth 15
      3.2.6 dataset (NA-MIMIx) 16
      3.2.7 on-hold 18
      3.2.8 author-submitted 18
   3.3 Publication cross-references (Xref) and cross-reference qualifiers 18
      3.3.1 PubMed ID (PMID) 18
      3.3.2 IMEx ID 18

4. Experiment Level **20**
   4.1. Experiment AC 20
   4.2 Experiment Short Label 20
   4.3 Host organism 21
      4.3.1 In vitro experiments 21
         4.3.1.1 Extracellular experiments 21

4.3.1.2 Experiments between intracellular proteins demonstrated ex vivo 22
4.3.2 In vivo experiments 22
4.3.3 Viral/Bacterial Strains 23
4.4 Interaction Detection Method 23
4.4.1 Examples of common interaction detection techniques 23
4.4.2 Enzyme Assays 25
4.4.3 RNA-protein or RNA-RNA interactions 26
Examples of common interaction detection techniques for RNA 27
4.5.1 Use of 'Predetermined' as a Participant detection method 28
4.5.2 Interaction/Participant Detection Methods 28
4.6 Experiment Annotations (NA-MIMIx except copyright) 31
4.6.1. Antibodies 31
4.6.2. confidence-mapping 31
4.6.5 copyright (MIMIx applicable) 32
4.6.6 data-processing 32
4.6.7 exp-modification 34
4.6.8 library-used 35
4.6.9 negative 35
4.6.9.1 Rules for annotation of negative interactions. 36
4.6.9.2 How to make an entry for a negative interaction. 36
4.6.10 uniprot-dr-export 37
4.6.11 variable and variable_2 37
4.7 Experiment cross-references and cross-reference qualifiers 38
4.7.1 method 38
4.7.2 see-also 38
4.7.3 source reference 38
4.7.4 target-species 39
4.8 Data not shown 39

5. Interaction **40**
5.1 Interaction AC 40
5.2 Short Label 40
5.3 Parameters 41
5.4 Interaction type 41
5.4.1 Association 41
5.4.2 Physical Association 42
5.4.3 Direct Interaction 42
5.4.4 Self or Putative Self Interaction 42
5.4.5 Colocalization 42

5.4.5 Proximity 43
5.5 Interaction Annotations (NA-MIMIX except Figure legends) 43
5.5.1 Commonly used Interaction Annotations 43
5.5.2 figure-legend (MIMIX applicable) 44
5.5.3 negative 45
5.5.4 uniprot-cc-note 45
5.5.5 url 45
5.5.6 3d Structural data 45
6.5.5.1 3d structure 46
6.5.5.2 3d-r-factor 46
6.5.5.3 3d-resolution 46
5.5.6 variable_condition_1/2 46
5.6 Cross-references and cross-reference qualifiers 47
5.6.1 identity 47
5.6.2 see-also 47
5.6.3 GO cross-references 48
5.6.4 intact-secondary 48

6. Participant **49**
6.1 Proteins 49
6.1.1 How to find a UniProt Identifier for your proteins 49
6.1.1.1 Tracing the Species of origin of a protein 50
6.1.1.2 Identification of naturally occurring fusion proteins 50
6.1.1.3 Artificial fusion proteins 50
6.1.1.4 Identifying a protein by sequence searching 50
6.1.2 Mapping of protein isoforms 51
6.1.3 Problems with mapping accession numbers/Special or Unusual situations 52
6.1.4 UniProtKB entries that contain multiple protein molecules 52
6.1.5 Synthetic Peptides 53
6.1.6 Proteins of unknown species 54
6.1.7 Protein update 54
6.2 Small Molecules
6.2.1 Glycans 55
6.3 Nucleic Acids 56
6.3.1 DNA 56
6.3.1.1 Curating ChIP experiments 57
6.3.2 RNA 58
6.4 Macromolecular complexes 58
6.5 Genes 58

6.6 Experimental and Biological roles 58
6.6.1 Experimental role 58
6.6.2 Biological role 60
6.7 Expressed-In (NA-MIMIX) 61
6.7.1 Heterologous protein expression 62
6.7.2 In vitro transcribed and translated proteins 62
6.7.3 Synthetic peptides 62
6.8 Stoichiometry 62
6.9 Participant name alias 63
6.10 Annotations (unique to Participant-level) 63
6.11 Experimental Preparation 63
Experimental preparation general annotation rules 64
Expression level*: 64
Sample process*: 65
Molecular source: 65

7 Participant Feature **67**
7.1 Fields required to describe a Feature 68
7.2 Examples of various features 68
7.3 Specific cases 70
7.3.1 Molecule is tagged 70
7.3.2 Molecule is radio-labelled 70
7.3.3 Molecule is post-translationally modified (PTM) 71
7.3.4 Molecule is a fragment, truncation or deletion construct 73
7.3.5 Molecule contains a Mutation 74
Rules for mutation increasing/decreasing short label notation: 75
Unusual feature: Polyglutamine Tract/Repeat: 75
Example of a mutated bait with differing effects on multiple prey 75
Frameshift and STOP lost mutations 76
STOP loss mutation 77
Frameshift Mutation 79
Complexes having mutant participants with stoichiometry more than one 80
7.3.6 Linking features 81
7.3.7 Fusion proteins/unusual Tags 82

8. BioSource **84**
n8.1. How to create a new BioSource/Host Organism 84
8.2 How to create a Cell or Tissue Type 84
8.2.1 Tissues vs cell lines 87
8.3 Cell Type 87

Immortalised cell lines are referred to as ‘cell types’ in IntAct. 87
8.3.1 Re-classified cell lines 88

9. Creating an Interactor **89**
9.1. Protein 89
9.1.1 Short Label 90
9.1.2 Source 90
9.1.2.1 Chemical synthesis 90
9.1.2.2 Chimeric protein/Fusion protein 90
9.1.4 Sequence 91
9.2 Small Molecule 92
9.3 Nucleic Acid 92
9.3.1 Examples 93
9.3.2 Species 94
9.3.3 Sequence 95
9.3.4 Cross-references 95
9.4 Gene 95
9.4.1 To Create a new Gene entry 95
9.4.2 Species 96
9.4.3 Sequence 96
9.4.4. Cross-references 96
9.5 Annotations specific to new participants 96
9.5.1 copyright 96
9.5.2 no-uniprot-update 96

10. Special Cases **97**

10.1 Prepublication or submitted data **97**
10.1.1 Creating a publication, lacking a PMID in the editor 97
10.2. Importing interaction data from other databases 99
10.3 Annotating data from partially curated source 99
10.4 Unpublished data submitted by the author as an extension of published data 99
10.5 Handling Amended Large Datasets 100
10.5.1 When the error is by the database 100
10.5.2 When the error is by the author 100
10.6 Data represented more than once in the database 100
10.7 Curation request 101
10.8 Causal Interaction Annotation – draft annotation rules 101

1. IntAct

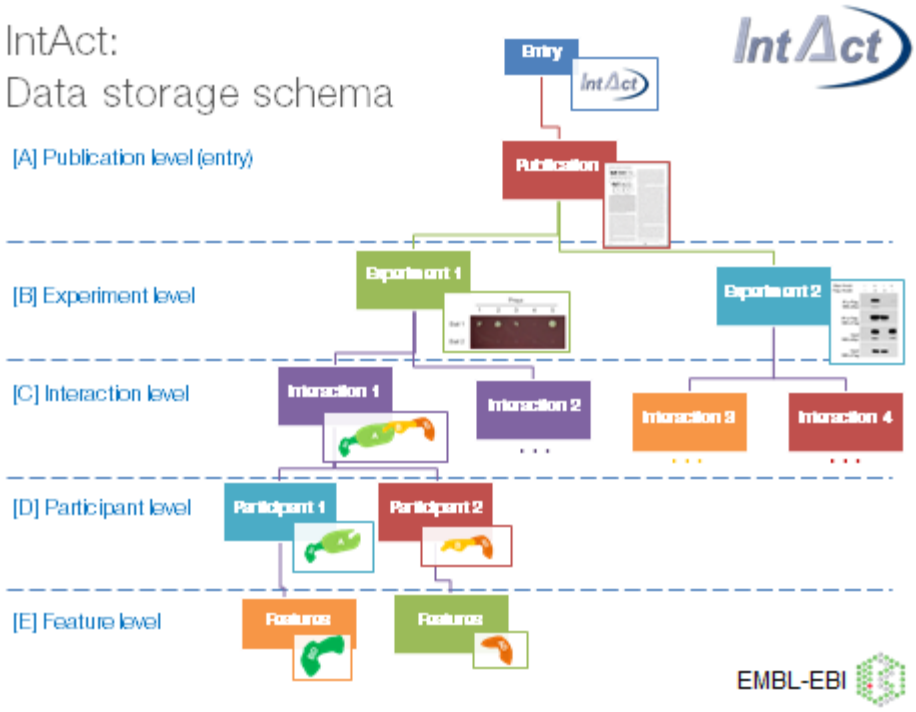
IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions. There are 2 database instances: PROD (production) and TEST (test). Curators add or edit data into the production database via the production editor. The test database is used to test new updates of the database and editor before putting them on the production editor. Large-scale data is always uploaded first on the test editor to ensure correct representation.

1.1 IntAct data model

Data curated in IntAct may be from peer reviewed publications, submitted pre-publication data or data submitted by authors as an extension of data related to a previous publication.

The IntAct data model is built on the following concepts:

- An entry is a publication or pre-publication.
- One publication contains one-to-many experiments;
- One experiment has one-to-many interactions;
- One interaction has one-to-many participants.



A Participant is an Interactor plus its features.
Interactors can be proteins/peptides, genes, nucleic acids, or small molecules.
A participant can have none-to-many features.
Features: these further describe the component of the interaction e.g. region of protein used, PTM present on the protein involved in the interaction, mutations or a tag.

1.2 Curation Depth

Curation can be carried out at two levels: IMEx (e.g. PMID:22453911) or MIMIX (minimum information for interaction exchange level (e.g. PMID:17687370).

The MIMIx rules can be found here:
http://www.nature.com/nbt/journal/v25/n8/box/nbt1324_BX1.html

The IMEx rules can be found at:
http://www.imexconsortium.org/sites/imexconsortium.org/files/documents/imex_curation_rules.pdf

IMEx IDs are assigned from IMEx Central (https://imexcentral.org/icentralbeta/), which the editor accesses to assign an IMEx ID. Once an IMEx ID has been assigned, the paper should either be completed in full or, if this subsequently proves impossible, a comment should be added to the publication record in IMEx Central explaining why not. Once an IMEx ID has been assigned, it cannot be removed from the paper in IMEx Central.

Brief Summary of IMEx vs MIMIx Curation

IMEx	**MIMIx**
Capture all PPI experimental data in publication - includes colocalisations but NOT genetic interactions	
Publication	
PMID + autocomplete	PMID + autocomplete
Author contact details	Author contact details
IMEx designator	MIMIx designator
Dataset	
Experiment	
Host organism-tissue/cell	Host organism-tissue/cell
Interaction Detection method	Interaction Detection method
Participant Detection method	Participant Detection method
Additional annotation – experimental modification, data processing	

Interaction	
Interaction Type	Interaction Type
Figure legend(s)	Figure legend(s)
Additional annotation – Comments, Caution	
Location xref (GO) – usually only colocalisations	
Participants	
Experimental/Biological role	Experimental/Biological role
Expressed in (in vitro expts only)	
Features – tags, radio/isotope labels, deletion mutants, point mutations	
Binding site xrefs (InterPro)	

In the subsequent rules 'NA-MIMIX' will be used as a tag to indicate that this annotation need not be added to a MIMIX level entry.

1.3 Controlled Vocabulary terms

IntAct uses the HUPO-PSI Controlled Vocabulary (CV) for interaction detection, Participant Detection, feature detection method, feature type, Interaction type, interactor type, database, interactor roles, and cross-reference types. The ontology lookup service (OLS) provides an overview of various PSI (CV) terms https://www.ebi.ac.uk/ols/ontologies/mi.

If an adequate term to describe the methods used in the publication is not available in the current CV hierarchy, a new CV term should be requested by creating an issue at the PSI-MI GitHub site: https://github.com/HUPO-PSI/psi-mi-CV/issues. Alternatively, an email to intact-help@ebi.ac.uk can be sent.

In cases where the controlled vocabulary term is pending, choose a possible parent of the requested term, give a description of the method used by the authors under annotation topic 'exp-modification' and a 'remark-internal'- 'Revisit when CV in place'. We run a periodic clean-up of the pending CV terms.
Additional annotation terms used by IntAct are available from ftp://ftp.ebi.ac.uk/pub/databases/intact/current/cv/.

1.4 Accession numbers

Each IntAct object is given an accession number (EBI-xxxxxxx), automatically generated by the editor. If a change is made which involves the transfer of contents from an entry to a new IntAct object, with the original IntAct object being deleted, then the accession

number of the old IntAct object will be cross-referenced in the new IntAct object using the Reference Qualifier 'intact-secondary'. This will ensure that the correct record/s is/are pulled up using both the former and the more recent accession numbers.

1.5 Short Label

Each IntAct object is given a mandatory short label. These use lowercase letters, numbers and special characters such as underscore (_), hyphen (-) and space. Any other characters are replaced by (_) underscore. There is a character limit of 256 on the Short Label. Whenever possible, short labels will be generated by the editor.

1.6 Full Name

Each IntAct object may be given an optional full name. The character limit on the full name is 1000 characters, and standard keyboard characters are allowed.

1.7 Annotation

Annotations are additional free text information which can be added to IntAct objects. They can be added at the level of Publication, Experiment, Interaction, Interactor or Feature. Again, only standard keyboard characters are allowed. Annotations are added under a series of specified headings – many of which are restricted to a particular level of an entry. Specific annotation will be detailed at the appropriate level, but the following may be added at any level

1.7.1 comment

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, and do not contain unusual (publication-specific) abbreviations as comments are visible externally.

Interaction AC	**Description**
EBI-6512836	The binding of the isolated FADD death domain to LRRK2 was weak.
EBI-6555250	Levels of H4K17Ac at the promoter level of this gene increased in UpSET (Q9VUB5) -/- cells
EBI-2658579	Atrx, but not Hira, is specifically required for H3.3 enrichment at telomeres. H3.3 colocalization with telomeres is lost in Atrx-null ESCs, but maintained in Hira -/- ESCs.

1.7.2 remark-internal

This annotation topic may be added at publication level if there is a common remark to be made about the entire publication. It can be added at Experiment and/or Interaction level – wherever is most appropriate.

‘remark-internal’ is for internal use only and is not made publicly available so comments and/or messages to oneself or other curators can be made here.

The remark-internal can be deleted if it is no longer appropriate.
This topic may also be used to describe why certain experiments described in a publication were not annotated. Some examples of ‘remark-internal’ used (at Publication level) are:

Did not curate Fig 3a/b/c (Kinase Assay) as LRRK2 was not purified other than by IP.
Awaiting reply from author - email sent 4/2/14
I am not sure if I entered all the co-localizations correctly?
Ang 3 in this paper must be UniProt Ang-4...
There is NO mention of an anti-HA antibody in the paper - so participant detection entered as "WB" rather than "anti-tag WB".
From Fig 7 I have entered the one mutation result stated to have statistical significance. (text states (fig 7a +B): "The G2019S, R1441C and I2020T mutations each enhanced binding of LRRK2 to MKK6," ??!

1.7.3 caution

This is used for warning the user:

- about possible errors 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.
- where there may be some difference in the sequence noted by the author and that present in the UniProtKB entry or method identification carried out by the curator.

Interaction AC	**Description**
EBI-2363999	A single band appears in the Rpt 2/3/4/5 identification region but the authors does not give an actual identification to this protein and it cannot be assigned from the image on the gel.
EBI-2939789	PMRT5 amino acids numbers have had +1 added to agree with UniProt sequences.
EBI-2464773	The PDB accession numbers given at the end of the paper for this crystal appear to be incorrect.
EBI-3505028	One IPI identifier was mapped to CDK2AP1 by the author, whereas in fact it mapped to CDK2AP2
EBI-2928481	Mutant Mdm2 (minus N terminal amino acids) does bind to p53 in Fig 2A, yet it does NOT bind p53 as shown in Fig 6E.

EBI-4373956	The cell type in which this interaction occurred was not specified.
EBI-5351142	myc-CREPT rather than flag-CREPT is stated to have been used in the figure legend.
EBI-6267663	Binding region length is given as 1101-1640 in the text. As the known length of the protein is 1460, it has been assumed that two digits have been reversed.
EBI-5326135	The molecule "TCF4" is the protein "Transcription factor 7-like 2" (Q9NQB0 (TF7L2_HUMAN)) whose Gene ID has a synonym of TCF4.

2. The Editor and publication life-cycle

For most small scale experiments (usually less than 100 binary interactions or less than 100 interactors per n-ary interaction) the IntAct data entry is made using the editor. The editor is available for curation only and needs a login and password. The editor can upload a list of interactors and enable the curator to assign common roles, expression details and annotations to the participants. A separate editor guide is available for anyone interested in curating within the IntAct database (request by mailing intact-help@ebi.ac.uk).

Every IntAct object, for example experiment, interaction, interactor, participant, feature, annotation, biosource, CV object is given an IntAct accession (AC) number when it is created. Every object is identified by an optional full name and mandatory short label. All objects in an editor page are allocated an Intact AC number on pressing 'Save'.

2.1 The publication life cycle

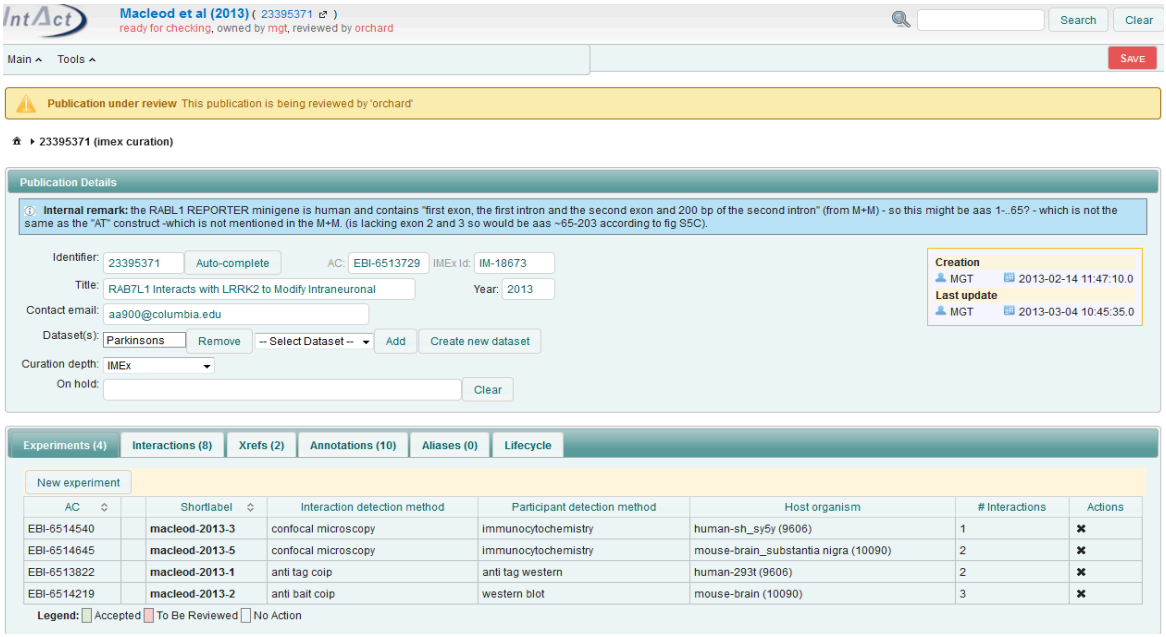
- A publication is initiated by entering a PMID and pressing <Auto-complete>.
- Once the entry has been saved, it is assigned to that curator as owner of the entry.
- When the complete publication has been curated, the curator clicks on the <Ready for checking> button. At this point, a reviewer takes over the publication and checks if the data is entered correctly. The status changes to 'To be reviewed'.
- Once this entry is checked and 'Accepted' it is 'Ready for release'. If the entry is not accepted by the reviewer it is returned to the curator tagged as 'Curation in progress' and the background colour changes to red. The curator must then address the points raised by the reviewer.
- Once the entry is ready to be rechecked, the curator presses the <Ready for (re) checking button>. Feedback from curator to reviewer can be added by pressing the <Correction comments> button. The background colour changes to yellow and the entry has a 'Ready for checking' status.
- This process continues until the entry is released.

3. Initiating curation – Publication Level

On starting curation of a paper the curator needs to 'claim ownership' of the publication by going to the "CURATE" section → Create a new publication → enter the PMID → Auto-complete.
Most details (Journal, year of publication, author-list) will be added automatically, but the curator will need to add:

- Manually add the corresponding author's email address
- Select a "Dataset" - if appropriate.
- Assign the "Curation Depth" as IMEx or MIMIx
- Give the entry an IMEx ID if appropriate by clicking on the <assign IMEx ID> button.

This is an example of a completed Publication level page



3.1 Duplication of Effort Avoidance

When the curator enters a publication for curation within the editor, a check is made in IMExCentral to ascertain whether the publication has been curated, or reserved for curation, by another partner database. If so, a message will tell the curator that this is the case. If another database has curated, or reserved, the publication already it is strongly recommended that the publication is not redundantly curated. If the curator still decides to curate the paper, an IMEx ID will not be issued.

If the publication is already curated within IntAct, the curator will be taken to the curated entry by the editor.

3.2 Publication annotations

Along with the common annotations described above, a publication may have any of the following annotations.

3.2.1 data-processing

This topic may be added at publication level where the data-processing description is common to ***all*** the experiments (*See Experiment Annotations* section for further info and examples). Briefly, this annotation topic is used to describe the steps in processing of data to obtain the ***identifiers*** described in the entry. This annotation topic is visible in the public database.

3.2.2 contact-email

Usually this is the contact given in the publication. When there are many corresponding authors the contact emails for correspondence can be added in the same line as a **comma** (not semi-colon) separated list without spaces. Example: *john@ebi.ac.uk,doe@ebi.ac.uk*. Please make sure you do not add any space, or extra symbols.

3.2.3 contact-comment

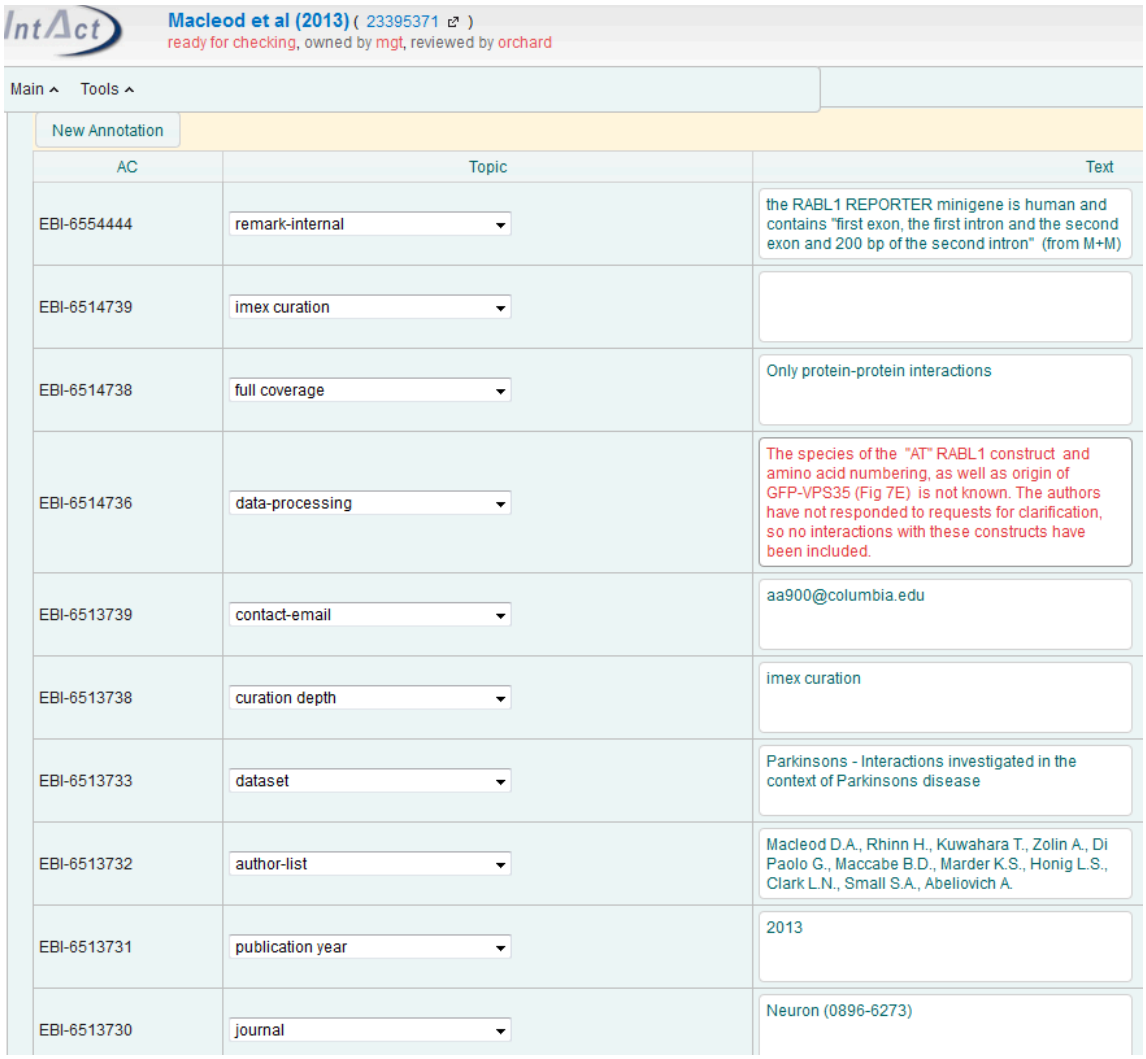
Example:
“The original address given in the publication was john@doe.com. This address is now invalid and has been replaced by that given in contact-email. The original address is stored here.”
This topic can also be used to indicate that the contact email is absent from the publication. This annotation topic is not viewed on the public database.

3.2.4 experimental modification

Experimental details are entered where experiments have a protocol that is **non-standard**, and cannot be easily entered elsewhere. This annotation is only added at the Publication level when the modification(s) applies to **ALL** Experiments. More commonly this term is used at the **Experiment** level, specifically for that experiment.

3.2.5 curation depth

Curators should indicate whether the entry has been curated to IMEx or MIMIx depth (see Section 1.2)



3.2.6 dataset (NA-MIMIx)

This annotation topic should be used to link various publications pertaining to a particular topic of interest.
You should ADD your publication to a relevant Dataset(s), if the paper contains interactions relating to that Dataset(s) - as listed below.
To do this click on the relevant Dataset, click on ADD, then SAVE. This will add the Dataset annotation to **every** experiment.
If you wish to create a NEW Dataset, please get approval for this first.

For further info regarding Datasets see:
http://www.ebi.ac.uk/intact/resources/datasets

A list of Datasets to which you can add publications:-

Cancer	Interactions investigated in the context of cancer

Alzheimers	Interactions investigated in the context of Alzheimers disease
Chromatin	Chromatin - Epigenetic interactions resulting in chromatin modulation
Cyanobacteria	Interaction dataset based on Cyanobacteria proteins and related species
Archaea	Interaction dataset based on Archaea proteins
Synapse	Interactions of proteins with an established role in the presynapse
Apoptosis	Interactions involving proteins with a function related to apoptosis
Parkinsons	Interactions investigated in the context of Parkinsons disease
Diabetes	Interactions investigated in the context of Diabetes
Affinomics	Interactions curated for the Affinomics consortium
NDPK	Interactions involving proteins containing InterPro domain IPR001564, Nucleoside diphosphate kinase, core
Cardiac	Interactions involving cardiac related proteins
Virus	Publications including interactions involving viral proteins
Neurodegeneration	Publications depicting interactions involved in neurodegenerative diseases.
Ulcerative colitis	Publications depicting interactions of proteins identified as having a link to ulcerative colitis
Crohns disease	Publications depicting interactions of proteins identified as having a link to Crohn's disease
Huntington's	Publications describing interactions involved in Huntington's disease
IBD - Inflammatory bowel disease	Publications depicting interactions of proteins identified as having a link to IBD
PDBe	Data obtained from the Protein Data Bank Europe by automatic import. Do not add manually added crystal papers to this dataset.

Do not add primary publications to dataset ‘PDBe’. These is reserved for direct imports from PDBe.

3.2.7 on-hold

Any experiment that is annotated with this topic will not be released to the public. This annotation is used to exclude submitted data which has not been published and integration of the data into IntAct is 'in progress'. This annotation should only be used when really necessary and the reason for the entry being on-hold should be entered in the free text box

Note: Subsequent to release of the PubMed ID or the completion of the entries from a PubMed ID the 'on-hold' annotation **must** be deleted. Experiments **will not** be released automatically after the date has passed.

This annotation topic is not viewed on the public database.

3.2.8 author-submitted

This refers to the data added to entries submitted by the author directly to IntAct prior to publication. The format of the free text field should strictly be as follows: *YYYY-MM-DD: J Doe, University of place*.

Experiment AC	Description
EBI-6476263	2012-09-27: M Boxem, Utrecht University
EBI-6430623	2012-12-05: M Meier, IMTEK, University of Freiburg

3.3 Publication cross-references (Xref) and cross-reference qualifiers

All added Cross references must have a qualifier which usually needs to be added manually. The primary reference will be added automatically at the publication level as part of the Auto-complete process.

3.3.1 PubMed ID (PMID)

This cross-reference is given the qualifier 'primary-reference', as it is the primary source of data in an entry. The journal details will be added to an entry when the PMID is used to create an entry, and the 'Auto-complete' button pressed. The PMID and any IMEx XRef will be carried over to any subsequent Experiments, as each Experiment within an entry is created.

3.3.1.2 Unpublished Data or data lacking a PMID

When the data has been submitted pre-publication or is a pre-print and a PubMed ID can only be assigned at a later date 'pubmed' should be the primary-reference and the primary ID column should be filled in with 'unassignedxxx, qualifier 'primary-reference'. On publication if a PubMedID is not available then DOI cross-reference may be used with qualifier 'primary reference'.

3.3.2 IMEx ID

IMEx IDs can be added to any paper curated to IMEx standards by members of the IMEx consortium, which have not previously been claimed as an IMEx record by another

consortium member. The addition of an IMEx ID will automatically trigger the assignment of the IMEx identifiers on all the experiments with the same Primary Pubmed Reference and all of the associated interactions. This accession is of the form IM-XXXXX. The qualifier will normally be IMEx-primary, unless an entry already has an IMEx ID, in which case a second number (which references the same entry) may be added with the qualifier ‘imex-secondary’. If an entry does not yet have a PMID, the IMEx ID must be added manually.

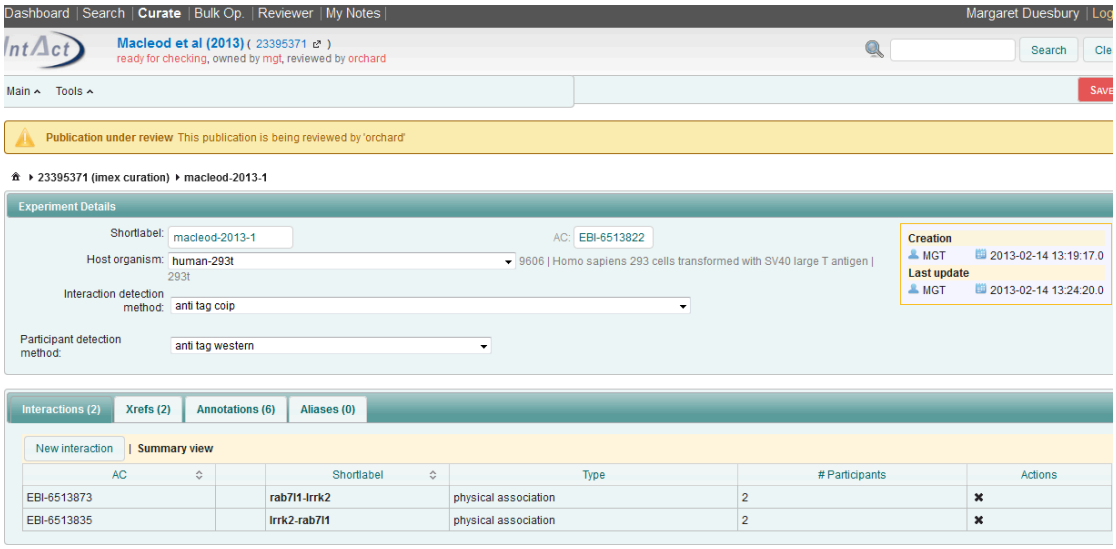
4. Experiment Level

The experiment records the experimental method used to demonstrate the interaction(s). One experiment may have many interactions linked to it; these interactions must all share the **same** experimental protocol. Within one Publication, multiple identical interactions can be found with every interaction being in a separate Experiment, due to differing experimental protocols being used for each individual interaction.

Note: The converse is not allowed i.e. one interaction may **not** belong to multiple experiments.

Experiments that demonstrate that an interaction does **not** occur can be entered in IntAct under special conditions. See: Annotation Topics 'Negative', section 4.6.9.

Example of an Experiment – with two interactions.



4.1. Experiment AC

A unique AC number is assigned to the experiment when you click the ‘save’ button. Searches on this number will bring up the entry.

4.2 Experiment Short Label

This is a mandatory field. It is normally automatically generated in the form “zhang-2012c-4”or similar. However errors can arise, so in order to correct these it is necessary to know that:

An experiment Short Label has the following format:
[first author last name]-[year of publication][optional char]-[integer]
Overall character limit is 256.
All characters that are not in [a-z] are replaced by _ (underscore).
The Short Label is lowercase.
When the name of an author has been used more than once (different PubMed ID) in the **same year**, enter year a, b, c etc. after the year of publication as illustrated below, based on the chronological order in which the data was entered in IntAct.
e.g. stargell-2003-1
stargell-2003a-1
The experiment number or final integer will be added based on the chronological order of entering the data in IntAct.

4.3 Host organism

This is a mandatory field.
The Host Organism indicates where the experiment has been performed. The Host Organism (plus tissue or cell-line, if relevant) is chosen from a drop down list of organisms. This menu uses the UniProt organism identifier code which provides a 5 letter code by which each organism may be differentiated. For example, the 5 letter code for "dog" = "CANFA" (the first 3 letters from **Can**is and first 2 from **fa**miliaris).
Additional information on the species level, for example details sub-species level, should be added in the annotation topic 'exp-modification'. For cell lines that have been modified, the host organism will be the unmodified cell line and the modification will be entered under annotation topic 'exp-modification'. Similarly when entering experiments from multiple cell lines which are very closely related (for example T-REX cell are a sub-line of HEK293 cells), you can enter one cell line, rather than several. Tissue and cell type should be used as host organism where applicable.

4.3.1 In vitro experiments

The interactions occur outside of the intact cellular environment. Experiments carried out literally "in glass", or commonly called "experiments in a test tube". Examples of when 'Host Organism' is *in vitro*:

4.3.1.1 Extracellular experiments

If one of the interacting proteins is extracellular or has an extracellular domain that participates in the interaction e.g. receptor-protein ligand interactions where the ligand is extracellular, the interactions are classed as **"*in vitro*"**. For a definition of extracellular see GO:0005576.
Interactions in which the interacting proteins are membrane-bound and the interaction is through their extracellular domains (e.g. by FACS analysis) are classed as **"*in vitro*."**

Example of the interaction of a receptor with a protein ligand by cross-linking and co-immunoprecipitation: EBI-80473

Example of interactions using cross-linking and 2D gels EBI-368460

[**Note:** The following receptor based interactions are curated as ***"in vivo"***
1. both the interacting proteins are membrane bound and it is not known if the interaction is in the extracellular, intracellular or transmembrane domain and cell lysate is used for determining interaction.
2. the study is carried out in a cell lysate and it cannot be determined if the interaction is between cells or within the same cell.]

4.3.1.2 Experiments between intracellular proteins demonstrated ex vivo

The interaction occurs out of the cellular environment e.g. on a column, in an instrumental cell, on a biosensor, a crystal. The components of the interaction may be purified or semi-purified and each component could come from any of the sources below.

i) Purified proteins: The proteins may be expressed in a heterologous system example: *E. coli*, yeast, insect, mammalian or/and plant cells and purified from these or from its natural source.
ii) Proteins are from more than one cell lysate
iii) Secreted protein
iv) *In vitro* transcribed and translated protein
v) Synthetic peptide
vi) Purified peptide
vii) Proteins embedded in the phospholipid bilayer, liposome or micelles

Details of heterologous protein expression for each protein must be entered in the 'Expressed In' on the interaction page.

4.3.2 In vivo experiments

The proteins interact within an intact cellular environment. All of the interacting proteins must be expressed in a cell or cell membrane; an exception would be micro-injected proteins. The proteins may be modified (tags, promoters, fragments, crosslinks). The cell may be subsequently lysed to enable access to an immuno-precipitating antibody or other detection agent.
The names of cell lines and tissue types used when the interaction is demonstrated to occur *in vivo* should be added in the experiment under "host organism". To create a 'Host Organism' or 'BioSource' please refer to Section 8.1. If the protein has been overexpressed or expressed at endogenous levels this information should be stored in annotation topic 'expression-level'. When the proteins are expressed in the same organism in which the interaction is occurring, the organism should **not** be redundantly added as 'expressed-in'
Example:
1. Human proteins A and B are expressed in Cos-7 cells and shown to interact by coimmunoprecipitation.

Experiment: Host Organism: Cos-7 cells (chlae-cos_7)
Note: In this experiment the host organism is C. aethiops (African green monkey) as the proteins were expressed in Cos cells and the complex was formed in Cos cells.

2. One cell expressing two or more transmembrane proteins and the intracellular domains or transmembrane regions are demonstrated to interact, for example by in vivo cross-linking using a cross-linker that crosses the cell membrane.
3. Detection of protein-protein interactions in live cells using techniques such as FRET (fluorescence resonance energy transfer) or BRET (bioluminescence resonance energy transfer).
4. Coimmunoprecipitation of proteins from a single extract or lysate.
5. Fractionated proteins - proteins are in a fractionated cell extract from cell lines, embryos or tissues. This refers to complexes pre-formed in cells and purified through subsequent steps. Example: cell extract, nuclear extract, vesicle fraction, Golgi membranes, membrane fraction.
6. If the membrane has been purified with the proteins and interactions within the membrane are studied this would then classify as *in vivo* interaction.

4.3.3 Viral/Bacterial Strains

For Viral or Bacterial Strains, we map to the strain level, if this is known.
If you do **not** know the exact strain used, as is frequently the case, take your UniProtKB identifiers from the Reference Proteome.

4.4 Interaction Detection Method

This is a mandatory field. Add a term by selecting from the drop-down menu.
Use the term from the hierarchy of the controlled vocabulary that describes the interaction detection as closely as possible. Experiments where the interaction detection technique used is a child term of ‘post transcriptional interference’ or ‘genetic interaction’, “Genetic Reporter Assay” or “phenotype-based detection assay” are not allowed by IMEx and not entered in the IntAct database.

4.4.1 Examples of common interaction detection techniques

Method	Experiment no.
molecular sieving	EBI-2911349, EBI-5668263 EBI-4305071
Fps	EBI-2941846, EBI-2877525
anti tag coimmunoprecipitation (in vitro)	EBI-2911101
anti bait coimmunoprecipitation (in vivo)	EBI-5378672
Cosedimentation (density gradient)	EBI-6127934
Cosedimentation (solution)	EBI-3989600
Far western	EBI-2943702
fluorescence spectr	EBI-1774733
Facs	EBI-3870457
pull down	EBI-2911426, EBI-6138583, EBI-5912162, EBI-4481555

protein crosslink	EBI-3960025
Itc	EBI-2655240
Spa	EBI-23597562
x-ray diffraction	EBI-2464773, EBI-2213576
Fret	EBI-4373847, EBI-6127663
Nmr	EBI-6264210, EBI-3507479
dhfr reconstruction	EBI-448165, EBI-1296, EBI-6313895
molecular sieving	EBI-4305063, EBI-4289462
PLA/pELISA-proximity ligation assay	EBI-4312047, EBI-6143088
BiFC	EBI-6097675
FRET	EBI-5544932, EBI-5240070, EBI-6162091
Crosslink	EBI-3959976, EBI-4533282
confocal microscopy	EBI-3950472, EBI-3863350
spr	EBI-4406354 , EBI-5452477
two hybrid (library screen)	EBI-6309959
Two hybrid	EBI-6313541
Tandem Affinity purification (TAP)	EBI-4374417, EBI-2130618

Further examples of interactions using a particular Interaction Detection Method in IntAct can be found by locating the MI number (PSI-MI accession number) for the method in OLS (Ontology Lookup Service): www.ebi.ac.uk/ols/ontologies/mi

After you have found the MI number, enter it into the “Advanced Fields” search on the IntAct web page, under the Field “Interaction Detection Method”.

If the method as described in the publication does not exist in the drop-down menu search for it in OLS as it may be a synonym.
Additional experimental detail, for example variations from standard protocols, should be added as an <Annotation> “Experimental Modification (Exp-Modification)”.

Purification of a protein complex may be a sequential process. If the process can be annotated to a common parent e.g. ‘chromatography technology’, this should be selected. If this cannot be achieved, the curator needs to select a key step in the purification process which separates the complexes out from a protein mix. 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. 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.

Cross-linking will almost always be taken as inferring any subsequent isolation technique, which will not be separately captured. In cross links, if the subsequent isolation technique

employee a protein as ‘bait’, participants will be annotated to have experimental roles ‘bait’ and ‘prey(s)’

4.4.2 Enzyme Assays

There has been debate as to whether an immunoprecipitated enzyme preparation is 'clean' enough to be used in an in vitro enzyme assay, or if the data should NOT be added as other enzymes may also be pulled down in the same preparation and may actually be the active factor in any subsequent assay. Enzyme assay data should therefore only be captured under one or more of the following circumstances:

1. The author (or commercial producer) describes further purification of an enzyme beyond just an initial IP.
2. The author shows by stained gel that the preparation is pure.
3. The enzyme activity is monitored by radioisotope incorporation, an autoradiograph is performed and the author shows the entire gel, and no other bands are present suggesting contamination.

The activity, and resulting process should also be captured using appropriate GO terms cross-referenced at the interaction levels, for example a protein kinase assay should have:

GO:0004672 protein kinase activity
GO:0006468 protein phosphorylation OR GO:0046777 protein autophosphorylation

Resultant ptms

The results of an enzyme interaction may be captured by adding the ‘resultant ptm’ as a feature on the Participant that is the “Enzyme Target”, using the following syntax:-

Shortlabel: enter the amino acid and residue number affected - if known, e.g “ser-99”. If the residue is not known enter “residue” and range “?-?”.
Feature Type: Enter correct Protein Mod term {OLS-MOD CV term}, e.g “opser”
Feature Role: Add “resulting-ptm” from the menu.

[**Note**: This replaces the legacy Annotation topic ‘resulting-ptm’ of the type e.g. P12345 ser-99 O-phospho-L-serine]

When you curate resulting-ptm features, please add a single feature for each amino acid/position affected. No multiple ranges should be added in these cases. This is because there are separate CV terms for each phosphorylated amino acid, so each one of them needs to be curated as a separate feature.

When enzyme assays are detected by a radio-labelled transferred molecule e.g. protein kinases, phosphotransferase, sulfurtransferases, the radiolabelled entity is not normally added e.g. ATP need not be added to every kinase assay.

In the case of transferases (e.g. phosphotransferase, sulfurtransferases) the biological roles should be the appropriate donor/acceptor rather than enzyme/enzyme target whilst we can only select one role.

4.4.3 RNA-protein or RNA-RNA interactions

Some Interaction Detection Methods are specific for RNA-protein or RNA-RNA interactions:

- **Clip (**MI:2191) and its variations: **Clip-Seq** (MI:2192); **iCLIP** (MI:2193); **Par-Clip** (MI:2188)

Combination of cross-linking and co-immunoprecipitation aimed to find protein-RNA interactions. The canonical method uses first a cross-linking procedure over a tissue sample or lysate, followed by immunoprecipitation using antibodies specific for the protein of interest. The protein is the bait and the RNAs are the preys. The detection method can be nucleotide sequence or Northern Blotting

- **CRAC** (MI:2194)

Similar to Clip, but the protein of interest bears a tag used for pull-down or immunoprecipitation.

- **CLASH (**MI:2195)

Used to detect RNA-RNA interactions that occur in the proximity of an RNA binding protein of interest. The RNA is crosslinked to the protein, that is immunoprecipitated. After RNAsi treatment the RNA hybrids protected from degradation by the protein are ligated and chimeras are sequenced, so the detection method is always nucleotide sequence.

The two RNAs are Neutral Components and the protein, when indicated, can be added as a comment in annotation.

- **Chemical RNA modification plus base pairing prediction** (MI:2224)

Used to detect RNA-RNA interactions that guide the chemical modification of the target RNA. RNA-RNA pairings are tested by knocking down one of the two interacting RNAs and then experimentally determine if its presence is required for the other to be chemically modified. The interaction type is a physical association and the participant

detection method is often primer extension assay (MI:2203). RNAs are neutral components and the modified base must be indicated in features with ChEBI ID (e.g. PMID:9891049)

- **Luciferase Assay (**MI:2285 miRNA interference luciferase assay)

The mRNA regulative region, often located at the 3' UTR of the mRNA, is cloned at the end of the luciferase gene, in suitable vectors. When the plasmid is co-transfected with the microRNA, the luciferase expression is significantly lowered. The interaction MUST be confirmed by mutating the microRNA seed sequence site in the 3' UTR of the mRNA to check that that the sequence is essential i.e. the binding is true (e.g PMID 26184978). The interaction type is a physical association and the participant detection method is tag luciferase for the mRNA and predetermined for the microRNA. RNAs are neutral components but there is a causality relation so that the microRNA is the regulator and the mRNA is the regulation target.

The mutation in the seed sequence is annotated as a feature of the mRNA.

Examples of common interaction detection techniques for RNA

Method	**Experiment no.**
Clip-seq	EBI-20978107
Crac	EBI-10835868
Clash	EBI-10954258
Chemical RNA modification plus base pairing prediction	EBI-11792756
Luciferase Assay	EBI-21006454

4.5 Participant Detection Method

This is a mandatory field. Add a term by selecting from the drop-down menu.
'Participant Detection' describes the method used for detecting the participants of an interaction. It may refer to detection during, or after the interaction has occurred.

If more than one method is used for Participant Detection in an experiment, the major or predominant method, or that identifying the prey, is entered in the drop-down menu at Experiment level. Any additional methods used can be added directly to the Participant under "Identification Method" from the drop-down menu at Participant level. e.g. in Co-localization experiments participants may be detected by a mixture of Tag Fluorescence and immunostaining (antibody staining).

4.5.1 Use of 'Predetermined' as a Participant detection method

Predetermined (or child terms) will be used whenever a group has introduced a known protein into a system and then utilised the fact that they know that this protein is already present in the system as a basis for identification. For example, tagged proteins are expressed in E.coli and purified. They are subsequently mixed and the comigration of the complex (with proteins identified purely by molecular weight following protein staining) down a column or in a gel is taken as evidence of an interaction. If using a matrix of known proteins in a two hybrid assay, the participant detection method is 'predetermined' unless the author has specifically stated that the molecules were subsequently re-sequenced or detected by another method.

Proteins from a lysate **cannot** be identified purely on the basis of molecular weight. 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 or dimers inferred by molecular weight from whole cell lysates. These should not be captured as there had been no purification step and other proteins may be responsible for the change in molecular weight. Curation will only commence when a complex has at least been partially purified. A reference to a previous paper is NOT acceptable. However, if a publication has participants identified by 'molecular weight estimation by staining' or a child term but 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 are used and participants identification using 'molecular weight estimation by staining' or child terms this data CAN be curated.

4.5.2 Interaction/Participant Detection Methods

Examples of specific cases, pitfalls for the unwary & some common errors or omissions:

Interaction Detection Method	**Comment**
Far Western Blotting	the protein probe is the bait and the protein interacting with it is the prey (which is often immobilised)
Crosslinking	The methodology is described as

	'cross-linking-study' irrespective of additional purification steps. Any previous/subsequent isolation technique(s) will not be separately captured, but entered as an "Exp-Mod". If the subsequent isolation technique employs a protein as 'bait', participants will be annotated to have experimental roles 'bait' and 'prey(s) ● If there is no additional purification, and the host organism environment is either whole cells or a cell lysate with binary pairs generated, the experimental roles=neutral, interaction type=proximity ● If a whole cell/cell lysate is crosslinked followed by coip (or any method employing a bait) to generate n-ary data, roles bait/prey (as per method), interaction type = association/physical (as per method). If purified by gel filtration (or any method not employing a bait protein) roles=neutral, interaction type=physical ● If a complex is purified from a whole cell/cell lysate followed by crosslinking, to generate n-ary data, roles bait/prey/neutral (as per method), experimental preparation= partially purified, interaction type = physical ● Crosslinking of 2 purified proteins, host organism= in vitro, roles neutral (unless a purification involved), experimental preparation= purified, type=direct ● Special cases e.g. an unnatural amino acid crosslinked with UV-activated azide, should be considered on a case-by-case basis with size of cross-linker taken into consideration
Two hybrid screens	When a prey cDNA library is used (for screening) the Participant Detection method is "nucleotide-sequence". (The "library used" should be added to the Experiment page.)
Two hybrid	Use "Predetermined" for participant detection for any further two hybrid assays after the initial screen when there has been NO re-sequencing of constructs.
Protein Complementation Assays (including multiple children e.g. Y-2-H,	Tags need not be added if they are present in the Interaction Detection Method. However if there is a choice of tag or the tags are non-standard, they should be added.

bifc, dhfr, ubiquitin-reconstruction etc……)	NB: all these assays normally have only Binary Interactions – one Bait-One Prey.
molecular weight estimation by staining	Only enter this data if further confirmation of the participants has been obtained by an alternative method such as MS or Western, if *in vitro* purified proteins have been used or in the case of radioactively-labeled, *in vitro* translated proteins obtained with reticulocyte lysate system, where the radiolabeled protein is not purified, but it is overwhelmingly dominant in the lysate.
Co-IP	If preys are detected by two different methods (e.g. anti-tag WB and WB) on the same blot enter data for Participant Detection at the highest level (Western blot). Alternatively the major or predominant method is entered in the drop-down menu at Experiment level. Any additional methods used can be added directly to the Participant under “Identification Method” from the drop-down menu at Participant level.
Fluorography	Enter as autoradiography
Mass Spec	Enter Participant Detection as “MS Participant” if no type of MS is stated or can be found. Tandem MS (or MS/MS) is entered as “Sequence Tag”.
Chip	Interaction Detection method is “CHIP” and Participant detection is usually “primer specific pcr”.
ChIP-seq	Interaction Detection method is “chip” and Participant detection is “nucleotide sequencing”.
re-ChIP (Sequential ChIP, Chromatin Re-IP and ChIP Re-ChIP)	Re-ChIP enables sequential chromatin immunoprecipitations to be performed using two different antibodies so that you can assay for the simultaneous presence of two proteins or distinct histone modifications at the same genomic region of interest, the protein pulled down by the second antibody used should be taken as bait, the first protein precipitated and the DNA sequence should be taken as prey.
EMSA	This should only be annotated if the sequence of the (labelled) probe is given and can be added to a created entry. A cross-reference should be added for the probe – if possible. Roles: The labelled probe is the bait, the bound protein is the prey.

Co-IP (generally entered as anti-bait or anti-prey)	Western blots: if it is not clear if the author has stripped and re-probed the same blot with multiple antibodies, OR repeated the co-immunoprecipitation multiple times and checked each with a single antibody, the experiment should be represented once as one bait, many prey interaction.
PLA	Participant's Experimental role should be "Neutral component". Interaction Type is "Proximity " and Participant Detection should be "Antibody Detection".

Further details can be found in the TABLE of PSI-MI dependencies –found here:-

https://docs.google.com/spreadsheet/ccc?key=0AtAwdqt9ekgYdFpIWTFrWEMxZGRXemdmWHNkS1RPRWc&usp=sharing#gid=0

[TAKE CARE not to type anything in this page, as anything added/altered will be retained!]

4.6 Experiment Annotations (NA-MIMIx except copyright)

Annotations on the Editor-Experiment page relate to the experimental conditions only. Annotations related to individual proteins should be entered on the participant. The following annotation topics (in addition to comment, caution and remark-internal, see 1.7) may be used while curating an experiment:

4.6.1. Antibodies

This annotation can list the specific antibodies used in the interaction detection or participant identification. This detail is not routinely added, but rather when an unusual entity is used.

4.6.2. confidence-mapping

This annotation topic gives a description of the method/s used by the author for confidence mapping the interactions attached to this experiment.
In large-scale experiments where the authors of the paper have assigned confidence values to the interactions, the experiment must contain an explanation of the author's definition of the confidence values. This must be written in the experiment annotation topic 'confidence-mapping'.
In addition, the author confidence values that are suitable for export to UniProtKB must be added in the experiment annotation topic 'uniprot-dr-export'
Example: In experiment EBI-332598 the annotation is as follows:

Experiment AC	Topic	Description

EBI-332598	confidence-mapping	The authors have assigned the following confidence values to the interactions: Core-1: found at least three times independently and the AD-Y junction is in frame. Core-2: found less than three times independently, retest positive, AD-Y junction is in frame. Non-core: all other interactions.
EBI-332598	uniprot-dr-export	CORE_1

Historically, we have captured the high and low confidence interactions and filtering is possible based on the values available in author-confidence annotation on interaction. From the Heidelberg 2011 IMEx meeting, only the high confidence interactions will be captured, where the confidence is expressed by the authors as high.

4.6.5 copyright (MIMIx applicable)

Unusually, 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.
Example: EBI-449125 and EBI-449126 have the annotation copyright “These interactions are the sole property of HYBRIGENICS, and shall not be used for any business or commercial purposes without the prior written license from HYBRIGENICS (http://www.hybrigenics.com)”.

4.6.6 data-processing

This annotation topic is used to describe the steps in processing of data to obtain the **identifiers** described in the entry, and placed at the experiment level when only applicable for a specific experiment. Information about the original number of interactions described by the authors in the paper and the corresponding number in IntAct, if different from the one published in the paper, should also be stored here.

Example:
A unique UniProtKB entry could not be associated with (author identifier/name) since (organism/subtype/exact gene product) of interactor cannot be ascertained.

Proteomics Papers: All too frequently with Mass Spec protein identifications, the wrong species is given in the results. E.g. the Experiment uses mouse-ES cells, and manages to identify a “dog” protein. This is likely to be a homologous protein, and a Data-Processing comment should be added of the type: “gi|51701919|sp|P63049| Ubiquitin is a dog protein - this has been remapped to mouse.”

When the method by which a protein has been identified is not immediately apparent from the original paper, this should be detailed. Format should be
Source of Protein X supplied by author
Source of Protein X taken from PMID: 1234567

Replies from authors confirming or stating the origin/source and/or identity of proteins should be entered here.

If the protein identification is doubtful, do not add the affected proteins to any interactions, and add a Data-Processing comment stating that the proteins have not been added, giving a reason

Some examples of other types of Data-Processing Comments:

EBI-6176354	The number of interactions found in the paper following the author's cut-off values is different from the one they give in the text. The authors were contacted to clarify this, but no answer was received.
EBI-6550704	The authors state that the Gfi1-Flag contruct used in Fig 5H contains the human GFI1 sequence.
EBI-6509682	Origins and details for the Flag-MKK3 and mutant constructs have been taken from the Addgene repository. The authors have previously confirmed that the FLAG-MEKKS are Human in origin. As the authors did not respond to subsequent requests for information, the origins of JIP1-4 (gifts from Dr. Roger Davis) have been taken from these sources (all quoting RJ Davis as co-author). JIP1 (PMID:11562351) = Murine JIP2 (PMID:10490659) = Human JIP3 (PMID:10629060) = Murine JIP4 (PMID: 15767678) = Murine
EBI-6548358	Tlr11 does not exist in humans. As the authors cloned both Tlr5 and Tlr11 it has been assumed both are murine.
EBI-6511141	Origin of ATXNL2 constructs according to PMID 11784712, of ATXN2 constructs according to 15663938.
EBI-6504820	The authors have confirmed that myc/gst tagged 14-3-3 proteins are murine in origin, as is PKA (obtained from New England Biolabs).
EBI-6513729	The species of the "AT" RABL1 construct and amino acid numbering, as well as origin of GFP-VPS35 (Fig 7E) is not known. The authors have not responded to requests for clarification, so no interactions with these constructs have been included.
EBI-6512291	Interactions involving the use of "PKC activity assay from Enzo Life Sciences" (Kinase Assay) cannot be entered as the reaction substrate is only described as

	a "peptide" and Enzo does not wish to further identify it. Interactions involving HSP90 have not been included due to lack of antibody specificity. The authors have confirmed that the IRS1 Kinase assay substrate is Rattine in origin (Millipore 13-124).
EBI-6554661	Source of Flag tagged Sirt6 and c-myc taken from previous work of the author (PMID: 23142079 and 20592034 respectively)
EBI-6270859	The authors confirm that both TLR3-Flag and TRIL-V5 are of human origin.
EBI-2657512	hnRNP G-T (ref XP_914797) the NCBI states "report removed", remapped to Q9DAE2 gi|51701919|sp|P63049| Ubiquitin is a dog protein- this has been remapped to mouse.
EBI-2696332	gi:480806 - "report removed", so remapped to P53534. Similarly, reports removed or unavailable for: gi:76496512 remapped to P09606, gi:34869154 remapped to P25286, gi:34879653 remapped to P26284, gi:62644670 remapped to P16086, gi:62664168 remapped to P60711, gi;631898 remapped to P61765, gi:62653035 remapped to Q6P9V9,gi:62653090 remapped to Q4FZU2, gi:62654117 remapped to Q7TP47, gi:70850 remapped to P62909, gi:34876714 remapped to B5DEN5, gi:62718633 remapped to P04256, gi:62659231 mapped to Q6URK4, gi:62644588 also remapped to Q6URK4, gi:62650074 remapped to P62083, gi: 56385
EBI-2550901	The authors have provided UniProtKB identifiers for most of the proteins. Where Ensembl identifiers were provided, mapping to UniProtKB identifiers was carried out based on cross-references in UniProtKB, HGNC or MGI records. ENSP00000382001 and ENSP00000386061 were mapped to UniProtKB accession numbers based on their gene names.
EBI-3964609	Ensembl protein IDs given in Supplementary Table 1 were mapped to UniProt IDs, at the sequence level, using the PICR mapping service.

4.6.7 exp-modification

Experimental details are entered where experiments have a protocol that is non-standard, i.e. where the CV terms for either ‘Interaction Detection’ or ‘Participant Detection’ are not adequate. Example: a modified protocol from the one described in the CV term used.

Examples

Accession Number	Annotation Description
EBI-6551052	Proteins were first coimmunoprecipited with an anti-Flag antibody before a second round with the specified bait.
EBI-6550960	Thymocytes were irradiated (5 Gy) and chromatin was generated either at 90 minutes later (Fig S4 K/L) or at 0, 30, 60, 90 and 120 minutes later (Fig S4 M).
EBI-6553615	Phoenix cells are a HEK293T variant.
EBI-313006	Binding of purified LRRK1 and LRRK2 to ATP was measured using four different ATP derivatives coupled to agarose beads. Binding to the beads was inhibited by the addition of free ATP to the binding reaction but not GTP.
EBI-6310657	Anti-Flag agarose beads containing 3xFlag-LRRK1 or 3xFlag-LRRK2 were incubated with GTP-α-33P alone or with cold nucleotides. After incubation, excess nucleotide was removed, and the amount of bound isotopic GTP was measured by scintillation counting.
EBI-6288136	Molecules tagged with a fluorescent label were detected by tag fluorescence.

4.6.8 library-used

This annotation is most frequently used with two hybrid screening assays.
The information about the library which was scanned to obtain the interacting protein clone is recorded under this annotation topic.
Example:

Experiment AC	Annotation Description
EBI-3925930	Human liver cDNA library
EBI-6502381	HeLa cDNA library
EBI-2298471	Arabidopsis cDNA library of seedlings subjected to various abiotic stresses as described in the text.

4.6.9 negative

This is used for annotation of experiments that demonstrate that an interaction does not occur, but is used only in very specific circumstances.

4.6.9.1 Rules for annotation of negative interactions.

1. The interaction must have supporting positive interaction data in the same paper before any negative information can be entered.
2. Only experimental data in the same publication or by the same scientific group in another publication that demonstrates the negative interaction is acceptable: do not enter negative interactions based on comments in the paper or 'data not shown' since there is no evidence for this shown in the paper.
3. Post-translational modifications inducing negative interaction should be entered as annotation to the positive interactions as this is really supporting evidence for the positive interaction.
4. Protein isoforms that do NOT interact can be entered as negative interactions, provided there is complementary positive interaction data for other isoform/s (and you must know the exact identity of the isoform that does not interact).

4.6.9.2 How to make an entry for a negative interaction.

Positive Experiment	**Negative Experiment**
Experiment - create as usual	Create a new separate experiment: Add the following annotation: Annotation topic: negative: 'This experiment relates to a negative interaction.'
Interaction – create as usual	Link negative interaction to positive interaction Add the following annotation to the Negative Interaction: Annotation topic: negative: 'The interaction does not occur under these experimental conditions.'
To the **Positive** Expt add an X-Ref with Qualifier "SEE-ALSO":- DataBase = IntAct + Accession Number of the **Negative** Experiment	To the **Negative** Expt add an X-Ref with Qualifier "SEE-ALSO":- DataBase = IntAct + Accession Number of the **Positive** Experiment
To the **Positive** interaction Add an X-Ref with Qualifier "SEE-ALSO":- DataBase =IntAct + accession number of the **Negative** Interaction	To the **Negative** Interaction Add an X-Ref with Qualifier "SEE-ALSO":- DataBase =IntAct + accession number of the **Positive** Interaction

All Negative Reactions can be seen on the Web site by searching in Advanced Fields for "Negative Interaction".
Note that to see the cross-referenced interactions you will currently have to view the interactions in the Editor.

A few examples:

EBI-6173513, EBI-4567600, EBI-6354102 - free text comments on the entry

4.6.10 uniprot-dr-export

Databases such UniProt, GOA and neXtProt take binary interactions from IntAct and these are selected following a strict set of rules. First, experimental evidences of interaction A-B are merged. Then, the evidences for interaction A-B are then scored, and the interaction exported provided the score crosses a certain threshold, and the molecule pair obeys certain rules, including that the interaction evidences must also include at least one evidence of a binary pair, excluding spoke expanded and colocalisation data, before export.
If a curator is entering two molecules as an effective binary pair e.g. bait and one prey in a coip, but knows from other evidence (which may be in a separate paper) that a third 'bridge' protein is required for these 2 molecules to interact, the uniprot-dr-export annotation topic can be used to prevent these interactions being used as evidence of a binary pair. Any interaction which should not be used as evidence of a binary pair, should have annotation topic uniprot-dr-export added at the Experiment level, with added annotation text "**no**". This will probably require the creation of a new Experiment if there are other interactions in the publication where this is not an issue.

Examples:
EBI-6393946: apparent binary pairs according to the limits of this experiment, but the interaction is known to involve other proteins (so are not true binary interactions).
EBI-5326135: in the absence of b-catenin the reaction (of two interactors) does not occur.

This annotation need only be added if there are only TWO interactors. When there are >2 interactors, the expanded binaries will NOT be exported to UniProt, in the absence of any additional information.

[This annotation is not to be confused with the "No-UniProt-Update" annotation which is only applied at the participant level, and is added to synthetic protein constructs to stop the Protein Update attempting to map them (unsuccessfully) to UniProt.]

4.6.11 variable and variable_2

These annotations are added to an experiment to describe a set of interaction(s) with one/two variable parameters, for example, a kinetic experiment in which the concentration of a substrate or the temperature may be varied, or the effects of increasing concentrations of an agonist may be monitored over time or combinations of any of these.

EBI-1255485 variable: Time after treatment with 1uM Phorbol 12-myristate 13-acetate (mins)
EBI-5237718 variable: Time after Sendai viral infection (hours).
EBI-4372312 variable: Insulin (nM)
EBI-1570754 variable: The stage of the cell cycle

EBI-4413025 variable: Gamma-irradiation (Gy)

Examples of variable data as shown on the Public Web site: (Dynamic Interactions)
http://www.ebi.ac.uk/intact/interaction/EBI-6253708
http://www.ebi.ac.uk/intact/interaction/EBI-6264089

Other examples with 2 variables:

EBI-1775371 variables: Rapamycin (ng/ml) and Insulin (mM)
EBI-2620909 variables: Time (mins) and IL3 (ng/ml)

4.7 Experiment cross-references and cross-reference qualifiers

All added cross-references must have a qualifier, which usually needs be added manually.

Database	Qualifier
pubmed	method reference
pubmed	see-also
pubmed	source reference
intact	see-also

4.7.1 method

The cross-reference qualifier "method" should be used with database ‘pubmed’ and primary ID the PubMed number, where there is a relevant paper describing the experimental technique used. It should not be used for information about the interactions.

4.7.2 see-also

- Used as a qualifier to add additional database links (e.g. for internally created proteins) which cannot be added using another qualifier.
- Used to add to entries obtained from an external source such as a database which has an associated publication describing the resource. The PMID should be given the reference qualifier ‘see-also’. Example: EBI-866743.
- Used while referring to an IntAct database cross-reference to an experiment used to show a negative interaction. See section 5.6.9.

4.7.3 source reference

This cross-reference is added to entries obtained from an external source such as a database which had an associated publication describing the data. The PMID should be given the reference qualifier ‘source reference’. Example: EBI-866743.

4.7.4 target-species

This cross-reference is added automatically and relates to the organism/s of the protein/s involved in the interaction. This cross-reference does not need to be changed by curators.

4.8 Data not shown

Normally if an author makes a claim based on ‘Data not shown’ this is not curated. The exception to this is screening data, usually two hybrid or MS proteomics following affinity purification. This may be captured, provided both methodology and results are clearly described in the text.

5. Interaction

After adding an experiment, you can add interactions using the Interaction tab on the Experiment page. These interactions will be created primarily attached to the experiment it was created under, however, the experiment that the interaction is attached to can be changed later, if needed. Each interaction is linked to a single experiment (if an interaction is linked to several experiments – this is an error).
The interaction page records the information about an interaction and the various factors that have an effect on the interaction.
Many interactions may be linked to a single experiment if they are all using the same experimental techniques. If a particular interaction is demonstrated several times in the same publication using the same experimental technique it need only be entered once, with all instances listed in the figure legend. Whenever there is a change in any one of host organism, interaction detection method or participant detection method, a new experiment must be created, even if the same interaction has already been demonstrated in the publication. Slight variations in the use of agonists/antagonists/ knock-out cell lines do not necessitate making a new interaction.

Interactions should be entered as binary interactions or n-ary interactions, as demonstrated within the specific experiment. All the interactions within a publication should be entered if possible, including positive controls. However it may not always be possible to enter all interactions in a paper, for a variety of reasons – see Table below:-
Unidentifiable interactors should be listed under the annotation “Data-Processing” (please refer to this topic for further details)

Common Reasons Why Interactions Cannot be Curated
A participant(s) has only been partially identified e.g. ”actin”, “myosin”, PKA/C, AKT, Histone (type 1/2/3)
A “pan” antibody has been used, for example to multiple closely related antibody targets:-“ERK” “RAS”
Antibodies to Histones type H1/2/3 used – many of these antibodies cross-react with a variety of histone proteins or modified histones. This is often a problem with Ch-IP assays (beware antibodies to H3K4me1/2/3)
A significant proportion of the interactors cannot be unambiguously identified

5.1 Interaction AC

An IntAct accession number is automatically assigned to each of the interaction, components, interactors, features, cross-references and annotations.

5.2 Short Label

The usual rules for Short Label apply. The interaction short label is automatically generated, based on any two gene names within the interaction (bait is always first in a bait/prey relationship) – but can be subsequently over-ridden so as to more accurately

represent the interaction taking place. The short label is NOT a stable identifier for an interaction as it can change after any modifications.

5.3 Parameters

Parameters on interactions can act as holding places for kinetic values and values for physical properties of interactions, such as the Kd (see www.ebi.ac.uk/intact/interaction/EBI-9873160). The unit used for Kd data should always be Molar (M). Additional information about the kinetics value described in parameters, or kinetic values which not on the parameter list should be added in annotation topic 'kinetics'. The experimental conditions under which a kinetic parameter has been generated (temp, pH) should be entered under the annotation topic 'kinetic_conditions' In cases where one or more mutant have an effect on the kinetics and either decreases or increases the rate of the reaction, the mutant(s) should be annotated at the features level for the protein as feature type "rate increasing/decreasing". The corresponding parameter can then be added to this feature.

Only ONE value per parameter entry should be added at the Interaction level, which should always been the wild-type (when given). Additional values of, for example, Kds obtained using multiple different mutations of one protein can be entered onto the associated Feature for each mutation or as a "Comment", if this is not possible.
However multiple (different) parameters can be added at the Interaction level.

5.4 Interaction type

The interaction type defines the type of interaction the molecules make, as demonstrated in each specific interaction.

Commonly used examples: association, physical association, direct interaction, colocalization.

This is a mandatory field.

5.4.1 Association

An association is an interaction in which more than one n-ary complexes are represented, assuming complexes as groups of molecules that are physically associated together at the same point. As a rule of thumb, any interaction with the Interaction Detection Method = affinity chromatography or one of its children (e.g. coimmunoprecipitation, pulldown, TAP) with 1 bait and MORE THAN 1 prey should be mapped to association, since more than one protein complex could be represented in the interaction.

No other experimental data has interaction type = association.

5.4.2 Physical Association

Any interaction in which the participants form a single n-ary complex (identified with techniques that can ascertain that only one complex is present, like x-ray

crystallography), or a binary complex which may involve more interactors than those identified in the experiment should be described as ‘Physical Association’. Any interaction with the Interaction Detection Method = ‘affinity chromatography’ or one of their children (e.g. coimmunoprecipitation, pulldown, TAP) with one bait and only ONE prey should be mapped to ‘Physical Association’ (unless the interaction occurs in vitro and consists of only 2 highly purified proteins in which case they should be mapped to ‘direct interaction’ (see below).

Examples:
In vivo IP with 1 bait and 1 prey
Yeast-2-hybrid assay
n-ary crystal
n-ary gel filtration

5.4.3 Direct Interaction

“Direct interaction” should only be used when there is no possibility of a third, unseen or ancillary molecule acting as a bridge between the two molecules of interest (known small molecules bridging much larger molecules are an exception to this rule).
NB: NO interaction occurring “in vivo” – in a cell or cell lysate – can be called “direct”.
An interaction should ONLY be described as direct if:

a. the number of interactors exactly equals 2 highly purified molecules (plus optional small molecule), or a homomultimer and
b. the host organism is "in vitro".

The "Experimental Detection Method" may be one of many types – most commonly biophysical methods such as X-Ray crystallography, enzyme assay or affinity chromatography methods.

5.4.4 Self or Putative Self Interaction

These are only used if 2 regions (each purified separately) of the same protein are interacting in vitro 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 and experimental role ‘self/putative self’ will then be used. 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.

5.4.5 Colocalization

Experiments which describe the purification of complexes where interactors may not interact (e.g. cosedimentation) or imaging studies (e.g. confocal/fluorescence microscopy) where it is difficult to assess if the interactors have a physical association must have the interaction type set to colocalization.
Gold immunostaining data (using Electron density methods/ electron microscopy) is acceptable as co-localization data when there is a clear distinction between the labelled particles, usually on the basis of size.

Colocalization between molecules will be captured only when, within the same paper, there is a physical interaction evidence.

Colocalizations where one of the proteins is only used because it is an indicator of a particular subcellular location, are NOT captured.

** Where the interactors co-occur in the same subcellular compartment these can also be annotated using the interaction type: colocalization and the appropriate GO term for the subcellular location of the colocalisation be entered as an Xref.

5.4.5 Proximity

Experiments which describe the co-occurrence of molecules within very close proximity (in the nanometer range), detected through molecule-level resolution methodology, but from which a physical interaction among those molecules cannot be inferred. Examples include FRET, PLA and BioID

5.5 Causal Interaction annotation (NA-MIMIX)

- See end of the document

5.5 Interaction Annotations (NA-MIMIX except Figure legends)

Any publicly available specific instances of “Interaction annotation” can be searched for in the public database using the “advanced_search”, selecting for the field “Interaction annotation” and entering the specific term needed e.g. “isoform-comment”.

5.5.1 Commonly used Interaction Annotations

Annotation	Definition/Comments	Examples
antagonist	Any molecule applied externally to cells or any type of environmental condition, such as hypoxia, that inhibits an interaction, potentially by alteration of amount or binding affinity of one or more of the interactors.	EBI-6397393: Neratinib, an irreversible inhibitor of EGFR/HER kinases. EBI-2879898: SB216763 - a GSK-3 inhibitor.
agonist	Any molecule applied externally to cells or any type of environmental condition, such as hypoxia, that stimulates an interaction, potentially by causing modification of one or more of the interactors. [Further description of the agonist should be included, if necessary]	EBI-6397393: PD0325901 – A MEK inhibitor. EBI-6395914: Amino acid starvation. EBI-6127767: Thapsigargin (an irreversible sarcoplasmic/endoplasmic reticulum Ca2+-ATPase)
3d-structure 3d-r-factor 3d-resolution	These terms are used in the curation of crystallographic data – see later	

comment	Add any info here that cannot be adequately described under any other type of annotation category	
inhibitor	Use to capture information on the action of a molecule which is directly inhibiting an interaction but you do not feel appropriate to add as an interactor. Usually only used with purified proteins where the inhibited interactor can be identified.	EBI-447320: This interaction was disrupted by chetomin EBI-447720 - Inhibition by autophosphorylation of IRAK-1 at undefined position(s).
kinetics	Use to capture additional kinetic data (such as effect of a mutation or a kinetic value not available as a Parameter). The units used should always be stated, and must be applicable to the kinetic value being given. Use SI units wherever possible.	EBI-591030: The affinity between TGN38 cytoplasmic tail and the mutant Y350A adaptin mu2 mutant is kd = 7.55 10-7 M EBI-519455: FIST protein which had a truncated kinase domain (142- 433) displayed reduced binding affinity for Daxx in this study.
resulting-ptm	Previously used to record the addition of a PTM on a participant as the result of an enzymatic reaction or child thereof. This should now be added as a feature role	EBI-2941958: Tyrosine phosphorylation of EGFR EBI-4479606: Sirt1 reduced the acetylation level of GST-c-Myc in the presence of NAD+. EBI-3890310: mono-methylation of RelA (Lys310me1).
stimulant	Use to capture information on the action of a molecule which is directly stimulating an interaction but you do not feel appropriate to add as an interactor. Usually only used with purified proteins where the stimulated interactor can be identified.	EBI-515302 Cross-linking of GPIV (Q9UIF2) in platelets increased associated FcR gamma chain (P30273).

The annotation comments Complex-properties, Curated complexes and Complex synonyms are only used for the curation of curated complexes and should not be used for experimental interaction evidence curation.

Rules for annotating Agonist/Antagonists:

- These are small molecules, or proteins or conditions like hypoxia, radiation, etc.
- siRNA without therapeutic significance should not be considered as agonists/antagonists.
- We don't have to add concentration or duration of exposure.
- Recommended format:

- **For small molecules,** use ChEBI: ChEBI name (ChEBI identifier) - Eg. 17β-hydroxy-5α-androstan-3-one (CHEBI:16330)
- For therapeutic antibodies: use ChEMBL:
- **For proteins** - UniProt gene name (UniProtKB identifier) - Eg. IL6 (UniProtKB:P05231), IL6 (Species Unknown)
- **When there is combinatorial usage** of agonists or antagonists use the + symbol. Eg. 17β-hydroxy-5α-androstan-3-one (CHEBI:16330) + IL6 (UniProtKB:P05231)
- **For physiological conditions**, use EFO X'ref: Eg. Radiation exposure (EFO:0020980)

5.5.2 figure-legend (MIMIX applicable)

This is the figure number in the paper where the interaction was shown. It is however essential that this be added where available.

Examples:
figure-legend: 1
figure-legend: 2a and 2b
A Figure-legend need NOT be added for screening data such as two hybrid or mass spectrometry, if none is given in the publication.
If the data is (also) given in a table, enter the table number instead of or in addition to the figure legend.

5.5.3 negative

This must be added on the entries made for negative interactions. See section 4.6.9 for usage.

5.5.4 uniprot-cc-note

This annotation is added to specify the text which is exported into UniProtKB as a "note" in the CC INTERACTION block. This is not currently exported for public viewing. Notes should be added using Swiss-Prot syntax.

5.5.5 url

The url of an external database from which the interaction data originated. This is generally only added to submitted data at the request of the submitter.

Interaction AC	Description
EBI-308790	http://www.kazusa.or.jp/tech-cgi/tablelist.ppi.cgi

5.5.6 3d Structural data

Crystals

There are 3 extra Interaction annotations, '3d-resolution', '3d-r-factors' and '3d structure'. Add the PDB accession code e.g. '3S6N' as xref, with the qualifier 'Identity'. This can usually be found in the publication, or can be searched for on the PDB site using the PMID.
The PDBe web site can be found at: http://www.ebi.ac.uk/pdbe/
Experimental roles are "Neutral" for crystals. Host Organism: 'in vitro' – (with "expressed-in" for participants).

Many details regarding the interaction can be found in the relevant PDBe entry, notably R Factors, resolution, mapping to UniProt, sequence numbering, presence of ligands etc. Binding regions can be LINKED.
Stoichiometry can be found under "Structure analysis"
"Assembly composition" was included in the PDBe import but is generally not added to manually curated entries.

Example entry:-

6.5.5.1 3d structure

This annotation topic can be used to comment on the structure of the 3D-complex but is optional.

6.5.5.2 3d-r-factor

This annotation topic should be used to denote the r-factor of the structure.

Interaction AC	Description
EBI-297231	working 19.1%, free 20.6%
EBI-539447	working 25.0%, free 29.6%

6.5.5.3 3d-resolution

This annotation topic should be used in conjunction with crystallographic data to denote the resolution of the structure. Units in angstrom (A) should be used; the text field however does not allow the entry of the superscript hence enter it as shown in the Table.

Interaction AC	Description
EBI-449117	1.46A

5.5.6 variable_condition_1/2

Describes the parameter varied throughout a set of interactions and should only be used when the annotation topic 4.6.11 “variable” has been added onto the associated **experiment**. Example: EBI-1255485.

5.6 Cross-references and cross-reference qualifiers

All added Cross references must have a qualifier: For the GO database the qualifier is added automatically, others must be entered manually.

A GO annotation(s) is normally added for the cellular location of an interaction for co-localisation experiments. For non-co-localisation experiments it can be added when an interactions occurs in a particular **CELLULAR COMPARTMENT**, e.g. nucleus.

A PDB cross reference is always added for crystals.

Note: References relating to the **protein** should be added on the **Participant** level using the Feature editor (e.g. InterPro Cross-Reference);

Databases:	**Qualifiers**
pubmed:	see-also
go:	e.g. component, process, function
pdb:	identity, see-also
intact:	see-also, intact-secondary
reactome	identity, see-also
omim:	see-also
intenz:	identity

5.6.1 identity

This cross-reference qualifier should be used when the cross-reference describes an equivalent entry in another database. Example: For a PDB entry, the database is ‘wwpdb’, the ID is ‘PDB AC’ and the cross-reference qualifier is ‘identity’.

Example: EBI-297231

5.6.2 see-also

This cross-reference qualifier should be used when there is additional information in another IntAct entry or external database which would add value to this entry for the user. For example, there may be information in a second publication which is not captured here but is relevant to the entry, in which case the PubMed xref should be given the qualifier 'see-also'.
When the same interaction is described in other databases but not in as much detail, cross-reference with the qualifier see-also.

5.6.3 GO cross-references

GO xrefs are added at the interaction level to give additional information about the subcellular location where an interaction occurs, and on enzyme assays to indicate the function and process demonstrated by that assay. The qualifier need not be added – that is added automatically on hitting Save. Under some circumstances the qualifier is not imported correctly or not imported at all and may need to be changed manually. Please always check the accuracy of the qualifier after saving.
If a required term is not available please request the term at:

https://github.com/geneontology/go-ontology/issues

5.6.4 intact-secondary

intact-secondary will be used as a qualifier on accession numbers associated with withdrawn interactions that have been remapped.

6. Participant

The **participant** defines the **interactor** construct when it enters an interaction.

Participant = Interactor + Features.

The feature imparts additional properties to the interactor.

Proteins, small molecules, nucleic acids, complexes and genes are currently interactors in IntAct.

6.1 Proteins

On entering a new UniProtKB AC (Accession Number e.g. **P01308**) into IntAct, a participant entry is created. Information such as the ID, UniProtKB entry name (e.g. INS_HUMAN), Gene Name, Organism and sequence are imported directly from the UniProtKB entry. An Intact AC number is assigned to every interactor. The Intact AC number can also be used to retrieve a protein entry.

6.1.1 How to find a UniProt Identifier for your proteins

Proteins are preferentially mapped to UniProtKB/SwissProt entries. If no UniProtKB/SwissProt entry is available, the mapping is to the longest UniProtKB/TrEMBL entry. Searches are performed for UniProtKB IDs based on the gene or protein name, identifier, cross-references, accession numbers or sequence provided by the authors.

(Hint - if the UniProt Database is down or slow, try replacing “www” in “http://www.uniprot.org” with “ebi”,“sib” or “pir”, for local access.)

The “ID mapping” service at https://www.uniprot.org/uploadlists/ can be available to obtain the UniProtKB identifiers should only non-UniProt identifiers be provided by authors (note -proteins with GI numbers as identifiers can only be cross referenced if the “GI” prefix is removed). Author given names which cannot be traced in UniProt can be added as an alias on the participant.

If your publication only provides exon info for your protein, find the Gene in Ensembl (Cross-Refs) and click on the appropriate TRANSCRIPT (compare aa length). Click on “PROTEIN link” at TOP LHS (left hand side) and this will give the amino acids from each exon in alternating colours. If necessary, you can blast the amino acid sequences to get the numerical range.

If the exact strain of the organism from which you protein of interest is missing, but there is a Reference Proteome for that species in UniProtKB, map to the Reference Proteome. To find the Reference Proteome, select the UniProt Taxonomy search menu and enter your organism e.g. Staphylococcus aureus. Use the ‘Advanced Search’ to filter the result

by ‘Reference Proteome (yes/no), and choose ‘yes’. This should be “strain NCTC 8325”, which has a TAX ID of 93061.
Proteins can be found for this organism by searching for the Name of your protein **and** the TaxID. E.g. “Nitroalkaneoxidase AND 93061” gives a UniProt ID of Q2FZX9.

6.1.1.1 Tracing the Species of origin of a protein

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

a. Reference chasing (potentially going back many years)
b. Curator writes to the author, and receives additional information (or not).
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. (The Addgene (non-profit) repository has a large collection of donated plasmids, that names depositors, and it is worth checking at http://www.addgene.org/ if your author(s) has donated plasmids here.)

In all cases, the decision tree used to assign a database accession number to the protein should be documented as a Data Processing comment.

6.1.1.2 Identification of naturally occurring fusion proteins

Examples: BCL-ABL1, MLL/AF4etc**.**
Interactions involving this type of protein are commonly found in cancer related journals. The exact identity or sequence of the fusion protein(s) is rarely given by the authors. In this case choose a UniProt entry best representing the features you wish to show and add a Data-Processing comment to the effect:
The fusion protein "BCR-ABL" has been entered as TrEMBL entry A9UEZ6. This is a reference entry only and may not be an exact match to the sequence of the protein mentioned in this publication.

6.1.1.3 Artificial fusion proteins

Interactions involving artificial chimeric/fusion proteins that are of NO biological relevance need NOT be curated, e.g. PMID: 22939624, chimeric kinase constructs in which one loop region/N-lobe was fused into another molecule – or domains reciprocally switched between molecules.
You may however choose to additionally annotate artefactual chimeric molecules when the sequence is available or can be reconstructed from the paper, but in most cases, this data will NOT be curated.
If an organism database or source database for the interactor identifier is not in the PSI CV and you need to cross-reference it, please request a term at the PSI-MI CV GitHub site by creating an issue (https://github.com/HUPO-PSI/psi-mi-CV/issues).

6.1.1.4 Identifying a protein by sequence searching

Where only sequences are available, Blast should be used for protein and/or short peptide matching. The best match showing a 98% sequence identity (or greater) that extends over 98% of the query or matched sequence should be used for annotation.

If there is a database identifier with a sequence attached to it but NO UniProtKB entry can be found after ID mapping, carry out a blast sequence alignment. If no sequence is seen to match any UniProtKB entry, create a new protein in the editor with a database cross reference, setting the database identifier qualifier to 'Identity'. A Comment should be made as to the UniProtKB orthologue in the nearest related species to enable manual mapping in the future. The sequence must be entered.

If a published protein sequence cannot be traced to a UniProtKB entry, the sequence information can be deposited at
http://www.ebi.ac.uk/swissprot/Submissions/spin/index.jsp
In the case of Mass Spec Data, contact Kate Warner, email: kwarner@ebi.ac.uk
Request for TrEMBL entries to be merged as isoforms into a UniProtKB?Swiss-Prot entry should be to http://www.uniprot.org/update
If the protein is found only in UniParc, contact Sandra Orchard
(Orchard@ebi.ac.uk)
If all else fails contact: help@uniprot.org

When a protein cannot be assigned from information in the paper or references, the author should be contacted to supply the missing information. If there is no response after ~ two weeks, curate all possible interactions in the paper, and make a note of missing proteins under the heading "data-processing". If the author contacts the database following release with the missing information the entry should be updated.

Where the peptide mapping by the author to the entries is problematic or the match of a mass spec peptide sequence to the identified protein is not 100 %, this information should be added to the data-processing and the interactions involving this protein should not be entered. This is one instance when there will be a discrepancy in the interaction numbers provided by the author and the ones in IntAct in a large-scale experiment.

6.1.2 Mapping of protein isoforms

If an author states that a specific isoform has been used or identified in an interaction, this information should be included. Should the isoform be identifiable within an entry the interaction should then be mapped to this UniProt AC, for example PMID:11517249 uses murine Jip1 isoform b which is described by Q9WVI9-1. If there is no attempt to identify a particular isoform, the parent/canonical UniProtKB AC should be used. If the authors list an isoform which cannot be identified from those currently described in the corresponding UniProtKB record, an "isoform-comment" should be added as an annotation.

However, if a sequence is given for this isoform, which would indicate it is either a novel sequence, not already in UniProtKB/Swiss-Prot or the required sequence is in TrEMBL, contact http://www.uniprot.org/update and request the isoform be added to the Swiss-Prot entry. If you curate to the TrEMBL entry, the isoform information will be lost when the entry is merged into Swiss-Prot as the interaction data is transferred to the canonical sequence. This will be particularly confusing if there are both positive and negative isoform interactions for the gene products.

Whilst waiting for the Swiss-Prot update, you may create the new isoform as an internal protein with a “no uniprot update” annotation - leave this annotation in place until the new isoform has been released (from UniProt) or the new isoform info will be lost on any IntAct protein update prior to that event. This should also be accompanied by a remark-internal ‘Update when UniProt AC available’

The update will take ~8 weeks to work through the UniProt release system.

Example:

Create an internal protein
Add identifier from UniProt (e.g. p07948-n) as the Short Label (n should be a number not yet used by UniProt - if in doubt use 99)
Copy the correct isoform sequence from UniProt into the "Sequence box"
Add a Uniprot xref to the isoform qualifier = identity and IntAct xref to the parent UniProt entry - in the example below to P07948:

Database	**Identifier**	**Qualifier**
UniProtKB	P07948-n	identity
IntAct	EBI-79452	isoform-parent

Sequence features, such as mutations, within a specific isoform should be mapped to the sequence of that isoform where it differs from the sequence as given in the canonical entry. If an isoform is shown to bind to an interactor and another isoform of the same interactor is shown not to bind, a negative experiment/interaction should be entered (see 4.6.9).

6.1.3 Problems with mapping accession numbers/Special or Unusual situations

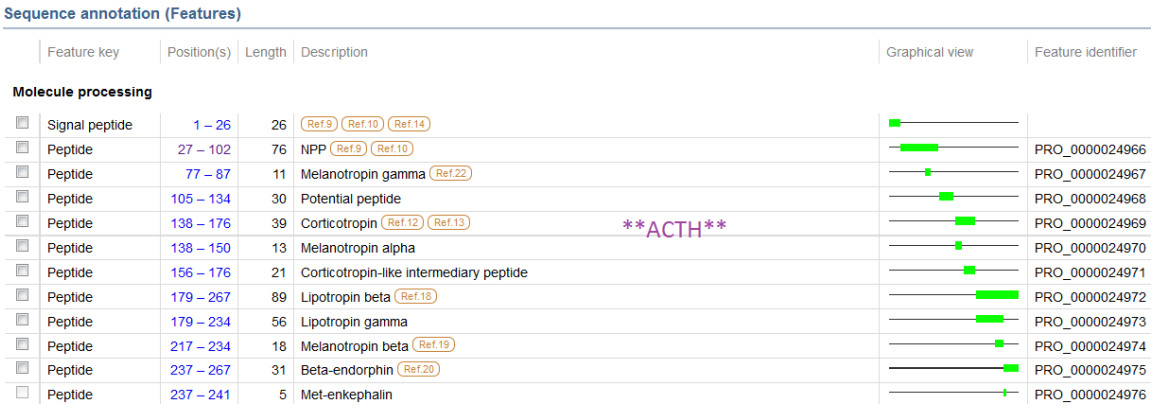
Ubiquitin: Ubiquitin is encoded by 4 different genes (in mammals) and cleaved from a larger precursor protein. If the Ubiquitin species is known, but not the exact origin, it can be entered as a pre-existing IntAct protein “ubiq_human” or ubiq_mouse etc…

IPI (deprecated) database isoform identifiers: do NOT enter isoform information **unless** peptide evidence is available to show that a specific isoform is present.

6.1.4 UniProtKB entries that contain multiple protein molecules

All molecules in UniProtKB/Swiss-Prot have the Feature ‘chain’ with a non-redundant PRO ID assigned to it. These indicate proteins that have been cleaved into one or more smaller proteins or bioactive peptides e.g. POMC – (UniProt Human entry: P01189). This precursor polyprotein is enzymatically split into 11 smaller proteins/peptides including ACTH and MSH.

Each of these peptides has a distinct identifier in UniProtKB, known as a “Feature identifier” and this ID can be found in the parent UniProt entry under “Sequence annotation (Features)” e.g. for ACTH the identifier is PRO_0000024969.



Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing						
☐	Signal peptide	1 – 26	26	Ref.9 Ref.10 Ref.14		
☐	Peptide	27 – 102	76	NPP Ref.9 Ref.10		PRO_0000024966
☐	Peptide	77 – 87	11	Melanotropin gamma Ref.22		PRO_0000024967
☐	Peptide	105 – 134	30	Potential peptide		PRO_0000024968
☐	Peptide	138 – 176	39	Corticotropin Ref.12 Ref.13 **ACTH**		PRO_0000024969
☐	Peptide	138 – 150	13	Melanotropin alpha		PRO_0000024970
☐	Peptide	156 – 176	21	Corticotropin-like intermediary peptide		PRO_0000024971
☐	Peptide	179 – 267	89	Lipotropin beta Ref.18		PRO_0000024972
☐	Peptide	179 – 234	56	Lipotropin gamma		PRO_0000024973
☐	Peptide	217 – 234	18	Melanotropin beta Ref.19		PRO_0000024974
☐	Peptide	237 – 267	31	Beta-endorphin Ref.20		PRO_0000024975
☐	Peptide	237 – 241	5	Met-enkephalin		PRO_0000024976

This “Feature ID” gives the exact sequence present in the molecule and is linked to the parent sequence. Naturally occurring peptides, processed peptides and polyproteins all have these “Feature ids” and these can be used to identify these molecules.

When entering a post-translationally processed chain, such as a viral protein or bioactive peptide cleaved from a longer precursor molecule, in the IntAct Database editor, enter a “FeatureIdentifier” into the import tool – NOT the parent UniProt identifier.

Example: 1. NFKB1
NFKB1 (KBF1) is in UniProtKB as accession number P19838

NF-kB1/p105	P19838	full length
NF-kB1/p50	P19838	from N term to end of glycine rich region range 1-433 has identifier P19838-PRO_0000030311

The processed form (the p50 subunit) of NFKB1 does not have a separate entry to the full-length protein in UniProtKB, but is differentiated by the PRO identifier.

Example: 2. Polyproteins
Polyproteins are proteins that are synthesized as a single polypeptide and then cleaved into several distinct proteins.
e.g. viral HIV-1 **Gag-Pol polyprotein Q73368**
Integrase has a Feature ID of PRO_0000250983 – typing this into the import tool of the editor will create the correct protein participant.

6.1.5 Synthetic Peptides

A chemically synthesised peptide can be added as an interactor in a variety of ways:
If the peptide maps to an existing UniProt ID, this can be entered and the exact amino acid sequence numbering entered as a Feature of the type “Sufficient to Bind”. The “Biosource” will be entered as “Chemical synthesis” under “expressed in”.

If the sequence is very dissimilar to anything found in UniProt, it can be added as a “New Interactor” of the Type “Peptide”, and “Organism” = Chemical Synthesis. The sequence of the peptide should be entered in this instance.

Alternatively, and generally only when a modified peptide has been created as a chemical agent, a new ChEBI entry can be requested, and this ChEBI identifier entered as an interactor [see section 9.2].

Many synthetic peptides will have been synthesised based on an existing UniProtKB protein/peptide entry but with a few additional linker amino acids. These linkers can be ignored as they do not participate in the interaction, so the UniProt ID with the range added as a feature can be used as an ID (see above). Linker amino acids can be added as a ‘mutant’ or ‘tag’ if important.

If there is an ambiguity in assigning a species to a peptide i.e. the identical peptide is conserved across several species, the species can be entered as the same as that of the other interactor(s) in the specific interaction. Map the peptide to the appropriate UniProtKB entry and enter the range as a “Sufficient to Bind” Feature.

6.1.6 Proteins of unknown species

Interactions cannot be entered unless we can be certain of the origin of the proteins used and can find a database entry. If the paper does not state the species of the protein, it cannot be traced by the curator, and there is no response from the authors of the paper then we will have to disregard these data. In this instance a Data-Processing comment should be added at either Publication and/or Experiment level stating that the proteins cannot be identified and why.

6.1.7 Protein update

This is a 4-weekly re-mapping of IntAct proteins to updated UniProt entries.
UniProtKB may update the proteins, merge, or de-merge entries, and assign them new accession numbers or some of the proteins may be retired because of changes to the underlying gene models. These are taken care of during the scheduled UniProtKB update. The retired proteins are identified, and where possible remapped. Where this is not possible, the protein is retained and an attempt will be made with each update to complete the remapping – the interaction will not be lost. When a TrEMBL entry is merged with, or itself becomes a Swiss-Prot entry, the data is mapped to the canonical sequence, even if the transcript the data was originally mapped to becomes an isoform. This could lead to the loss of interesting information, so curators are encouraged to request isoforms only present in TrEMBL to be merged into Swiss-Prot entries prior to commencing curation. Where the update results in two participants being identical these are merged in an interaction and the interaction may consequently have fewer participants. This will be recorded on the interaction, or should be done manually by the

curator if observed when first creating the entry. The ranges – where assigned – are also checked for their validity and a range update is carried out with respect to the newer accession.

The schematic of the update procedure can be found at https://drive.google.com/file/d/0B3S9Q1JQ2DygZTFkMGZiMDgtMDk0Ni00OTAzLWEzZjktMjEwNDA2OGQ4NzMx/view.

6.2 Small Molecules

IntAct uses the Chemical Entities of Biological Interest (ChEBI) database definition of a small molecule or molecular entity. ChEBI is a freely available dictionary of molecular entities focused on 'small' chemical compounds. The term 'molecular entity' refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The molecular entities in question are either products of nature or synthetic products used to intervene in the processes of living organisms. All small molecules should have a ChEBI xref.

Interactions involving small molecules will be annotated wherever possible, but it is not our usual policy to include papers which describe only interactions involving these molecules, unless they are received as a direct submission. Data will only be added when sufficient detail is available to be sure that the small molecule is directly binding to the molecule in question, for example molecules present in a buffer will not be added even if the author states they play a role in the interaction, unless experimental evidence of this is given in the paper.

1. Crystallisation papers – For a small molecule ligand to be seen as binding within a crystal it must be within 3.5Å of the protein, and should not be included in crystals produced at lesser resolution. Modelled ligands should not be included. Care should be taken to separate out binding molecules from crystallisation solution contaminants – normally this should reflect the detail in the paper rather than the wwPDB entry.
2. Cofactors – these will only be added if appropriate data is available to be confident that the molecule is binding to a protein i.e. titration curves, competition analysis etc.

Protein/peptide derivatives do not count as small molecules in the ChEBI definition and we should attempt to map these to parent proteins as tagged or fusion molecules when possible. Example: Experiment EBI-371686 and Interaction EBI-371689.
In this experiment, purified TBP was bound to a TATA box sequence, and then purified Mot1 was added. The interaction between Mot1 and TBP was detected by the change in DNA anisotropy when Mot1 was added to TBP pre-bound to DNA.

6.3 Nucleic Acids

6.3.1 DNA

IntAct policy is only to annotate interactions involving DNA when either the sequence is given in the publication or an exact match can be found in the Ensembl or RefSeq databases.

Unlike for any other interactor type, features are not to be referenced using the internal coordinates of the referenced entity, but using genomic coordinates, since every piece of DNA is actually part of a long chromosomal thread.

The annotation must adhere to the following rules:

I. If no sequence-related information is available and no Ensembl or RefSeq ID can be attributed to the interactor, it cannot be curated.
II. If no sequence-related information is available:
    A. If the DNA entity can be attributed to an Ensembl identifier, use this to import the interactor through the editor as for proteins.
    B. If no Ensembl ID is available, but a reasonable RefSeq entry can be identified, create an internal interactor using the RefSeq sequence and cross-reference it to RefSeq with an ‘identity’ qualifier. As a rule of thumb, RefSeq can only be used if reasonable efforts to map to Ensembl have failed and no sequence is available.
III. If sequence-related information is available:
    A. Full sequence information: BLAST/BLAT it against Ensembl:
        1. >98% identity, 100% coverage of Ensembl entry: use Ensembl ID to import interactor. Deviations from the reference sequence in cases of not 100% identity can be captured as ‘mutation’ upon curator judgement.
        2. >98% identity, partial coverage of Ensembl entry: use Ensembl ID to import interactor and add a ‘sufficient to bind’ feature with undetermined range at the participant level. This feature needs to be annotated with a ‘genomic coord’ annotation that provides the genomic coordinates for the aligning region. Deviations from the reference sequence in cases of not 100% identity can be captured as ‘mutation’ upon curator judgement.
        3. <98% identity or no match to Ensembl entry, hits a genomic position(s) with >98% identity and it is identified in the paper as a regulatory region/promoter for a given gene: as before, use Ensembl ID to import interactor and add a ‘sufficient to bind’ feature with undetermined range at the participant level. This feature needs to be annotated with a ‘genomic coord’ annotation that provides the genomic coordinates for the aligning region. Deviations from the reference sequence in cases of not 100% identity can be captured as

'mutation' upon curator judgement. The sequence provided in the paper can also be captured as 'original sequence' at the feature level. A cross-reference to the 'promoter' term in the Sequence Ontology/SO (SO:0000167) or equivalent is also required.

4. <98% identity and no match to any genomic reference, not even suggested in the paper: Create internal interactor providing the sequence.

B. Partial sequence information:
   1. Genomic coordinates: If the paper provides a link to a given gene, please import it from Ensembl and add a sufficient to bind feature stating the genomic coordinates. If no link to a given gene is provided, please extract the sequence from the corresponding genome build and proceed as stated in point III.A.
   2. Probes/primers: Please attempt to map probes/primers to genomic sequences and coordinates. If that is possible, then proceed as in point III.B.1. If that is not possible, map the interactor best you can with the information available and add a 'sufficient to bind' feature with 'probe sequence' annotations where the probe/primers sequences are given.

6.3.1.1 Curating ChIP experiments

ChIP experiments are the most common case where gene participants are found, so here is a quick guide on how to represent the gene element of the interaction following the rules stated above.

ChIP experiments usually monitor the binding of a regulatory protein (transcription factor) to a regulatory element in the genome (promoter or equivalent). In these cases the gene is usually identified in the manuscript and an Ensembl identifier provided. The promoter region coordinates are rarely stated, but often one can find primers provided in the methods section.

In the most common case, these are the steps to take:

1. Identify the gene in Ensembl and import it using the editor as usual
2. Create a 'sufficient to bind' feature for the amplified DNA region, with undetermined ranges. Shortlabel should be meaningful and state if this is a promoter region or any other significant regulatory element.
3. If genomic coordinates are available, add a 'genomic coord' annotation at feature level and put them in. The underlying sequence can be added as 'original sequence' at feature level.
4. If primers are available, BLAT them to Ensembl's reference genome and put the genomic coordinates in as indicated above.
5. If the primers do not map to the reference genome, please add a 'probe sequence' annotation and give the primers' sequence there.
6. Finally, add a cross-reference to the Sequence Ontology/SO (www.ebi.ac.uk/ols/ontologies/so) if the binding region is identified by the authors as a promoter (SO:0000167) or regulatory region (SO:0005836 or specific term, see ontology).

6.3.2 RNA

For interactions involving RNAs, non coding RNAs entries are created in IntAct using RNA central ID.
mRNAs entries are created using Ensembl transcript ID plus RefSeq, in case no Ensembl transcript ID is not available.
For mRNAs, if the identity of the specific transcript for a given gene is unclear, the reference transcript for that gene should be used if possible. Manually selected reference transcripts for human and mouse, the most problematic species, according to Ensembl can be found on the GIFTS project website (www.ebi.ac.uk/gifts), marked with a blue sign ('Ensembl select').

6.4 Macromolecular complexes

Please use Complex Portal identifiers (www.ebi.ac.uk/complexportal) to represent complexes in IntAct. If a complex is missing, you can request it via www.ebi.ac.uk/support/complexportal.

6.5 Genes

Genes should be given an Ensembl/EnsemblGenomes XRef.

6.6 Experimental and Biological roles

The Protein roles are split into biological role and experimental role in PSI2.5. **See** www.ebi.ac.uk/ols/ontologies/mi.

Note that this refers to the role that entity plays within the specific experiment i.e. a protein will only be given the biological role 'enzyme' in an enzymatic reaction, not in two hybrid experiment.

6.6.1 Experimental role

The role played by the participant in the context of the experiment.

Experimental Role	**Comment**
Bait	Bait must always have one or many prey(s). If there is an interaction with prey(s) there must be a bait. The bait is in general the protein which is used to isolate other proteins which bind to it. The bait may be immobilised during an experiment, be tagged or have an antibody raised against the protein itself (anti-bait) to aid purification.

Prey	Protein(s) isolated by binding to a bait molecule.
Unspecified	The experiment is such that proteins should have a bait/prey relationship but the author has not specified which is which.
Self	Only proteins that have been proven to interact intra-molecularly should be given the role 'self' (Example: autophosphorylation or disulphide bond). Where the same protein has been modified in different ways, (Example: tagged with different tags) and these molecules are shown to interact, these are considered as non-identical subunits. They should be entered as separate proteins with the role neutral component or bait/prey as appropriate to the experiment and the tags used described as features of the protein. See 11.11 for further details
Putative Self	Where the interaction is shown to involve molecules of identical sequence and modifications but a possibility of inter-molecular interaction exist are given the role 'putative-self'. The term "putative self" is used more frequently than "self", as it is often not known if a molecule is interacting inter- or intra-molecularly or both.
Neutral	Used when there is no directionality (e.g. bait/prey) in the experiment – all molecules play the same role, for example when a protein complex is isolated by molecular sieving. Role neutral component is often observed with stoichiometry determination or oligomerisation studies. Unmodified identical proteins may be shown to self-associate using techniques such as density gradient sedimentation, non-denaturing gel electrophoresis, mass spec, crystallography, and gel filtration (size exclusion chromatography). These will be entered as neutral component components. The number of molecules interacting will be indicated by stoichiometry (section 5.11)
Acceptor fluorophore/Donor fluorophore	Used in conjunction with fluorescence transfer experiments such as FRET. Some of the common donor acceptor pairs are CFP/YFP, BFP/GFP, GFP/Rhodamine and FITC/Cy3. This method detects interaction by demonstrating proximity of 1-10 nm. In the case of BRET and HTRF assays there is a fluorescence acceptor and the protein coupled to the molecule giving the fluorescence is assigned the role fluorescence donor.
Fluorescence acceptor/donor pair	Used in conjunction with fluorescence transfer experiments when a single molecule is tagged with both tags to demonstrate an intra-molecular interaction

6.6.2 Biological role

This is used to capture the physiological role of an interactor in a cell or in vivo environment, which is reproduced in the current experiment.

Biological Role	Comments	Example
Self	Only proteins that have been proven to interact intra-molecularly should be given the role 'self' (Example: autophosphorylation or disulphide bond). Where the same protein has been modified in different ways, (Example: tagged with different tags) and these molecules are shown to interact, these are considered as non-identical subunits. They should be entered as separate proteins with the role neutral component or bait/prey as appropriate to the experiment and the tags used described as features of the protein. See 11.11 for further details	
Putative Self	Where the interaction is shown to involve molecules of identical sequence and modifications but a possibility of inter-molecular interaction exist are given the role 'putative-self'. The term "putative self" is used more frequently than "self", as it is often not known if a molecule is interacting inter- or intra-molecularly or both.	
Electron acceptor/donor	Use where the interaction involves a transfer of electrons.	
Phospho-donor/acceptor	Use when the interaction involves a transfer of phosphate groups. This term is NOT used for Protein Kinase Assays, but can be used for Phosphotransferase assays /bacterial 2 component transfer systems.	EBI-6429292, EBI-6429316
Stimulator	Use when a molecule increases an interaction by interacting with one or more of the participants and it is not essential for the interaction.	
Inhibitor	Use when the interaction detection technique indicated interaction between the inhibitor and inhibited molecule. This could be at a prey/enzyme target binding site or a site that modifies the binding site.	
Electron acceptor/donor	Use where the interaction involves an exchange of electrons.	
Enzyme/Enzyme Target	An **enzyme** must always have an **enzyme target** (and vice–versa). The enzyme (example: kinase) modifies at least one of the interactor/s and the target is the modified interactor. In the case of autocatalysis e.g. Autophosphorylation, the protein may have both roles indicated by using self/putative self. Trans-catalysis: 2 molecules Experimental role – neutral	

	Biological role – enzyme/enzyme target Autocatalysis (when proven): 1 molecule (unless differentiated by tags, mutants etc.) Experimental role – self Biological role – self Autocatalysis (uncertain): 1 molecule (unless differentiated by tags, mutants etc.) Experimental role – putative self Biological role – putative self	
Cofactor	This will normally be used in conjunction with enzyme/enzyme target for enzymes where the presence of cofactor has been shown to be required for the interaction.	
Competitor	This interactor binds to the bait molecule in competition with other prey molecules.	
Unspecified	Used when the biological role of the interactor is difficult to ascertain.	
Ancillary	Additional interactor within an interaction whose role is not clearly defined – used in conjunction with molecules which do have clearly defined roles	

6.7 Expressed-In (NA-MIMIX)

'Expressed-in**'** refers to the source/origin of the protein when this is not the same as the Host organism in which the experiment was carried out. If possible, this field should always be filled-in for '*in vitro*' experiments. Leaving the field on --select-- indicates a protein expressed by the system in which the interaction is being measured, i.e by the host organism in which case it need not be entered twice

.
The protein may come from any of the following sources:

Example: Protein A is a human protein expressed in baculovirus-infected Sf9 cells, then purified and mixed with Protein B which is an endogenous protein from a nuclear extract from HeLa cells
Experiment Host organism: *in vitro*
Interaction: Protein A 'Expressed In' Sf9 cells
(enter spofr-sf_9 -*Spodopterafrugiperda* insect cells)
Protein B: 'human-hela'

6.7.1 Heterologous protein expression

Heterologous expression means that a protein (or more usually, the corresponding cDNA) is experimentally added into a cell that does not normally express that protein. Proteins

may be given the additional annotation ‘Experimental Preparation’ and a choice from the drop-down menu e.g. ‘electroporation’.

6.7.2 In vitro transcribed and translated proteins

In vitro transcribed and translated are entered in the ‘Expressed in’ box as “in vitro”, for example the production of [35S]-Methionine radiolabeled proteins.

6.7.3 Synthetic peptides

Please see section 5.8.2.8. The ‘Expressed in’ is ‘chemical synthesis’.

6.8 Stoichiometry

Only integer values are accepted. These values are entered as stoichiometry on each individual participant. Where only two purified components have been shown to interact and the experimentally determined stoichiometry is reported as a fraction, approximate the stoichiometry of the interactors to the nearest integer. In the case of assemblies a real fractional stoichiometry may be observed between the components of the interaction, here the stoichiometry must be entered rounded up/down to nearest whole integer. The editor now has the capability of entering a stoichiometry range but in the case of experimental data, the min and max value should generally be identical.

Where a higher assembly is clearly indicated but the number of molecules involved in the complex is not clear, the interactor must be entered twice with stoichiometry of ‘0’ for each and a comment made on the interaction to indicate a higher order complex formation. Examples would be a complex eluting in void volume for a molecular sieving chromatography or an observed higher order gel pattern. Multiple interactions must be entered where the stoichiometry of individual interactors is different e.g. a protein has been shown to exist as both a homodimer and homotetramer.
If two 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 must be entered twice with stoichiometry ‘0’. The Experimental Roles of the interactors are ‘putative self’ and the Interaction Type ‘self interaction’.

Identical molecules but with differing tags should be treated as distinct participants and the stoichiometry of each separately entered.

Briefly: for oligomers/dimers etc.
If the number of molecules is KNOWN: enter molecule ONCE; stoichiometry = number of molecules.
For homo-oligomers, if the number of molecules is NOT KNOWN: enter molecule TWICE, stoichiometry = “0”.
If a dimer and a trimer are both formed, enter two separate interactions.

The fields "Min Stoich" and "Max Stoich" should have the same values entered – a range is only entered for Complex curation.

Additionally, the GO cross-reference 'protein homooligomerization' GO:0051260 can be added at the Interaction level.

6.9 Participant name alias

The author-identifiers, when different from the gene name or ID, may be added onto a participant as an 'alias' under "author given name". This is important in the curation of large scale datasets to keep a record of the mapping of the original data. This is a defined field and can be added using the drop-down menu in the tabbed field 'Alias'

6.10 Annotations (unique to Participant-level)

The annotations for individual participant properties should be added on the participant.

6.11 Experimental Preparation

This is the method by which a molecule is delivered or engineered into a cell. This annotation can include the terms electroporation, microinjection or nucleic acid delivery.

This is a defined field and can be added using the drop-down menu in the tabbed field 'Experimental preparation(s)'. Additional information may be added as a Comment.

It is **mandatory** to annotate (according to the case) information that cannot be inferred from the experiment itself:

'**expression level**' for protein expressed in living cells, fixed cells or in cell lysate:

-over expressed
-under-expressed
-physiological level

'**sample process'** that indicates the sample context in which each interacting molecule is presented to its partner.

- Cell lysate
- Subcellular preparation
- Fixed cell
- Purified
- partially purified
- In vitro translated protein
- Conditioned medium (important for secreted proteins or shed ectodomains from membrane proteins)

Experimental preparation general annotation rules

Expression level*:

- Physiological level
  - A molecule whose synthesis is under control of its natural gene promoter or estimated to be expressed at a similar rate.
  - Includes all the proteins that are expressed using the wt gene, promoter region, etc…
  - Include virus proteins in host cells
  - In case of Native promoter + stimulation, the curator should judge if use 'Overexpressed' or 'Physiological level' accordingly with the specific case and add a comment.

- Overexpressed
  - A molecule is estimated to be expressed at higher levels than in physiological condition.
  - Can generally be used in transient or stable transfection systems unless otherwise stated by authors.
  - Normally includes any tagged molecule and any protein which appear as transfected, although some techniques (e.g. CRISPR) allow for the creation of tagged constructs under control of the native promoter. In this case, the expression level is 'Physiological level'.
  - In case of native promoter + stimulation, the curator should judge if use 'Overexpressed' or 'Physiological level' accordingly with the specific case and add a comment.
  - Viral proteins not engineered are not to be annotated as overexpressed
  - It is not annotated when sample process is: in vitro translated, purified or partially purified

- Under expressed level
  - A molecule is estimated to be expressed at lower levels than in physiological condition.

Sample process*:

- cDNA library: This term should be associated only to nucleic acid interactors not to their protein products.
- Cell lysate: Usually associated to *in vitro* affinity chromatography (especially Pull Downs) techniques. To use also for *in vivo* affinity chromatography, such as pull down and COIP.
- Subcellular preparation: A procedure to extract a particular compartment, where the interaction has been detected. If the paper provide information about the compartment where an interaction or colocalization occurs, this information

should also be captured using appropriate GO Cell Compartment terms cross-referenced at the interaction levels.

- Fixed cell: Generally associated to imaging techniques. It must be explained in methods and in the text.
- In vitro translated protein: The fact that a protein is translated in vitro does not necessarily imply it has also been purified, so the use of this type of preparation should not be used to infer that the interaction is direct. Judgement must be made from the information presented in the paper.
- Living cell: Usually associated to Y2H techniques; Sometimes associated to in vivo imaging techniques. This term should be used only if the technique doesn't imply that is living cell, as for example, in vivo imaging techniques such as FRET, BRET and PLA. Judgement must be made from the information presented in the paper.
- Phage Library: A bacteriophage library of genes encoding proteins or peptides fused to a phage coat protein that are expressed on the surface of the phage virion.
- Purified: Include proteins labelled as 'recombinant'. Generally, well described in material and methods
- Partially purified: Includes immunopurified proteins and proteins whose purification process has not been exhaustively explained in text. Partially purified proteins **cannot** participate to 'direct interactions' or children.
- Conditioned medium: molecules present in media harvested from cultured cells.

*Important because it helps determine the 'Interaction type' (i.e. an interaction is 'direct interaction' or any child term only when both participants are 'purified' and or 'in vitro translated proteins').

Molecular source:

naturally occurring/engineered: Only add engineered when there is evidence the gene/protein has been tampered with.
Only use when there is a case of ambiguity, e.g. level of a physiological protein is increased by an agonist (over-expressed, naturally occurring), transfected molecule, no change made (engineered). Engineered need not be used when there is a clear change to a feature, e.g. deletion mutant, point mutant, tag.

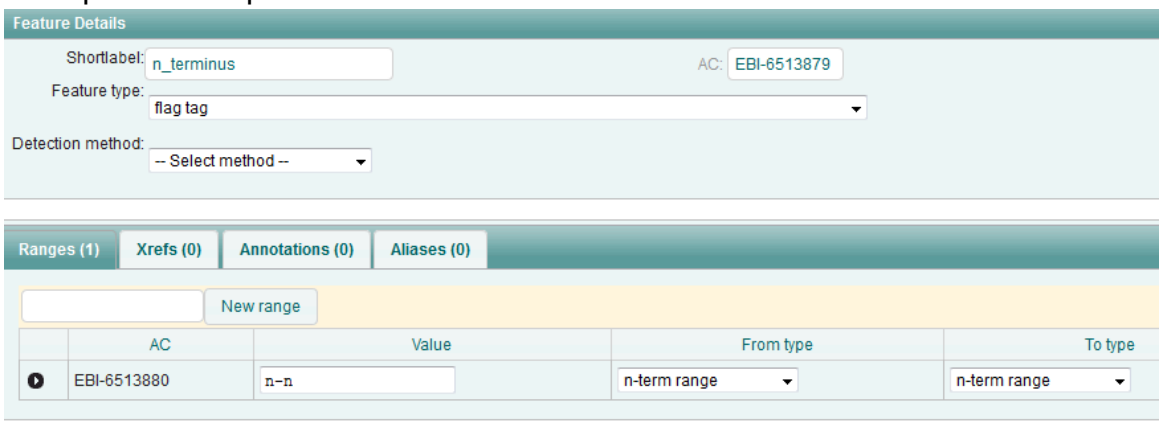
Definition of all terms related to experimental preparation can be found in OLS (Molecular Interactions): www.ebi.ac.uk/ols/ontologies/mi.

7 Participant Feature

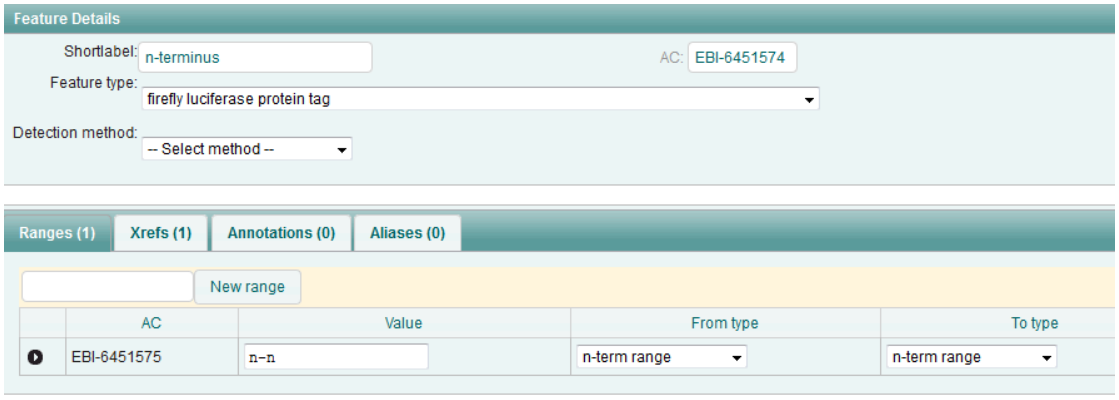
A feature is added on the participant to more fully describe the construct used in the experiment and will contain information such as binding region, mutation or deletions, tags.

On submitting the features, these will be visible on the interaction page. Linking of the features will have to be carried out from the interaction page.

Example of a simple feature:



A more unusual feature:



A Feature with “Fuzzy” type range: the authors have not given the range for the interacting region, so this has been deduced (generally from UniProt Sequence Annotation):

7.1 Fields required to describe a Feature

Short label (obligatory)	This does not auto-generate. It may be based on the InterPro name of a feature e.g. SH3 domain. The default term “region” is used if no more informative term can be found. Mutations are a special case, as described in section 7.3.5.
Feature Type (obligatory)	Use the most appropriate CV term to describe the feature or modification
Detection Method (optional)	e.g. deletion analysis
Range (obligatory)	Ranges can be specific or indeterminate and describe the position of the feature within the sequence of the molecule. A feature may have MORE than one Range.
X-Refs	When a range binding site feature can be matched to an InterPro domain, enter InterPro Xrefs at the feature level: these can be found from the UniProt entry page under Cross References/InterPro/ Graphical view. The IPR number is entered with qualifier “identity” (Example IPR000001 for Kringle)

7.2 Examples of various features

Feature type	**Feature AC**	**Short Label**	**Range**	**Feature detection**	**Xref (qualifier)**
required to bind	EBI-6128250	region	51-372	deletion analysis	
sufficient to bind	EBI-4373855	fadd-dd	93-191	-	IPR011029 (identity)
mutation decreasing	EBI-922561	p.Tyr39Phe	39-39	-	
opthr (Phosphothreonine)	EBI-6162530	thr-33	33-33	-	-
35s radiolabelled	EBI-2911435	region	?-?	-	-
flag tagged	EBI-2616889	n-terminus	n-n	-	-

tagged molecule (non-standard tags)	EBI-627 0838 EBI-625 0853 EBI-437 0535	sumo tag halo_biotin_tag s-peptide tag	?-? c-c n-n	-	

Note: binding site has two children, 'sufficient to bind' (binding observed with construct shorter than full-length molecule) and 'required to bind' (binding not observed with region deleted from full-length construct.
If amino acids are sequentially mutated to alanine, the feature detection method is alanine scanning.

Features can also be added to nucleic acids and genes. Gene features are currently not mapped to sequence, use range= ?-?

Some examples of Nucleic Acid Features

Short label	**CV term**
promoter_region	Binding region
tp53 response element region	Binding region
hypoxia_response_element_promoter_region	Binding region
promoter_cpg_island	Binding region
promoter vdre (vitamin d response element)	Binding region
transcription start site	Binding region

Feature Type: RNA chemical modification id MI:2202
for Ribonucleotide modifications (such as methylation, pseudouridylation etc.**)**. The specific modification is then indicated as an Xref with database = ChEBI.
Example: *N(6)*-methyladenosine: CHEBI:21891

A useful Database for RNA Mods:-http://mods.rna.albany.edu/mods/
Also http://modomics.genesilico.pl/modifications/

Feature Type: DNA chemical modification, id: MI:2201
Example: 5-methyl-2'-deoxycytidine: CHEBI**:**47876 (as an Xref).

For both, RNA and DNA modifications, the full name of the mod - e.g. "N6-methyladenosine" can be added as the short-label.

7.3 Specific cases

The Feature Editor should be used in the following cases:

7.3.1 Molecule is tagged

The term “Tag” includes small peptide i.e. “epitope-tags” e.g. Flag, V5 and also larger fusion proteins e.g. GFP, GST.

Curators need **not** enter feature type ‘tag’ or children of this term where the Interaction detection method is a standard ‘protein complementation assay’ or a child term AND the tag is explained in the definition of the method used or is used regularly with the method.
For example Interaction Detection Methods such as:-
Two Hybrid, DHFR reconstruction, Luciferase Complementation assay.

If, however, the tag is non-standard this DOES need to be added.

Short Label
The short label is used to describe the position of the tag i.e. ‘’ or ‘n-terminus’-when this information is available.
N.B. The term “terminus” is used to show the location of a TAG, whereas the term “terminal” is used to describe the location of a protein sequence.
If the position of the tag is known specific amino-acid positions should be entered. If the position is not known, the range ‘?-?’ is added. If a novel or non-standard tag is used, i.e. one not present in the current controlled vocabularies, the short label will give this information so- ‘xyz-tag n-terminus’ with Feature type ‘tag’. (see examples in Table above)

If the protein has multiple tags, they should be entered as separated features.
Examples:
EBI-6250852 – both a His and V5 tag
EBI-4370371 – both an s-peptide and His tag

7.3.2 Molecule is radio-labelled

Short Label – This is usually “region”, unless a specific residue/region is known to be labelled, in which case this residue or region should be entered in the short label, e.g. “arg-15” or “n-terminus”.
Feature Type: Choose the exact “radio-label” used from the drop down list; example 131i radiolabel or 35s radiolabel
Range - The actual position of the radio-labelling, if specifically known, should be entered in the range – using amino acid positions. Otherwise, Range will be ‘?-?’ if the position is unknown or the protein was globally labelled.
Example: Protein expressed in bacteria grown in 35S growth medium
EBI-465724.

7.3.3 Molecule is post-translationally modified (PTM)

Different types of modifications should always be added as a separate line.

Short label: The should describe the amino acid position using the amino acid 3 letter code-sequence position (e.g. ser-123). If the position is unknown, the short label 'residue' should be used instead. If the type of amino acid(s) modified is known, but the exact site is not, add amino-acid-? e.g. ser-?, tyr-?
If multiple amino acids are modified, list and separate with a semicolon e.g. ser;ser-99
Ir a resultant cleavage is being described, add the amino acid positions either side of the cleavage, separated by an underscore e.g. ser-99_arg-100

Feature Type – select the appropriate term from PSI-MOD (the Ontology can be found in OLS under "**Protein Modifications'** (MOD - Ontology Lookup Service)
and find the matching "Short label curated by PSI-MOD synonym" in the Editor feature drop-down menu.

Feature role: For PTMs, the additional field 'Feature role' needs to be added.
When the PTM has been shown to be required for the interaction, add the Feature role 'prerequisite PTM'
When the PTM has been shown to be present but it is not apparent if it is required for the interaction of not, add the Feature role 'observed PTM'
If the PTM has been shown to alter the strength or rate of an interaction, add the appropriate Feature role from 'decreasing-ptm/disrupting-ptm/increasing-ptm'
If the PTM is caused as a result of an interaction (an enzymatic reaction) the Feature role 'resulting-PTM should be used. If the PTM is a cleavage the more specific 'resulting-cleavage' should be added instead.

e.g.
Editor short label: ser-129
Feature Type: "opser" (O-phosphorylated L-serine) (from PSI-MOD short label)
Range 129-129
Feature role: observed-ptm

Editor short label: lys-123
Feature Type: "NacLys" (N-acetylated-lysine)
Range 123-123
Feature role: required-ptm

If Range is *unknown* enter ?-?

'Feature detection method' may also be entered, if described in the paper.
Commonly used Feature Detection Methods:
'deletion analysis', 'mutational analysis'

Special case: Ubiquitin. This small protein can be linked to another protein as a PTM –ubiquitination. Alternatively, it can be entered as a protein participating in an interaction. As Ubiquitin can be derived from multiple genes, and the exact source is often unknown, entering as the IntAct internal protein ubiq_human, ubiq_mouse, ubiq_arath etc. circumvents the need to know this. Calmodulin is also derived from 3 genes (human) and Histone H4 from 1 but these have not yet been demerged into separate entries by UniProtKB/Swiss-Prot.

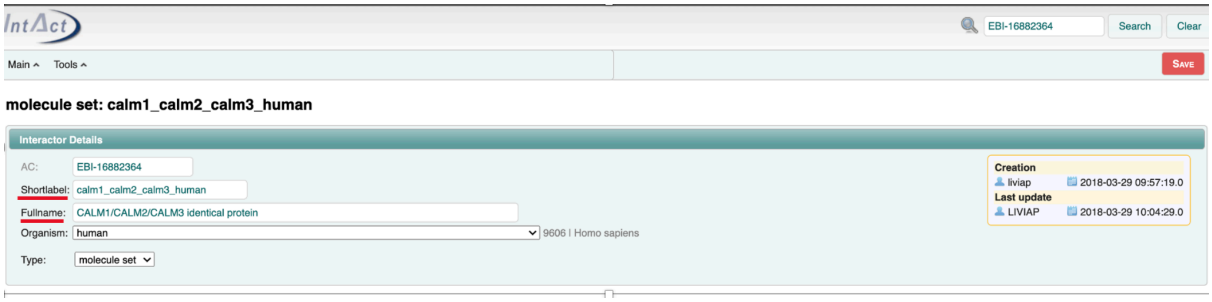
Identical proteins may be added as protein sets.

Curation of sets

Step 1: Conformation of the proteins being identical by aligning the proteins in and checking the similarity

Step 2: In the editor, go to 'Main'>'New'>'Interactor' and choose the 'type' as 'Molecular set'

Step 3: Creating the Short label
The short label has to be created by proving the names of the proteins in the sets and the species separated by underscore. Kindly check the example EBI-16882364 as below.

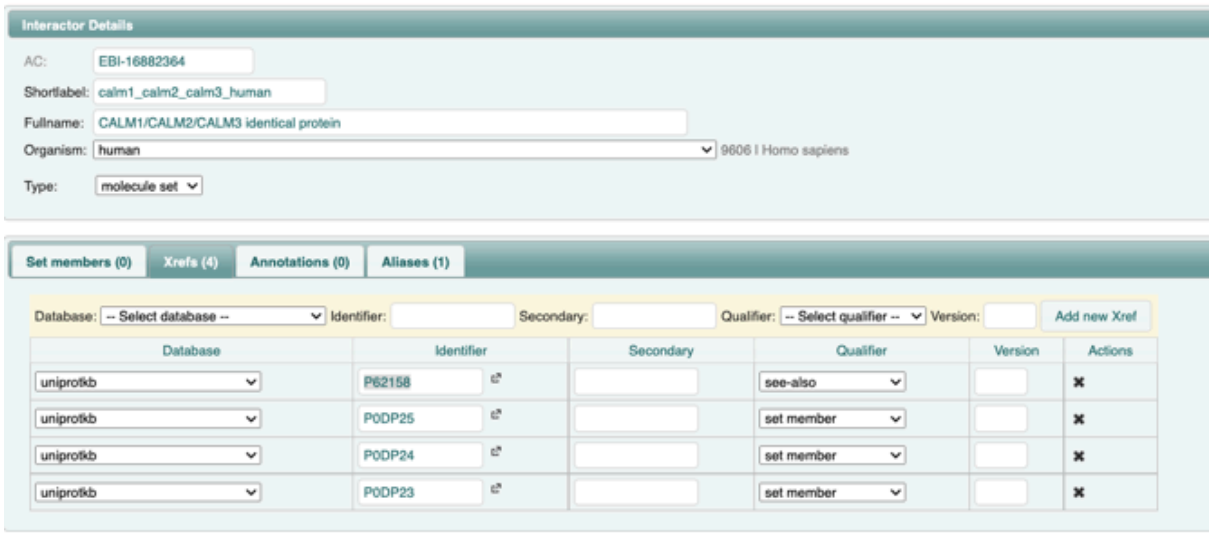


Step 4: Creating the full name
Full name has to have all the gene names of the set proteins separated by '/' symbol followed by the term 'identical protein'

Step 5: Specify the species in the 'Organism' field and the 'type' as 'Set' field

Step 6: Addition of cross-references
The UniProtKB accession number which has been obsoleted, has to have the qualifier as 'see-also' and the other valid accession numbers of the set proteins should have 'set member' as qualifier.

molecule set: calm1_calm2_calm3_human



Sumoylation, neddylation, myristoylation, farnesylation, glycosylphosphatidylinositol (GPI) anchor formation, pupylation and ISGylation result in PTMs formed by the addition of a peptide/small protein or a hydrophobic group.
We do NOT enter the formation of PTMs as interactions, e.g. the ubiquitination of proteins, unless these are described as isolated enzyme assays. Intracellular ubiquitination, for example, is not added.

7.3.4 Molecule is a fragment, truncation or deletion construct

If a paper describes a series of deletion constructs, annotate to the **shortest region** that interacts. It is NOT necessary to list each separate deletion mutant used by an author. Deletion mutations should be described as featureType = “required to bind” or “sufficient to bind” where possible.

For Example: Protein A (length 50aa) binds to Protein B. A deletion mutant of Protein A is constructed from regions 1-20 and 30-50 and this fails to bind. Annotate the fragment 21-29 as the ‘**required to bind**’ with feature detection as ‘deletion analysis’.
Example: EBI-465428

Short Label: This may be described just as ‘region’ or the Short Label may be used to give a more detailed description or features within the fragment Example: ‘sh3 domain’, ‘heat repeat’, ‘nad’ (for NAD-binding site), where the fragment contains the domain, repeat or binding site respectively.
For C-terminal fragments, the Short Label will be ‘c-terminal’
For N-terminal fragments, the Short Label will be ‘n-terminal’
For the cytoplasmic region, the Short Label will be ‘cytoplasmic region’
For transmembrane fragments, Short Label will be ‘transmembrane region’

Example: EBI-77516
Where an InterPro domain, repeat or binding site has been defined, the Short Label can be derived from the names of these domains. Example: ‘sh3 domain’, ‘heat repeat’, ‘nad’ (for NAD-binding site). Use the InterPro short name if the InterPro domain name exceeds 256 characters.

FeatureType = ‘binding site’ (this does not imply all domains within a fragment are necessarily involved in binding). Children of this term – ‘required to bind’ and ‘sufficient to bind’ – should be used where possible.
Range – range defines the binding fragment of the protein.
Cross-reference - if this binding site has a domain, repeat, well defined binding site or active site relevant to the interaction, this should be added in as a cross-reference with the database = ‘interpro’, primary id = ‘InterPro accession number’ and qualifier = ‘identity’. Example: EBI-457906

7.3.5 Molecule contains a Mutation

There are many different varieties of “mutations”, some of which are naturally occurring, however ‘mutation’ in the PSI-MI CV refers to an artefactual change to a sequence, the term ‘variant’ is used for polymorphisms.

GENERAL POINTS regarding Mutations:
XML3.0 contains a defined field to describe the sequence change created by a point mutation and this **must always be filled in**. This is a defined field and can be added using the drop-down menu in the tabbed field ‘Sequence and resulting sequence’

XML2.5 lacks this field, we therefore use the **SHORT LABEL** to give the location of the original un-mutated amino acid(s) and the identity of both the original and resulting amino acid(s) produced by the mutation. Short label should be reported according to the http://varnomen.hgvs.org/recommendations/protein. Examples and further development of the HGVS rules as applied in IntAct curation can be found here: https://docs.google.com/document/d/1OTIv0ygFaU_G4bAB0ej6TfbU7acjhMweNMLopMTA3bw/edit.

N.B. When entering the position of a mutation, as quoted in a publication, it is advisable to check in UniProt that the relevant amino acid(s) really ARE at the position quoted, for the appropriate species. If not, remap to the correct positions.

If the experiment used both, a wild type fragment and a mutated fragment:

- if the mutation impairs/increases an interaction, it should be entered on the positive interaction with a Feature type “mutation decreasing/increasing” etc...
- if the mutation does not change the interaction: please curate as ‘mutation with no effect’.

- if the mutation changes the kinetics (e.g. as in mutations of an enzyme), it should be ‘mutation increasing/decreasing the rate of an interaction’

Mutations that are **not** in the same construct i.e. a series of point mutations have been made, that have **separately** been shown to have an effect, these should be individually entered as a separate feature for each mutation on the same interaction. However, if the effect has only been shown when mutations have **all** been made within the same instance of the molecule i.e. several mutations **in the same construct**, this should be described as a single feature.

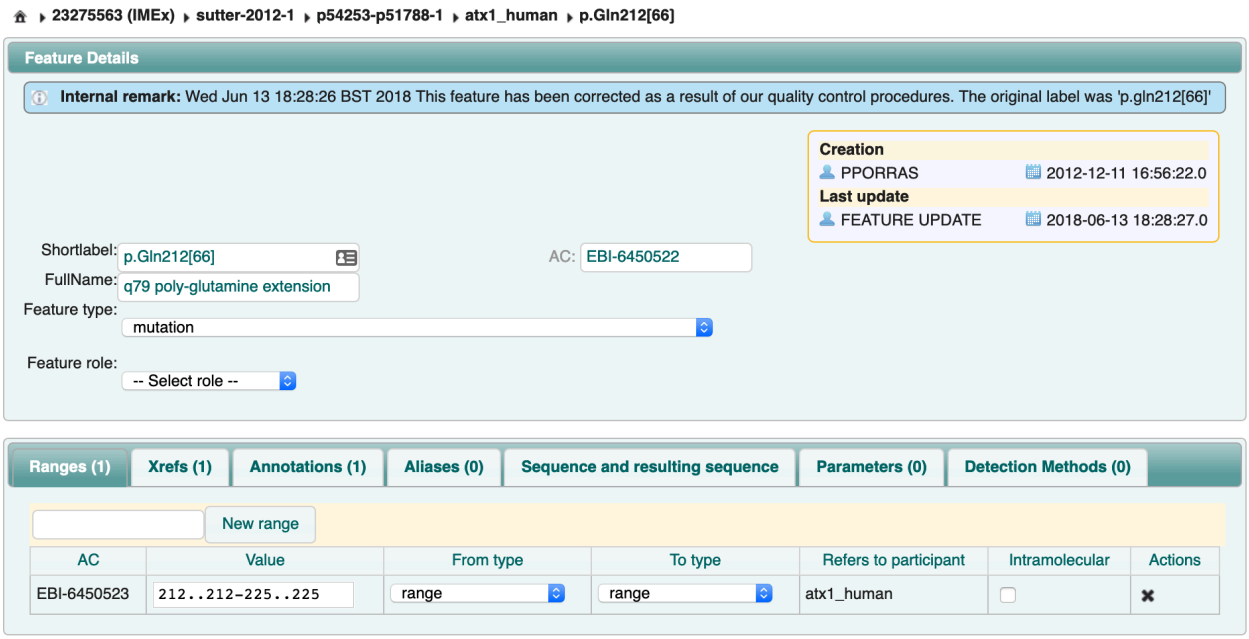
Rules for mutation increasing/decreasing short label notation:

Please see http://varnomen.hgvs.org/recommendations and https://docs.google.com/document/d/1OTIv0ygFaU_G4bAB0ej6TfbU7acjhMweNMLopMTA3bw/edit.

If an interaction contains molecules with multiple mutations and it is not clear what the combined effects of the mutations are, a Comment (preferably taken from the text) can be entered instead of multiple individual mutations.

Unusual feature: Polyglutamine Tract/Repeat:

The exact content of the PolyG tract can be described in ‘Sequence and resulting sequence’ by adding the appropriate number of Gln molecules.



Example of a mutated bait with differing effects on multiple prey

The paragraph explains how to handle cases where a mutation is shown to affect an n-ary interaction in a complex way, meaning that not all participants of the interaction are affected in the same way (e.g. some may disappear, some new may appear).

There are 3 Interactions with **3** different baits – WT, Mutant X and Mutant Y

Bait A	**Wild Type**	**Mutant X**	**Mutant Y**
Prey B	+	+	
Prey C	+		
Prey D	+	+	+
Prey E	+	+	
Prey F			+

This case requires 3 entries to be entered:

1) Bait with Preys A,B,C,D and E (all the preys that interact with the WT protein… F is excluded), where Bait protein has 2 mutations:
    a) mutant X MI:2333 - mutation with complex effect
    b) mutant Y MI:2333 - mutation with complex effect
2) Bait with Prey B, D and E, where Bait protein has 1 mutation:
    a) mutant X MI:0118 - mutation
3) Bait with Prey D and F where Bait protein has 1 mutation:
    a) mutant Y MI:0118 - mutation

WT and mutant interactions should be reciprocally cross-referenced at the interaction level, cross-reference using ‘intact’ as DB and the ‘see-also’ qualifier.

Frameshift and STOP lost mutations

The resulting amino acid sequence for Frame-shift and STOP lost mutations may not be available with the publication. These mutations tend to be of clinical significance, so it will be better if we are able to get the mutated sequence.

We can make use of an online tool, Mutalyzer (available at https://mutalyzer.nl/), for predicting the resulting sequences. Basically this tool verifies if any given HGVS notation is correct or not. Mutalyzer can predict the resulting change for an input, which is a HGVS notation of a mutation at cDNA/genomic level, on RefSeq identifiers. It does not support HGVS for protein sequences or HGVS nucleotide notation on Ensembl.

In many cases, the authors may not give us the original mutation at the nucleotide level, and we might need to do a bit of detective work on the web. As mentioned above, these mutations have

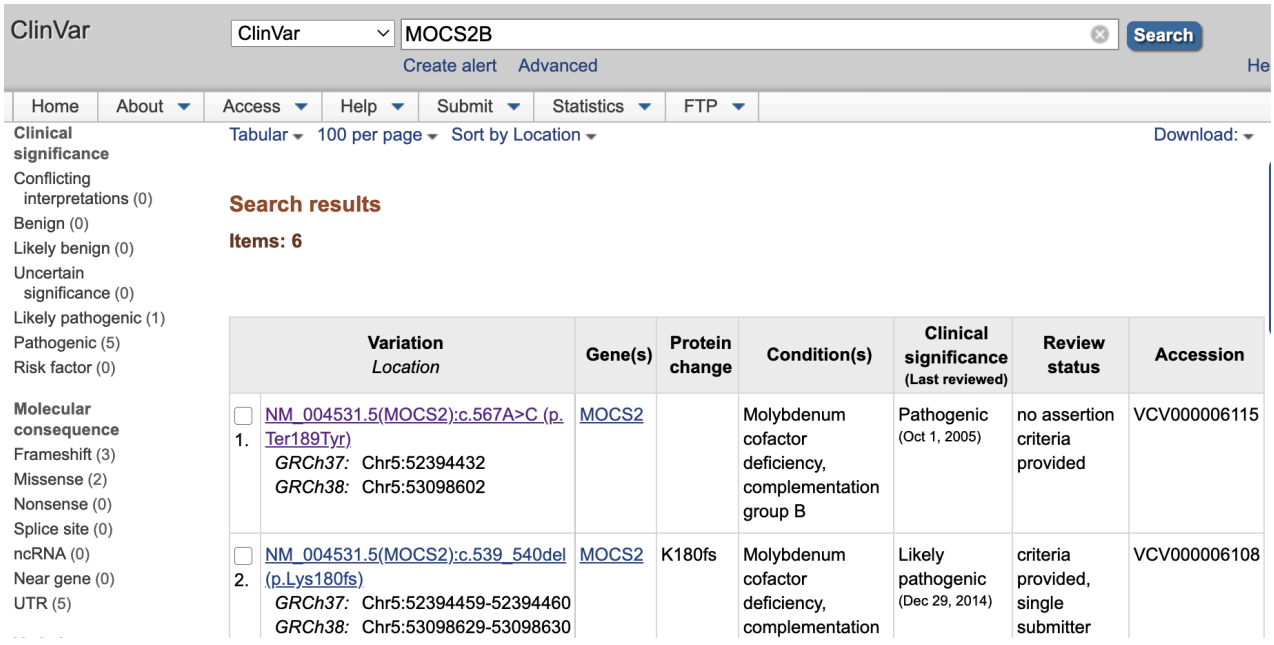
been, so far, clinical which means any basic search in ClinVar with gene name and by using filters we can get the nucleotide HGVS notation easily.

We hosted a session in which details were given about how to use these tools, with examples: https://drive.google.com/drive/folders/1QqncsnF1zgiJkl4sF1wq7AjBcLMGKPBZ?usp=sharing

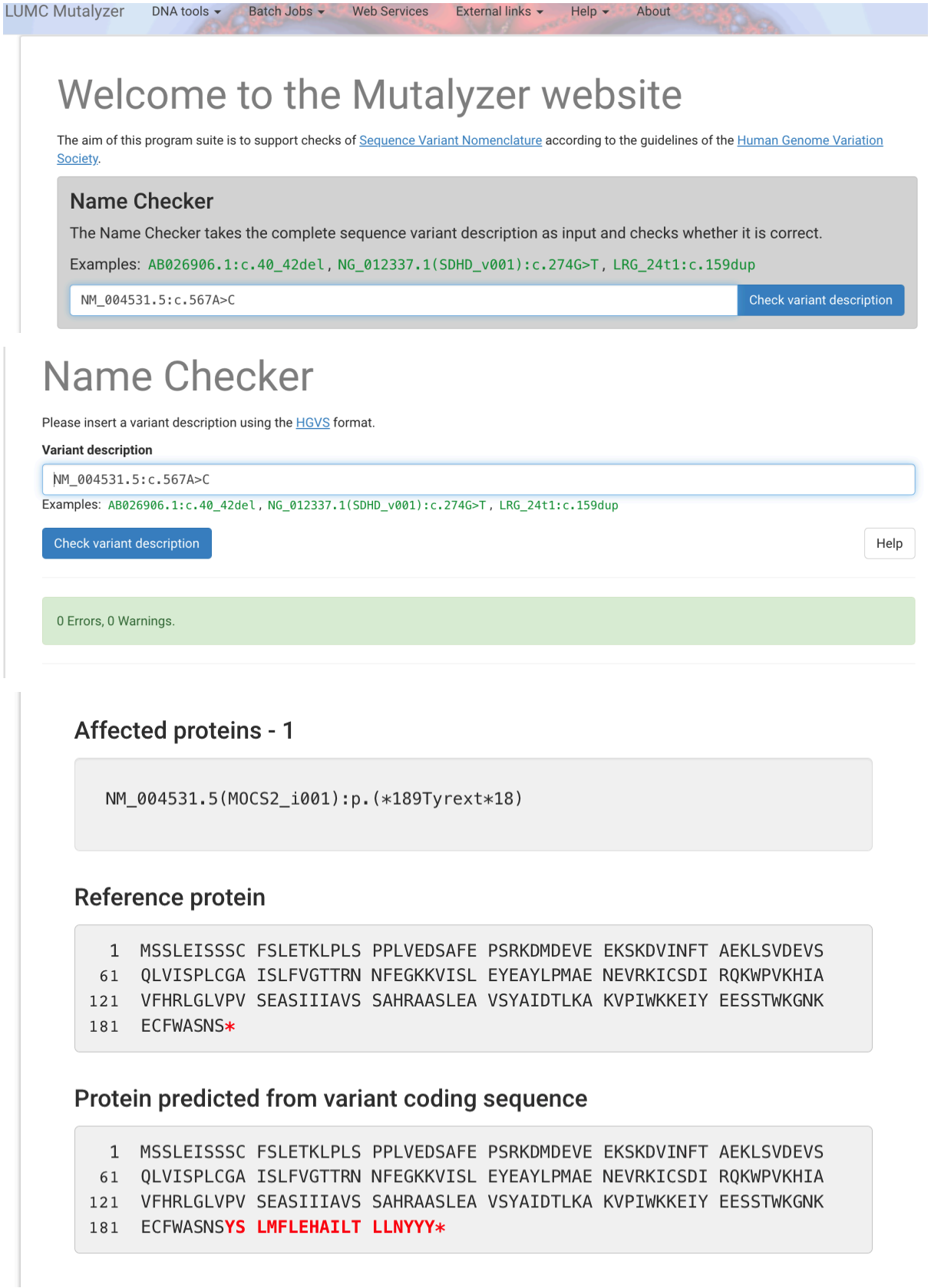
STOP loss mutation

For example, in 16021469, the authors have documented the impact of STOP loss X189Y mutation in MOCS2B, on its interaction with MOCS2A
Captured as interaction EBI-26564681, as feature EBI-26564702 in IntAct

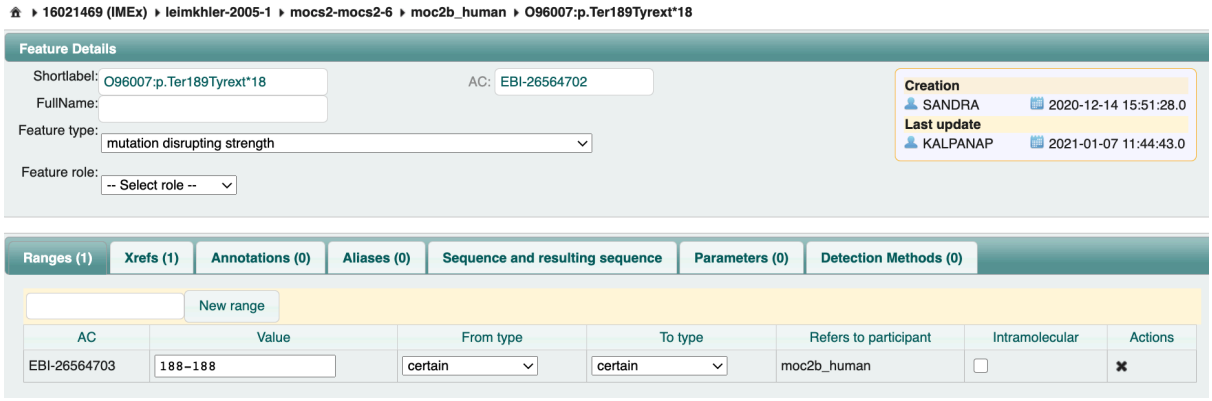
Step 1: Now, to find the nucleotide HGVS version of the mutation, a quick search in ClinVar for the gene name 'MOCS2B' gives us the list of mutations reported for the gene, and luckily for us the first hit turns out to be the one which we are looking for.



We can make use of the filters at the left side to narrow down the results.
Step 2: The nucleotide HGVS, in the above case is NM_004531.5:c.567A>C, can be keyed-in as input and Mutalyzer predicts what would be the resulting mutated sequence as shown in RED, below the original sequence.

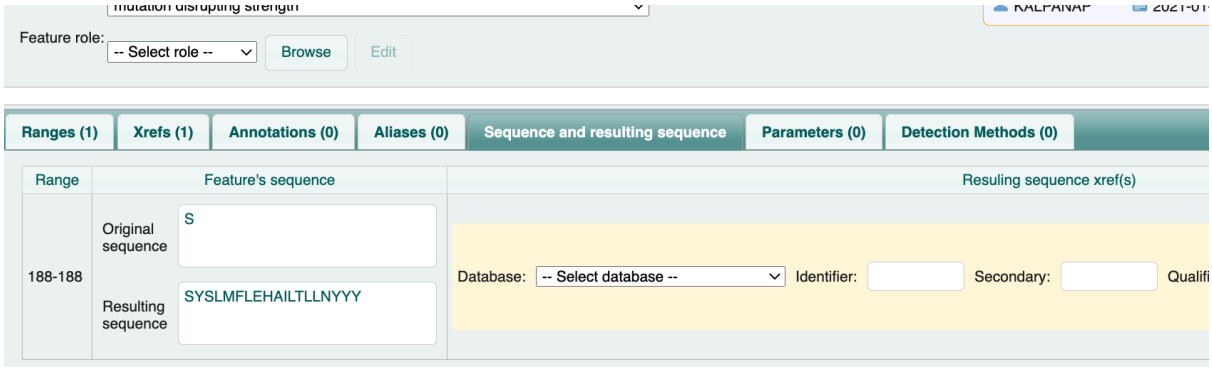

LUMC Mutalyzer
DNA tools
Batch Jobs
Web Services
External links
Help
About
Welcome to the Mutalyzer website
The aim of this program suite is to support checks of Sequence Variant Nomenclature according to the guidelines of the Human Genome Variation Society.
Name Checker
The Name Checker takes the complete sequence variant description as input and checks whether it is correct.
Examples: AB026906.1:c.40_42del, NG_012337.1(SDHD_v001):c.274G>T, LRG_24t1:c.159dup
NM_004531.5:c.567A>C
Check variant description
Name Checker
Please insert a variant description using the HGVS format.
Variant description
NM_004531.5:c.567A>C
Examples: AB026906.1:c.40_42del, NG_012337.1(SDHD_v001):c.274G>T, LRG_24t1:c.159dup
Check variant description
Help
0 Errors, 0 Warnings.
Affected proteins - 1
NM_004531.5(MOCS2_i001):p.(*189Tyrext*18)
Reference protein
1 MSSLEISSSC FSLETKLPLS PPLVEDSAFE PSRKDMDEVE EKSKDVINFT AEKLSVDEVS
61 QLVISPLCGA ISLFVGTTRN NFEGKKVISL EYEAYLPMAE NEVRKICSDI RQKWPVKHIA
121 VFHRLGLVPV SEASIIIAVS SAHRAASLEA VSYAIDTLKA KVPIWKKEIY EESSTWKGNK
181 ECFWASNS*
Protein predicted from variant coding sequence
1 MSSLEISSSC FSLETKLPLS PPLVEDSAFE PSRKDMDEVE EKSKDVINFT AEKLSVDEVS
61 QLVISPLCGA ISLFVGTTRN NFEGKKVISL EYEAYLPMAE NEVRKICSDI RQKWPVKHIA
121 VFHRLGLVPV SEASIIIAVS SAHRAASLEA VSYAIDTLKA KVPIWKKEIY EESSTWKGNK
181 ECFWASNSYS LMFLEHAILT LLNYYY*

Step 3 - The range: The mutation had occurred at 189th CODON which is the stop codon. But as the editor can consider only till the coded sequences, the range of STOP loss mutation in the editor should be the position of the last amino acid. In the above example, MOCS2B is 188 aa length, so the range has to be 188-188



Step 4: The resulting sequence

For filling the resulting sequence, start filling the last coded amino acid followed by the predicted resulting sequence obtained from mutalyzer - In this case 'S' followed by YSLMFLEHAILTLLNYYY = SYSLMFLEHAILTLLNYYY.



Frameshift Mutation

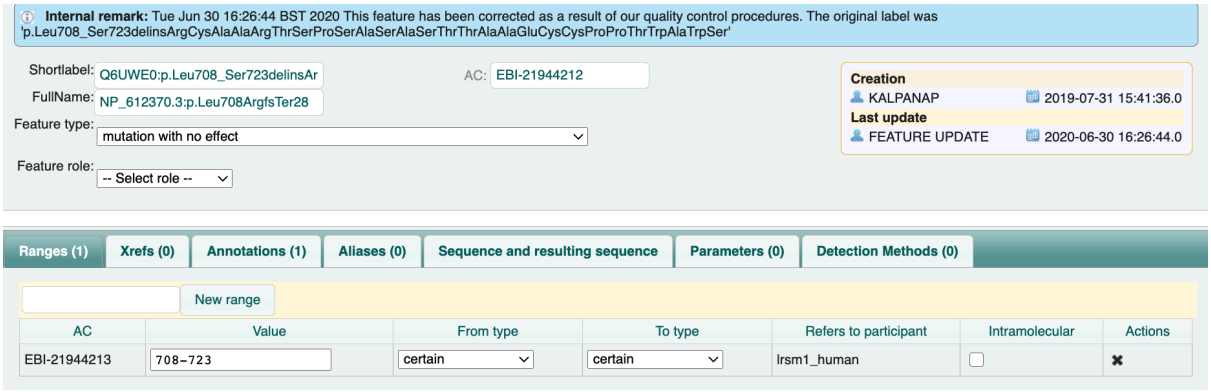
Let us consider **LRSAM1 (c.2121_2122dup, p.Leu708Argfs) = NM_138361.5(LRSAM1):c.2121_2122dup**;

We can find the **resulting sequence** for frameshift mutation by **following steps 1 and 2 using ClinVar/dbSNP and Mutalyzer.**

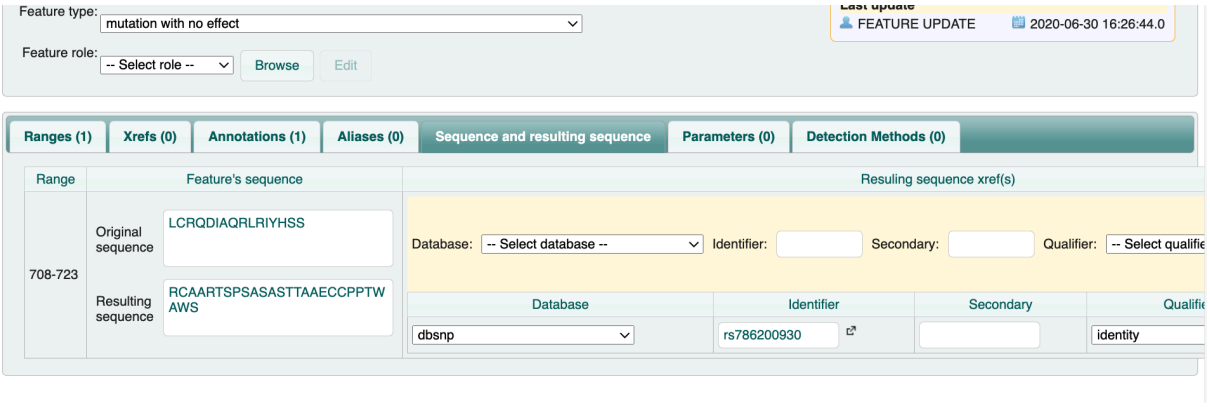
Step 3: The range has the position of amino acid where the mutation had been detected to the last amino acid number. Eg. interaction EBI-21944208 in PMID 28335037
For frameshift mutation in LRSAM1 (c.2121_2122dup, p.Leu708Argfs) the range should be 708-723. (canonical LRSAM1 is 723 aa long).

VERY IMPORTANT: Our current Mutation update pipeline cannot understand the HGVS of a frameshift mutation, so it is recommended to give the HGVS in the 'FullName' data field in the and the 'Shortlabel' should have a deletion-insertion (delins) format.

https://www.bioinformatics.org/sms2/one_to_three.html gives you the conversion of single letter amino acid sequence to triple letter sequence, so that it becomes easy to frame the 'ShortLabel' in delins format.



Step 4: The resulting sequence - fill the predicted sequence obtained from Mutalyzer



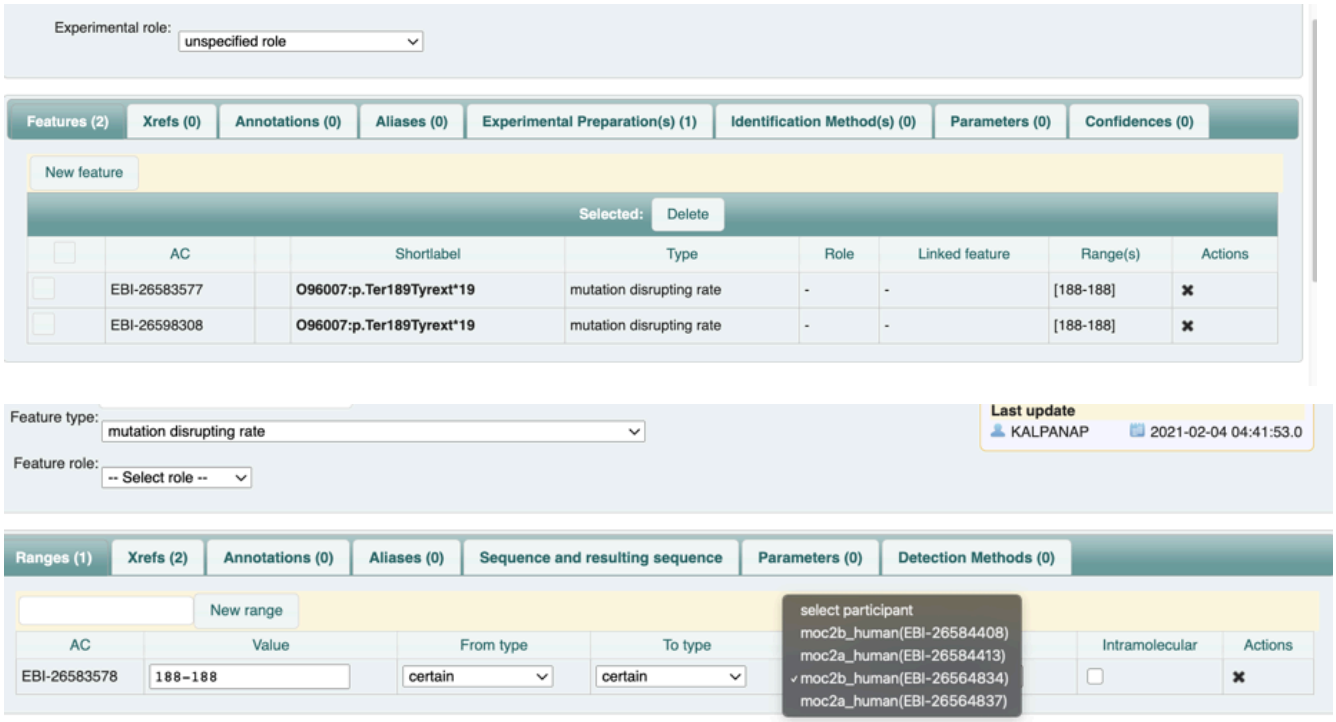
Complexes having mutant participants with stoichiometry more than one

We have started encountering these cases wherein there is a mutation in a component of a complex and the mutant component has stoichiometry more than one.

For **example** interaction **EBI-26564846** - EBI-26564847 is a complex of two participants MOC2A and MOC2B both having stoichiometry 2.

MOC2B carries a STOP loss mutation O96007:p.Ter189Tyrext*19.

In this case the **mutation should be annotated for two molecules of MOC2B**, by annotating the mutation feature twice and tagging to two MOC2B individually.



7.3.6 Linking features

An experiment may show specific residues within two proteins interact, or a residue on Protein A interacts with a domain on Protein B.

For example, domain 50-100 on protein A has been shown to interact with domain 200-250 of protein B

Protein A will have feature Short Label ‘region’, feature type ‘sufficient to bind’ range ’50-100’

Protein B will have feature Short Label ‘region’, feature type ‘sufficient to bind’ range ‘200-250’.

The above features should be linked if they interact, using the ‘Link Feature’ button on the Interaction page. Select the relevant features using the checkboxes next to each Feature, then press ‘Link Feature’. Domains must be shown to interact in the same interaction,

and should be precisely defined – a fragment of unknown coordinates should not be linked.
If one domain "A", from protein X is shown interacting with full length Protein Z in one interaction, and domain B from Protein Z interacts with full length Protein X, they should not be linked if there is no direct evidence that domain A binds to domain B.

Some examples where Features should be linked are:

- one amino acid of one of the proteins is shown to interact with an amino acid of another
- Phosphorylated amino acid is shown to interact with a SH2 domain
- When PTM(s) on proteins are important or necessary for the interaction, these should be linked to the interacting region.

Features within a participant in an interaction may be linked to each other. Example: Interaction: EBI-519830, Features EBI-540445 and EBI-540447 are linked and interacting.

7.3.7 Fusion proteins/unusual Tags

(See earlier section for Identification of naturally occurring fusion proteins e.g. BCR-ABL)

If a tag or fusion protein is not on the drop-down menu, 'Create a Feature' on the protein.
Add Short label: e.g. "xanadu protein" (using the name of the tag or protein fused)
Feature type: tag or fusion protein

As the Term 'tag' is a child of 'fusion protein', the term 'fusion protein' might be applicable to large tags, but only if the tag/fusion protein does not participate in the interaction or has NO biological properties. The tag protein may have been added to assist in isolation, identification of solubility of the protein to which it has been used.

In cases where a fusion protein or chimeric construct has biological properties and participates in the interaction, this data will normally not be curated. However, if there is a reason for doing so, then a new protein has to be created in the editor with the entire sequence of the fused proteins.

7.4 Adding feature to a Complex as a participant

See:
https://docs.google.com/document/d/1yDoHnxKi6JY16H3jq9l7A2VzEu01FNvyVFYSpsWZfeA/edit?usp=sharing

7.5 GAGs

- If GAG is interactor: use new feature "carbohydrate chemical modification" and capture nature of modification in short label, xref to MOD if required.

- GAG as protein modification: New feature type “attached carbohydrate” with MOD as xref to the modification species and chemical modification in short label of feature
- Chemical modification on GAG affecting interaction: New feature type “attached carbohydrate affecting interaction” and a reference to WT that has feature type “attached carbohydrate”. Effects can be reported with appropriate children terms (e.g. “attached carbohydrate increasing/decreasing interaction”), in a similar way mutations are handled.

8. BioSource

The BioSource is used for:

- Host Organism of the experiment
- Expressed in (Biosource used for expressing the participant)
- Interactor Organism: the Organism from which the interactor originates

A Biosource may be any of the following:
a) An organism (tissue/cell line unspecified) Example: *Homo sapiens*
b) A tissue type, which includes cell suspensions derived from tissues.
c) A cell line
d) in vitro
N.B. in the IntAct editor 'Cell type' denotes 'Cell line'.

n8.1. How to create a new BioSource/Host Organism

In practise this only needs to be done if an organism is not currently listed on the drop-down menu for "Host Organism" – most common organisms are already present.

From the publication or Experiment level, click on Main → New → Organism

The main source for organism information is http://www.uniprot.org/taxonomy/

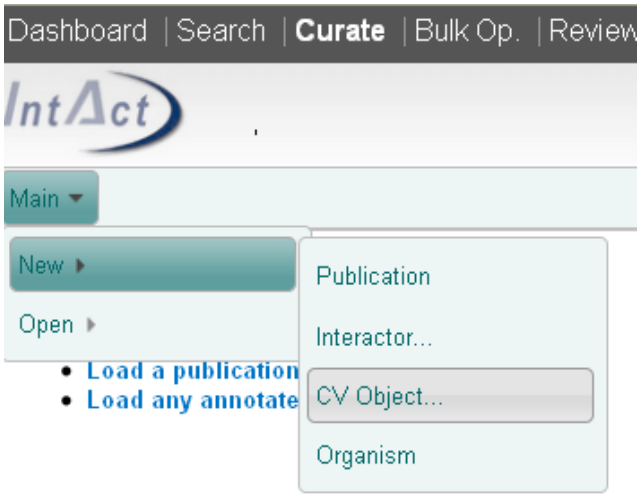
Add the NCBI Tax Id of the new organism e.g. 40324 (for *Stenotrophomonas maltophilia*), and enter it in NCBI Tax Id field in the IntAct editor.
Press 'Auto-complete', and the other fields should be filled in. (short label is "stema" taken from the UniProt "Mnemonic"/5 letter organism identifier, and the UniProt Xref should have qualifier set as 'identity'.)
Then Save.

8.2 How to create a Cell or Tissue Type

There are two stages to this:-

First, check the cell line or tissue is NOT present in the Biosource drop-down menu already - so SEARCH for it in the editor.

If the cell line/tissue is not currently present in the editor
Select Main → New → CV object/
Then Select 'Cell Type' or 'Tissue' – as appropriate
Click Create



Shortlabel: Add the name of cell line/tissue (in lowercase letters and using underscores instead of spaces)
Fullname: Repeat the name of cell line/tissue and add a very brief description

Then add an **Xref or Annotation**:

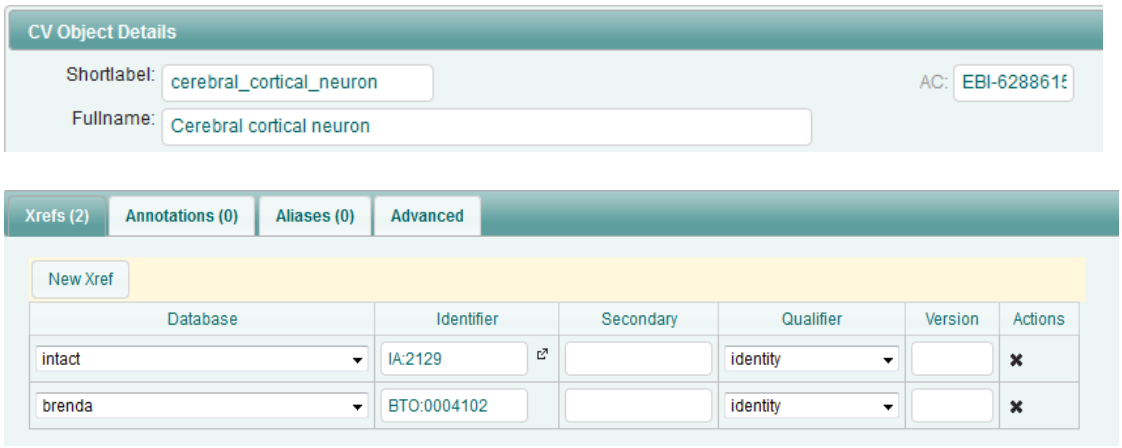
Tissue:
Database = “Brenda”
Identifier = “tissue identifier” (BTO:xxxxxxxx) or EFO
Qualifier = “Identity”

http://www.brenda-enzymes.info/ontology/tissue/tree/update/update_files/BrendaTissueOBO
or
http://www.ebi.ac.uk/ols/ontologies/bto
or
http://www.uniprot.org/docs/tisslist.txt

Cell Type (immortalised cell line):
Cabri or other Database:
Database = “Cabri” or other appropriate DB
Identifier = “Cabri or other DB identifier”
Qualifier = “Identity”
ATCC cell lines:
Do NOT add as Xref, but
Annotation topic = “url”
Description = “url of entry”(if you cannot do either of these, add the catalogue number as a Annotation topic = “comment”.)
AND add an annotation (comment) giving further info about the cell line.

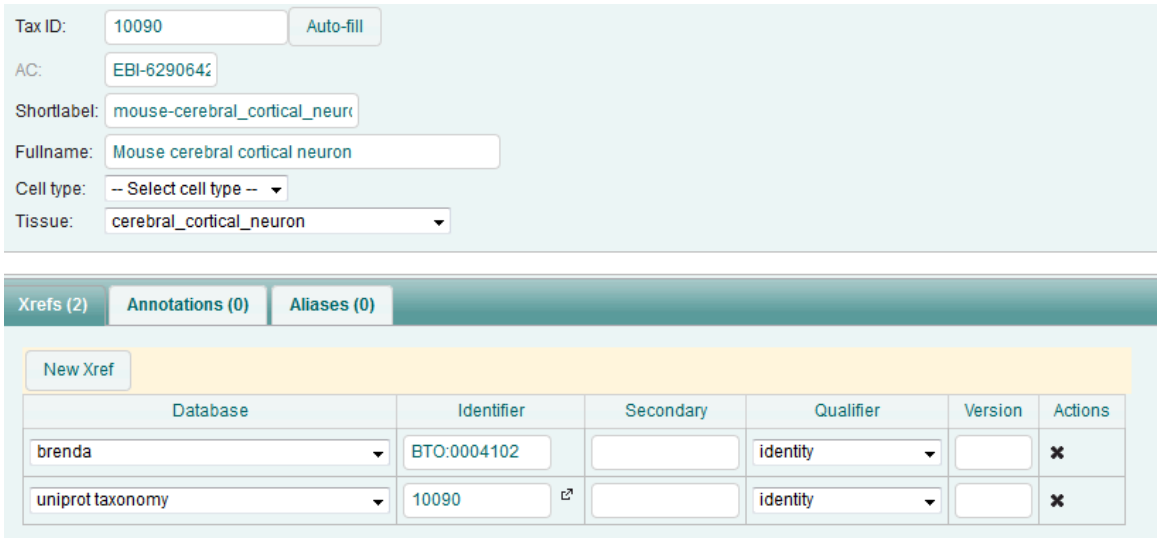
To locate your cell line in Cabri/ATCC it is probably easiest to do a Google search such as “293 cabri” or “293 atcc”.

Some cell lines are now present in Brenda.



Link this new cell type/tissue to a BioSource. If you do not, you will not be able to see it on the Host-Organism drop-down menu.

Select Main → New → Organism
Enter the TaxID of the species this cell line/tissue belongs to
Click on Auto Fill
Short-label: ‘Species-name of cell line/tissue’ e.g. ‘human-293. Use underscores instead of spaces in the cell type/tissue name.
Fullname: Repeat the name of cell line/tissue and add a very brief description, e.g. Human epitheloid cervix carcinoma HeLa cells.
Add either the Cell type (cell lines) OR Tissue (tissue or primary cells) from the drop-down list near the bottom of the page
NB: Start the species name with upper case, unless there's a good reason not to. Do not add a full-stop at the end of the full name.
Save



The new cell line/tissue should now appear on the Host organism drop-down menu and can be used in an experiment.

8.2.1 Tissues vs cell lines

There can be confusion as to what constitutes a 'tissue'. For the purposes of the IntAct database, tissue includes primary cells which have been directly derived from an organism, then passaged to provide a single cell type. A 'tissue' will usually have an entry in either the Brenda tissue ontology (https://www.brenda-enzymes.info/ontology/tissue/tree/update/update_files/BrendaTissueOBO or https://www.ebi.ac.uk/ols/ontologies/bto) or UniProtKB tissue list: http://www.uniprot.org/docs/tisslist.txt
When a specific cell type cannot be found in either list, use the closest term available and add the detailed description in the 'exp-modification' annotation on experiment.
Example: colonic circular smooth muscle – add as (Tissue) 'circular smooth muscle' and add 'colon' as an exp mod.
Examples: blood, aortic smooth muscle, adipocyte, melanoma, neuron

An embryo becomes a foetus when organogenesis commences and the dpc (days post coitum) of the organism should be considered when using these terms.

In the case of a specific bacterial strain, take the strain name from UniProtKB species list: http://www.uniprot.org/docs/speclist.txt or http://www.uniprot.org/taxonomy.

8.3 Cell Type

Immortalised cell lines are referred to as 'cell types' in IntAct.

8.3.1 Re-classified cell lines

If a cell line has been re-classified or proven to be contaminated it must be annotated to the corrected BioSource.
Example: In EBI-367374 (PMID:15109305) the authors state that KB cells were used in the experiment. Subsequently these cells have been re-classified as a HeLa sub-clone (See definition in CABRI database). Therefore, the BioSource for this experiment must be entered as HeLa cells and a note in the annotation under 'CAUTION' to explain the discrepancy.

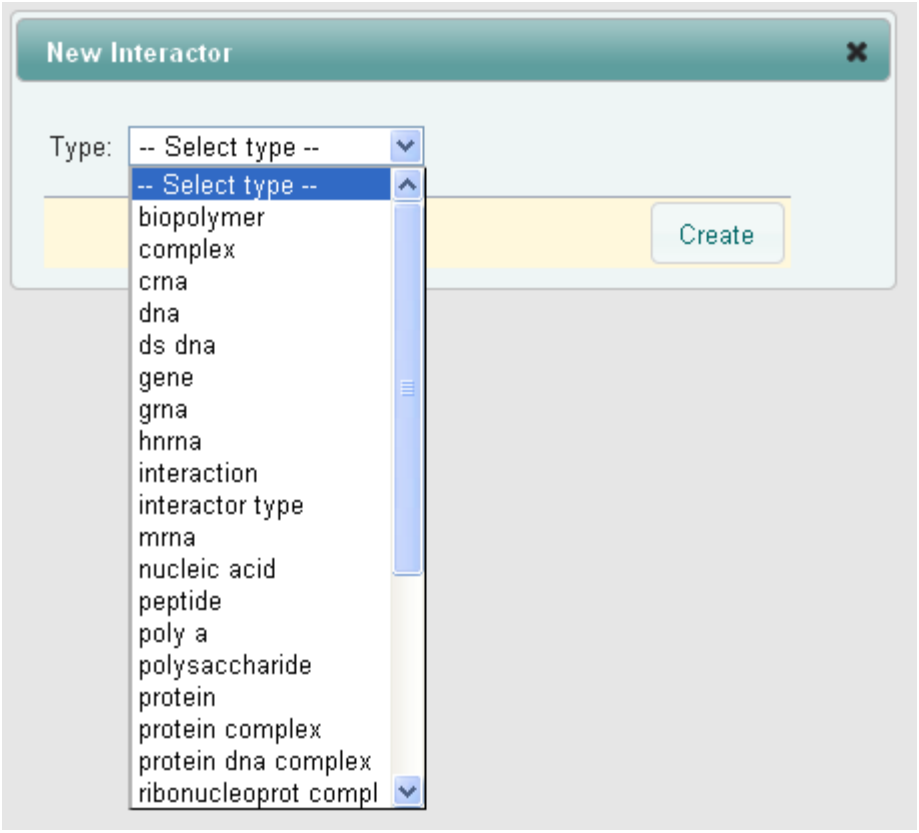
8.4 Special cases

In the case of host organism for Y2H assays, it was agreed to use the top level hierarchical yeast term (yeasx TaxID:4932), but if a yeast protein was identified as an interactor, this should be mapped to the correct strain or reference proteome.

In the case of a crystal, the host organism is always taken as the crystallisation solution i.e.'in vitro' even in the case of a homomultimer.

9. Creating an Interactor

Go to Main → New → Interactor
Select the interactor type from the pulldown menu, then 'Create'.



If your molecule is a peptide, do you need to create this as a new entry, or can it be mapped to part of an existing protein in UniProtKB? **BLAST** your sequence to find out if this is the case. If it corresponds to a partial sequence from an existing protein of the correct species, use the canonical protein as your participant and add a Feature (type = Sufficient to Bind) with the range corresponding to your peptide sequence.

9.1. Protein

A new protein interactor is required when either:
1. The protein is synthetic and does not have a biological equivalent with a UniProtKB accession number assigned to it.
2. The protein is a chimeric protein which cannot be described adequately within the feature editor. Such a chimeric protein may be made up of two (or more) peptides from different gene products or species which interacts with other proteins or peptides (for further info regarding relevance see below).
3. The protein does not have an available sequence at present, such as a protein recognised by an antibody specific for a homologous protein or a protein from a

proprietary database where the sequence is not available. Add a comment “similar to UniProt P9999”.
4. The protein has a sequence available in another database but does not have a UniProtKB entry (this rarely occurs – only use if blasting a protein does not give any UniProt matches >98% identity with full sequence match.) An Xref should be added – qualifier = identity – for the non-UniProt entry.

9.1.1 Short Label

The short label should, wherever possible, use the protein name of a homologous protein (ortholog or paralog), followed by _organism name and _protein or _peptide as the case may be.
Examples:

EBI-1563412	loc653596_human_protein
EBI-6868540	s10aa_chlae_protein

In cases where this is not possible, for example the peptide is synthetic with no biological counterpart having a UniProtKB accession number, the identifier used by the author of the paper should be used as the short label, followed by underscore peptide ‘_peptide’. As a last resort, an identifier consisting of the word pep and the first 17 amino acid single letter code characters should be used. In the latter case the short label should take the form: pep your amino acid sequence (lowercase).

9.1.2 Source

The source of the protein should be selected from the drop down list. It should always be a species **not** a tissue or cell type organism. If the source organism of the peptide or protein is not within the drop down list, a new BioSource entry must be made for it. If a peptide has multiple biological sources (an identical peptide is a conserved protein in several organisms), the peptide should be matched to the species of the other interactor(s) in the assay and the appropriate entry created.

9.1.2.1 Chemical synthesis

The peptide has been synthesised through a series of physical and chemical manipulations usually involving one or more chemical reactions *in vitro.* The “expressed in” for this peptide on the interaction will be chemical synthesis.

9.1.2.2 Chimeric protein/Fusion protein

The peptide or protein has been produced by a splicing together of two or more complete or partial DNA or protein sequences to produce a chimeric or mosaic protein. In cases where a newly created protein (fusion protein/chimera) has biological properties and participates in the interaction, create a new entry using the protein editor with the sequence of the fusion protein. If possible, the UniProtKB accessions of the two proteins should be added in with the cross-reference qualifier ‘multiple-parent’. Example: EBI-1264950 in table below

If the proteins are naturally occurring fusion proteins do **NOT** create a new protein: see section 6.1.1.3 earlier e.g. BCR-ABL –‘Identification of Naturally occurring fusion Proteins’.

Interactions involving fusion proteins that are of NO biological relevance need NOT be curated. For example, PMID: 22939624, a chimeric kinase construct in which one loop region/N-lobe was fused into another molecule or cases in which domains reciprocally switched between molecules.

Table of created fusion proteins

Short Name	Long Name	Species	Annotation	x-ref qualifer
tp53_rara_human_prot EBI-1264950	p53-RAR alpha fusion protein	Human	No-uniprot-update	Multiple parent (UniProt P10276 and P04637)
oct1_oat1 EBI-5261357	Oct1 Oat1 fusion protein	Rat	No-uniprot-update	Multiple parent
Daap-4 EBI-2912160	Human fusion protein: Ig2 and Ig3 domains from VEGFR1, and Tie2 aas 1-348.	Human	No-uniprot-update	

9.1.4 Sequence

Any novel sequences used to create a new protein interactor, which are over 20 amino acids in length, should first be analysed by InterProScan to determine if further information can be obtained about the protein sequence. This information can be entered as an InterPro cross-reference to InterPro via an InterPro accession number.

Any discrepancy between the sequence in the publication/submission and the protein in UniProtKB should be noted as a Caution comment in the annotation section (as a sequencing error or polymorphism if this can be identified).

Note: If a protein is entered via the Protein Editor and has **no accompanying sequence**, numeric **feature ranges** should not be used to describe domains, PTMs etc. even if given in the paper. Non-numeric characters “n”, “c” and “?” may be used, numeric information should be stored in a Comment until a sequence becomes available.
For example a protein may have no sequence either in UniProtKB or written out in the relevant paper, but the authors may state they have sequenced the protein and that the binding site is an SH2 domain. In such a case, the feature range will be “?-?” and the database cross reference will be IPR000980.

9.2 Small Molecule

Small molecules are entered directly into IntAct using ChEBI identifiers in the import tool.

If your molecule does NOT have an existing CHEBI entry, you will have to **create** one at www.ebi.ac.uk/chebi/submissions/login. This will immediately give you a new CHEBI Identifier. You must then create a new small molecule in IntAct using the new identifier as a database Xref.

If your entry is rejected by ChEBI, you can still create a reference entity in the IntACt database (see below).

Whilst ChEBI prefers not to include genome-encoded peptides, as quoted on the ChEBI Homepage: ” The qualifier 'small' implies the exclusion of entities directly encoded by the genome, and thus as a rule nucleic acids, proteins and peptides derived from proteins by cleavage are not included.”, it is possible to create a ChEBI entry for very small, synthetic or modified peptides.
Examples:

Boc-DON-Gln-Ile-Val-OMe	CHEBI:72703
Leu-Leu-Leu	CHEBI:74541
Ala-Leu-Thr-Pro	CHEBI:73292
N-acetyl-L-tyrosylglycylglycine	CHEBI:388118
Ac-D-Phe-His-D-Pro-NH2	CHEBI:73016

To create a new small molecule in the Editor: Go to Main → New → Interactor

Select and enter the correct type of molecule from the drop-down box (see below).
Press Create.

9.3 Nucleic Acid

Nucleic acids are normally only created if a sequence is available or if an exact match for the sequence can be found in ENA or another nucleotide sequence (INSDC - Genbank/DDJB), a genomic database e.g. Ensembl/UCSC/NCBI or RNAcentral (www.rnacentral.org)

(An exception is when you creating a “gene” for Ch-IP experiments: for example “human_bax_gene”; this only requires an Xref to a genomic database, preferably Ensembl).

First, ensure your molecule is not already in the IntAct Database by searching for it in the Editor. If it is not present:

Go to Main → New → Interactor

Select and enter the correct type of molecule from the drop-down-box.

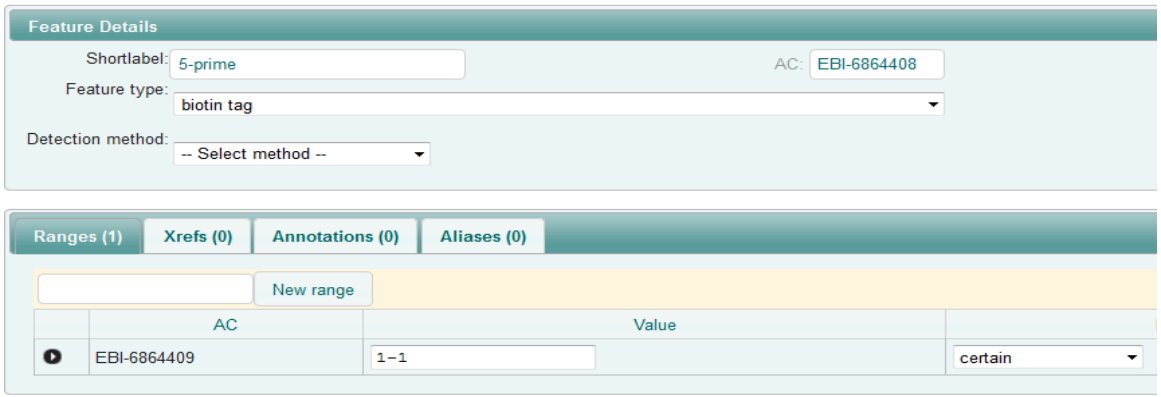
Press Create.
The shortlabel and fullname could be based on the author's description or, for short sequences, the sequence itself can be used for nomenclature.

9.3.1 Examples

Nucleic Acid	Shortlabel	Fullname
DNA	pri-mir-29b2_29c_regulatory _region	Human pri-miR-29b2/29c Regulatory Region
ssdna	m13mp18	M13mp18ssDNA
ssdna	igh_mouse_dna_1.4	Mouse IgH lamina-associating sequence
GENE	Bax_human_gene	Human BAX Gene
tRNA	Trnamet	tRNA(Met)
dsDNA	dsdna_dcdg:dcdg	dsDNA_dCdG:dCdG

Simple double stranded DNA molecules can be entered as Type of Interactor = "ds dna".
If a sequence is given in the publication this should be entered, as should an XRef, if possible.
If neither a sequence nor XRef are available, no entry should be made.

In most instances the sequence of the forward (+) strand only should be entered (5→3). The sequence of the reverse (-) (3 →5) strand is ASSUMED to be complementary and need not be entered.
If there is a tag e.g. Biotin, on the positive/forward strand this can be entered as in the example below (EBI-6864408):

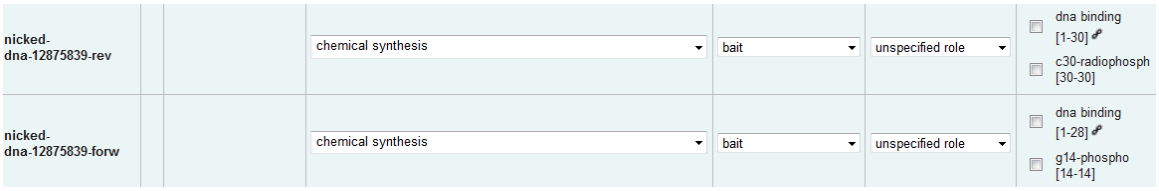


Shortlabel = 5-prime (or 3-prime);
Feature type = type of tag;
Range= numerical only -_no "n" or "c" terminus entries as these terms are protein specific.

If the Tag is found on the negative/reverse strand:
Shortlabel = 5-prime (or 3-prime)_reverse_strand;
Feature type = type of tag;
Range= numerical only (see above)

More complicated double stranded molecules will have to be created as TWO individual single DNA strands, then linked within the interaction, using the linked feature option (create a feature for each which corresponds to the full-length of the molecule) see example:

http://www.ebi.ac.uk/intact/editor/interaction/EBI-6678631



9.3.2 Species

If the species of origin for a Nucleic Acid is NOT stated in your publication, but you have some idea regarding its probable source, BLATTING any primer, siRNA or other sequence given may confirm this.
BLAT search as found at: http://genome.ucsc.edu/cgi-bin/hgBlat
You will need to enter a species and a sequence of >20 bases. The algorithm will only tell you if the sequence entered matches the species chosen. The Blat algorithm is much, much quicker than Blast.

9.3.3 Sequence

This should be entered in uppercase without gaps.

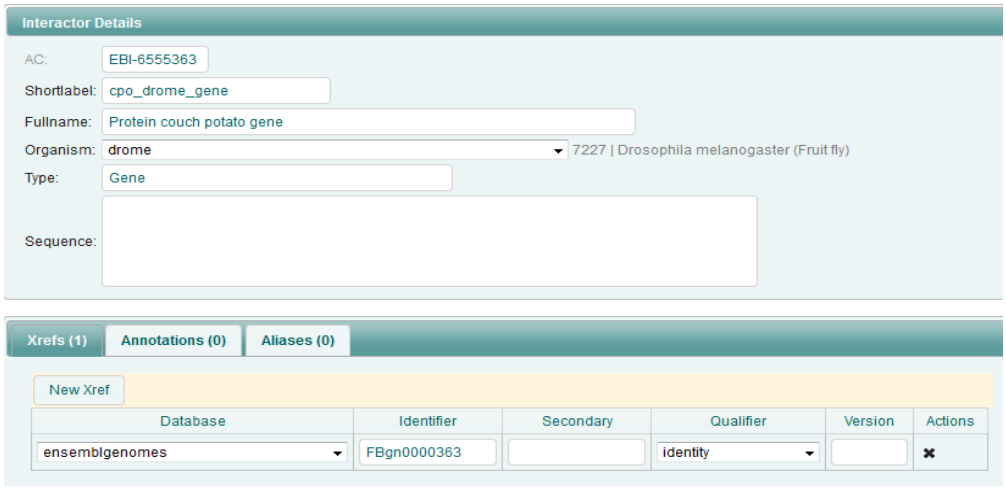
9.3.4 Cross-references

The databases often cross-referenced are ‘embl/genbank/ddbj’, ‘ensembl’ ‘rnacentral’ and various organism databases with qualifier= ‘identity’.

9.4 Gene

Example of a Gene Entry

Gene



9.4.1 To Create a new Gene entry

Provided your gene is in an Ensembl species, entering the Ensembl gene ID into the import tool will create the entry. Note, this does not work for EnsemblGenomes.

To create a new gene, first ensure your “gene” is not already in the IntAct Database by searching for it in the Editor using the Ensembl identifier as a search term . If it is not present:

Go to Main → New → Interactor

Select and enter the correct type of molecule from the Drop-down-box (see below)
Press Create.
You will now have a new page to enter details for your molecule.
Short Label - should be generated in the following format for a Gene:
UniProtKBID_organism_gene

e.g. bax_human_gene or xyz1234_mouse_gene

“**p53**_human_gene – rather than **tp53**_human gene, **tmps2_human_gene** rather than **tmprss2_human_gene**

If the protein is not in Swiss-Prot, the format genename_organism_gene should be used instead.

9.4.2 Species

This **MUST** be added and should be an organism, not a tissue or cell type.

9.4.3 Sequence

This should be entered in uppercase without gaps where you do not have a cross-reference for the sequence.

9.4.4. Cross-references

The databases cross-referenced are ‘ensembl’ or ‘ensemblgenomes’ with qualifier= ‘identity’.

9.5 Annotations specific to new participants

9.5.1 copyright

This topic provides information relevant to copyright statements attached to peptide or proteins (e.g. patents). This topic is added on the interaction.

9.5.2 no-uniprot-update

This annotation topic needs to be added on protein entries that for any reason should not be touched by the protein update mechanism.

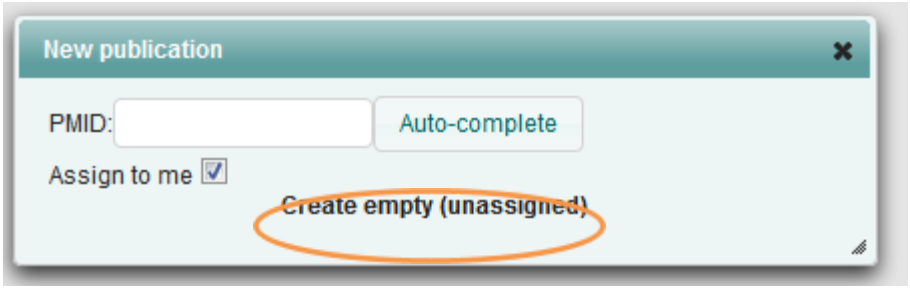
10. Special Cases

10.1 Prepublication or submitted data

Data submitted by an author prior to publication has the highest priority in the curation pipeline. Confidentiality is critical in these cases. The data submitted by the author should be stored at $PRO6/intact/local/data/curation-materials/year directory in a separate directory with the format firstauthor-year, for example smith-2010.

10.1.1 Creating a publication, lacking a PMID in the editor

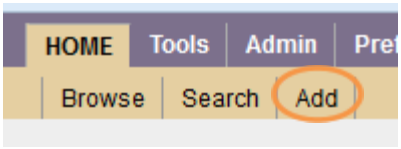
(Note, this is normally only required for handling direct submissions or pre-publication papers)
Press button labelled ‘Curate’
Select ‘Create a new publication’
Select ‘Create entry (unassigned)



An empty entry will be created with an ‘unassignedxxx’ identifier in place of the PMID.
An IMEx accession number also needs to be associated with the entry.

Log on at https://imexcentral.org/icentralbeta/user

Select ‘Add’



Add details of the authors and the eventual publication from the submission. ‘Add’ the paper to the database.

On the next screen, assign an IMEx ID.

Transfer this IMEx ID back onto the entry in the IntAct editor. The database will be 'imex', the primary ID the IMEx accession number and the qualifier will be 'imex-primary'. Then create the first experiment in the normal manner.
Two additional annotations should be added at the publication level.
1. 'submitted' with the annotation in the format:
'2004-09-10: JS Choudhary, The Wellcome Trust Sanger Institute, Hinxton, UK'
2. 'on-hold' added to the experiment. The experiment may be checked and accepted after curation, however the on-hold remains until the authors give us a written permission to make the data public or the publication containing the data is released by the journal. At this point the on-hold should be removed, the correct PubMed ID should be assigned (if available) and the accuracy of the data should be checked with respect to the public paper. The journal, year, author list etc. may be filled up using autocomplete while entering the PubMed ID.

Once the entry is complete, check with the authors that the data is the way they would like it to appear. This is normally done by the curator in charge of the entry. When the author contacts the curator to make the data public or the paper is in the public domain the entry may be released using the unassigned number, a DOI or PubMed ID. If only the DOI is available this becomes the primary-reference. When the PubMedID is available the PubMedID now becomes the primary reference and the DOI qualifier is changed to 'see-also'

A request should also be made to the submitter for inclusion of the following sentence in the publication:

'The protein interactions from this publication have been submitted to the IMEx (http://www.imexconsortium.org) consortium through IntAct [x] and assigned the identifier IM-xxxxx.'

X. PMID: 24234451

If the data is submitted by the author post online publication but prior to publication of the print version, confidentiality does not have to be maintained and a request for inclusion of the above sentence referring to the references to IntAct and IMEx accession numbers in the final publication should still be made.

10.2. Importing interaction data from other databases

If an interaction is imported from another database (Example: KIAA and Riken) put the accession number or reference number for that interaction from the source database in the interaction annotation cross-references. Do not just add the URL of the source database.

10.3 Annotating data from partially curated source

This data was obtained from sources other than the original publication. There are currently only two examples of this – an early FlyNet and a PDBe import.

FlyNet consisted of a set of binary interactions, linked to a single experiment from a single holding publication. The papers these interactions came from were added as xrefs on the interaction. If you curate a paper Aug 2005 or earlier which contains drosophila proteins, check the PMID. If it brings up an interaction attached to the jacq-2005-1, (EBI-866743, unassigned4), it is part of the FlyNet set. If you curate the paper, delete the PMID xref from the FlyNet interaction and, if it is the only xref on that specific interaction, delete the interaction as well.

Any entry marked with the dataset PDBe has structural wwPDB data imported from PDBe. This contains all the information pertaining to the structural studies from the publication; however other experimental data contained within the publication is not covered and needs to be manually annotated. When annotating these papers do the following:

1. Remove the dataset annotation on the Publication page.
2. Derive stoichiometry from the publication and add this to the interactors.
3. Check that the features and sequences are appropriately represented in the interaction.
4. Add any additional experiments.
5. Add an IMEx ID (assuming the entry is now of IMEx standard).

10.4 Unpublished data submitted by the author as an extension of published data

Occasionally authors ask us to hold additional information which was not included in a particular publication but was generated by the same methodology. The original publication where the methodology and some data have been published must be curated in IntAct before curating the additional data.
Where such additional data is being curated the annotation topics ‘submitted’, ‘url’, ‘data-processing’ must be filled in where appropriate. An annotation ‘caution’ must be added with the description that reads as follows:
‘The data described by this entry is not part of an independent publication and contains data produced as an extension of the work described in PMID: xxxxxxx using the methodology described therein.’.

This type of data will not have a PMID, it will retain the ‘unassigned1234” identifier in its place. The original article should be cross-referenced with its PMID and a reference qualifier see-also.

10.5 Handling Amended Large Datasets

10.5.1 When the error is by the database

a. **The data was correct but a subset is being withdrawn** (e.g. a low confidence subset). The interaction subset should be deleted. A Caution comment should be added to the experiment, and the original dataset may be made available at a static ftp site should the originating database wish. A news item/tweet should be made announcing this change. No change to the IMEx ID required.
b. **The data was incorrect** – the interaction dataset (and experiment if necessary) should be deleted and replaced by a new dataset (and experiment). A Caution comment should be added to the experiment. A news item/tweet should be made announcing this change. The IMEx ID of the paper should not change but interaction IDs should be subsequent to the original set, rather than a repeated set. To avoid this problem, please import the new interactions before deleting the old ones, so the IMEx assigner will give them new, non-redundant IMEx IDs.

10.5.2 When the error is by the author

a. **The erratum is published** – the publication should be entered as a separate entry with a new IMEx ID. If the original entry still contains data additional to that which has been withdrawn, it should be maintained with a Caution comment added to the publication, and the PMID/DOI and the IMEx ID of the new publication of the erratum added, using qualifier see-also. The erratum publication should have its own PMID/DOI as primary-reference, its unique IMEx ID as IMEx-primary, and the PMID/DOI and IMEx ID of the original paper as qualifier see-also. A news item/tweet should be made announcing this change.
b. **The author does not wish to publish the erratum.** The interaction dataset (and experiment if necessary) should be deleted and replaced by a new dataset (and experiment). A Caution comment should be added to the experiment. A news item/tweet should be made announcing this change. The IMEx ID of the paper should not change but interaction IDs should be subsequent to the original set, rather than a repeated set. To avoid this problem, please import the new interactions before deleting the old ones, so the IMEx assigner will give them new, non-redundant IMEx IDs.

10.6 Data represented more than once in the database

Data may be represented more than once in the database because it is shared between independent studies. If at all possible, try to avoid this situation unless it is a direct request from an author submitting data and you cannot persuade him/her that this is not a good idea. To enter such data, you need to identify the publication that is the primary source of the data. This could be done by contacting the author/s. The data on this

primary publication alone must be considered for DR and CC line export. This data in all the other places should be attached to an experiment with the annotation 'uniprot-dr-export: No'. The primary and subsidiary publications with the data should have a cross-reference to each other using database = "pubmed", identifier = "PubMed id" and qualifier = "see-also"; and database = "intact", identifier = "IntAct Accessions" and qualifier = "see-also". Annotation topic = "comment" on the experiments must clearly describe the primary and subsidiary dataset.
Example: EBI-1256426, EBI-1190357

10.7 Curation request

Request for curating published work whether by the authors of the publication or scientist in the field should have the annotation topic 'curation request' with the annotation in the format:
'2004-09-10: JS Choudhary, The Wellcome Trust Sanger Institute, Hinxton, UK'

10.8 Causal Interaction Annotation – draft annotation rules

Causality consists on an additional information than can be added to a physical interaction describing the effect that one molecule has on its binding partner (for instance, the phosphorylation of MEK by ERK results in the activation of MEK).

Annotation of Causality is NOT mandatory in IMEx.

Causal Interactions can be captured (in IMEx interactions) only when they meet the following criteria:

1. The interaction is binary [or can be exported as binary].
2. Interactions to which this annotation is added should be 'interaction type = direct' (or child terms), with the exception of 'Ch-IP' and 'mirna interference luciferase assay'.
3. Causality must be experimentally demonstrated in the same paper. Inferred causality, or references to external publications, should not be captured.
4. To define directionality, Participants need to have prerequisites to be annotated with biological roles:
    a. Modulator entity: regulator (and children) or enzyme
    b. Modulated entity: regulator target or enzyme target
5. Causality is an annotation captured at the Interaction level, using the annotation topic: 'Causality Statement'
6. Causality statements will be made using the following controlled vocabulary, with NO additional text, copy/paste from the OLS is STRONGLY recommended (for further detailed terms, please see also all the children of 'causal statement':

https://www.ebi.ac.uk/ols/ontologies/mi/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMI_2234)

up-regulates

up-regulates activity

up-regulates quantity
up-regulates quantity by expression
up-regulates quantity by stabilization

down-regulates

down-regulates activity

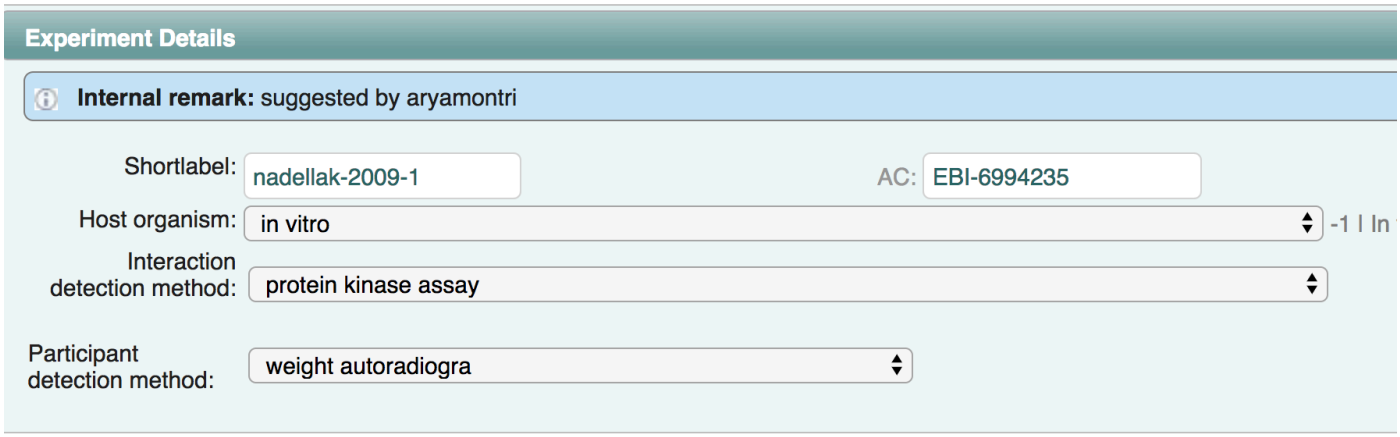
down-regulates quantity
down-regulates quantity by expression
down-regulates quantity by stabilization

7. When an interaction results in a PTM being added to the substrate, capture this using the following syntax

Annotation topic 'resulting-ptm' free text Identifier amino acid-three letter code-residue number [MOD CV term]
e.g. P12345 ser-99 O-phospho-L-serine

Example1: enzymatic assays
(interaction EBI-6994245, phosphorylation)
PMID: 19424295, protein kinase A phosphorylates and activates Limk1

19424295 (IMEx) nadellak-2009-1



Interaction: **limk1-prkaca-1 (EBI-6994245)**

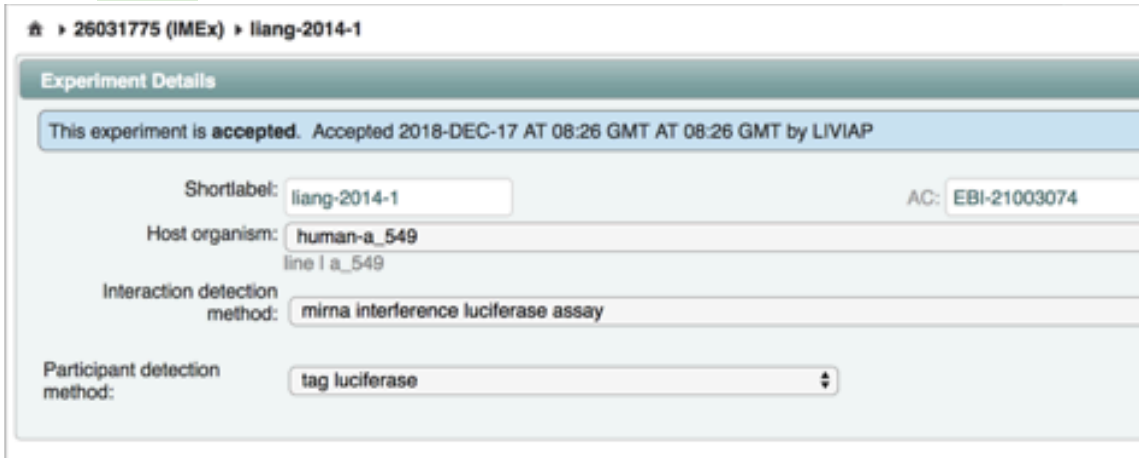
Participants: 2

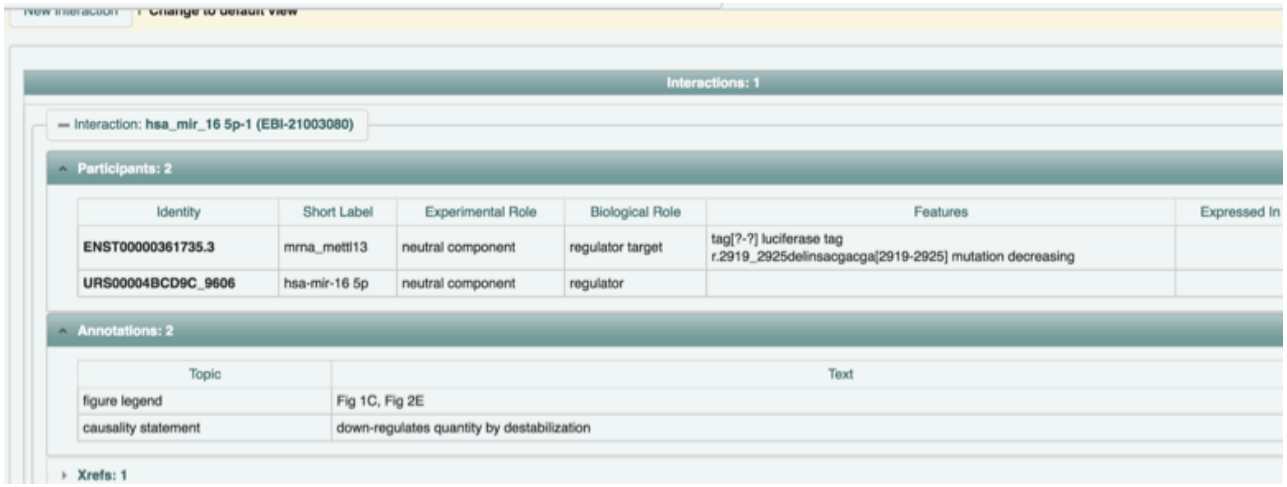
Identity	Short Label	Experimental Role	Biological Role	Features
P53667	limk1_human	neutral component	enzyme target	gst tagged[?-?] gst tag p.Ser596Ala[596-596] mutation decreasing rate p.Ser323Ala[323-323] mutation decreasing rate
P17612	kapca_human	neutral component	enzyme	

Annotations: 4

Topic	Text
causality statement	up-regulates activity
figure legend	F3

Example2: mirna interference luciferase assay
(interaction EBI-21003080, physical association)
PMID: 26031775, miR-16 targets FEAT mRNA and induces its degradation





Interactions: 1

Interaction: **hsa_mir_16 5p-1 (EBI-21003080)**

Participants: 2

Identity	Short Label	Experimental Role	Biological Role	Features	Expressed In
ENST00000361735.3	mrna_mettl13	neutral component	regulator target	tag[?-?] luciferase tag r.2919_2925delinsacgacga[2919-2925] mutation decreasing	
URS00004BCD9C_9606	hsa-mir-16 5p	neutral component	regulator		

Annotations: 2

Topic	Text
figure legend	Fig 1C, Fig 2E
causality statement	down-regulates quantity by destabilization

Xrefs: 1