Curation Manual for Macromolecular Complexes

1. What can be described as a complex

A stable set of (two or more) interacting macromolecules such as proteins which can be co-purified by an acceptable method and have been shown to exist as an isolated functional unit *in vivo*. Any interacting non-protein molecules (e.g. small molecules, nucleic acids) should also be included.

1.1 What should not be captured

1.1.1 Enzyme/substrate, receptor/ligand or any similar transient interactions (see exceptions at the end of the manual).
1.1.2 Proteins associated in a pulldown / coimmunoprecipitation with no functional link or any evidence that this is a defined biological entity rather than a loose affinity complex.
1.1.3 Any literature complex where the only evidence is based on genetic interaction data.

Note:
You may curate the paper(s) detailing the experimental evidence for the existence of the complex into IntAct so that a link to this can be provided from the curated complex. Then curate the complex as described below:

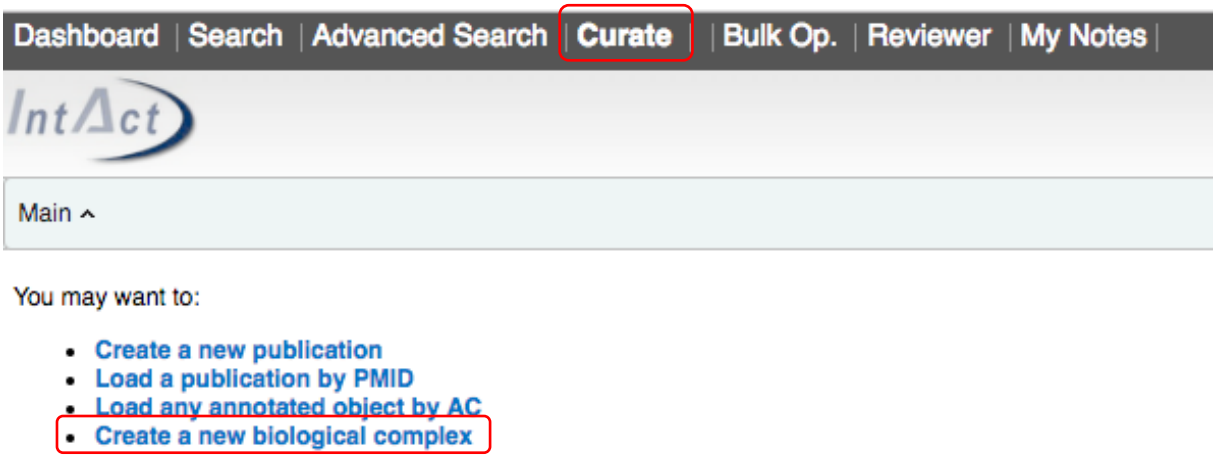
2. Curating Macromolecular Complexes

Notes:

- The complex appears on the Dashboard under its "complex recommended name".
- When using the new CPX-xxxx complex IDs as cross-references do not add the version number. Versioning is getting handled automatically.

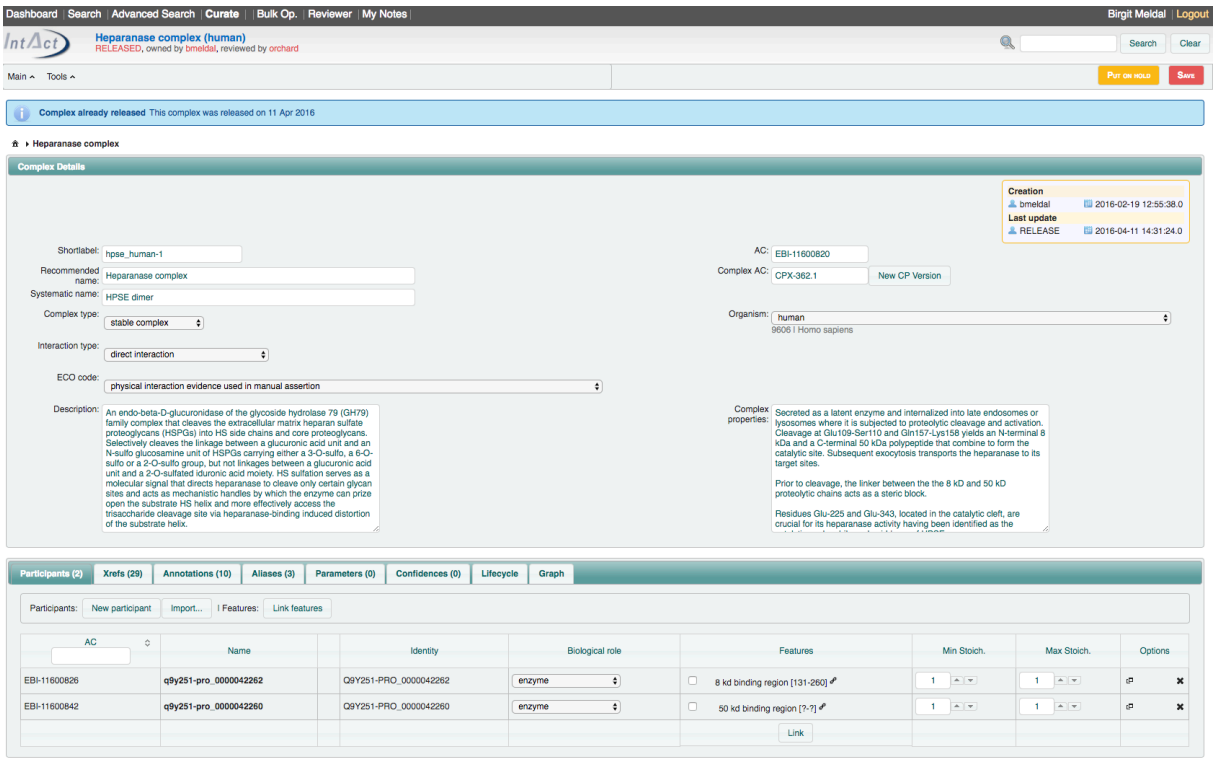
2.1 Creating the basic complex

2.1.1 Go to http://www.ebi.ac.uk/intact/editor/login.xhtml to login.
2.1.2 On the Dashboard, select <Curate> <Create a new biological complex>.

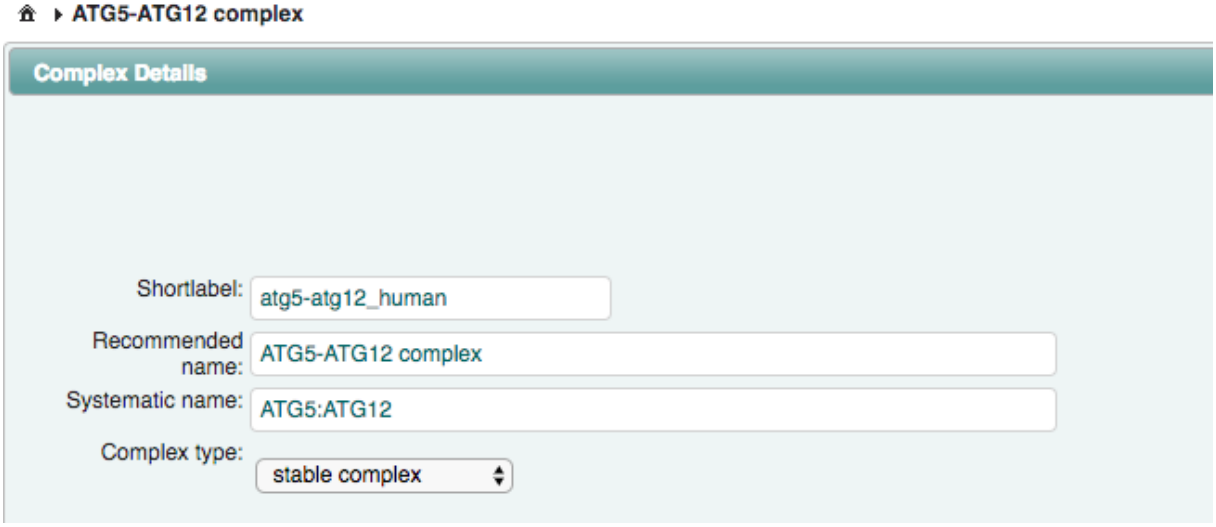


The curation page has got two sections, a static top section (or 'Header') and a bottom section consisting of several tabs. All fields in the Header are mandatory except for <Complex properties>. Some of these fields are repeated in the tabs but greyed out. You can only annotate them in the Header. This functionality ensures that no mandatory fields are forgotten and no fields are annotated in more than one place.

e.g. Heparanase homodimer (EBI-11600820/CPX-362):



e.g. ATG5-ATG12 complex – no specific recommended name (EBI-11598325/CPX-356):



2.2 Annotations in the Header

2.2.1 <**Shortlabel**>: an appropriate designation for the complex with species indicated using the UniProt five letter code e.g. fibrinogen_human, tfiid_mouse. When the same complex is present in multiple organisms, the root name should be maintained across all entries e.g. fibrinogen_human, fibrinogen_mouse, fibrinogen_bovin. The <Shortlabel> should be related to the <complex recommended name> (see below) but might have to be abbreviated. If no <complex recommended name> exists, use the <complex systematic name> (or abbreviation thereof). Note: This field does not appear on the Complex Portal website but is stored in the database and xml file and forms the GO annotation object's symbol and is searchable in the Editor.

2.2.2 <**complex systematic name**>: a string of gene symbols of the participants of the complex separated by a colon (e.g. "fiba:fib:fibg"). Ignore non-protein participants. The order of the gene symbols is alphanumerical, and numbers precede letters, e.g. AB1D:ABCD. For large complexes the systematic name can be abbreviated with <complex recommended name> terms (but so far we always had enough characters to spell it out). Note: A complex can only have one systematic name. If multiple gene names exist, chose one version for the systematic name and add all variants as <complex synonym> (see section 2.5.3), e.g. haemoglobin HbA complex (CPX-2158).

2.2.2.1. Mixed-species complexes (e.g. host-pathogens): retain the alphanumerical order irrespective of the taxon of each gene symbol.

2.2.2.2. Isoforms and UniProt PRO chains: use gene symbols; the complexes are distinguished by their participant IDs and the recommended names (see below).

2.2.2.3. Stoichiometry: If a component occurs more than once add the number and "x" in front of the gene symbols (e.g. 2xGABRA:2xGABRB:GABRG has a 2:2:1 stoichiometry).

2.2.2.4. Oligomers/polymers: use the gene name followed by 'oligomers/polymer'.

2.2.2.5. HOMOmers: use the assembly term, e.g. dimer, instead of the numerical stoichiometry (e.g. 2xJAK2 becomes JAK2 dimer).

2.2.2.6. Complexes with complexes as participants: list all gene symbols for all proteins at all levels, e.g. for 2 different homodimers, "ProtA dimer" & "ProtB dimer" that form a heterotetramer the systematic name becomes "2xProtA:2xProtB".

2.2.2.7. Polycistronic genes and viral polyproteins: use the common name for each protein coded by the gene, e.g. "2xMOCS2A:2xMOCS2B" (CPX-6341) or "NSP3:NSP4:NSP6" (CPX-5691).

2.2.3 <**complex recommended name**>: If a suitable GO cellular component term (see below) exists, use this. This may be used as a root term if several variants of a complex exist e.g. "mitochondrial respiratory chain complex IV variant 1". If no GO term exists use the most informative, well accepted name in the literature, something that is intuitive to the user, ending with 'complex' (e.g. "fibrinogen complex"). If no suitable recommended name exists, copy the <complex systematic name> here, replacing the colon (:) with a hyphen (-), e.g. CPX-356: ATG5-ATG12 complex. Note: A complex can only have one recommended name. Please add any variants as <complex synonym> (see section 2.5.3), e.g. haemoglobin HbA complex (CPX-2158).

2.2.4 <**Complex type**>: choose from drop-down menu. In most cases this will be "stable complex (MI:1302)". If unsure, use parent term "complex (MI:0314)".

2.2.5 <**Interaction type**>: ALWAYS "physical association (MI:0915)"

2.2.6 <**Organism**>: select the species the complex is found in. This should be the canonical species, e.g. *Saccharomyces cerevisiae* (strain ATCC 204508 / S288c) (NCBI taxonomy: 559292), rather than a particular strain the experimental evidence came from.

2.2.6.1. Exception 1: For viral strains where a species has many strains/isolates, curate the complex in the species and also for the strain(s) for which there is specific evidence. E.g. human SARS coronavirus (694009) is the species but complexes of the 2019 SARS-CoV-2 strain (2697049) have also been curated.

2.2.6.2. Exception 2: for mixed-species, host-pathogen complexes choose the pathogen as the organism. This allows searches for all complexes with a given pathogen taxId.

2.2.7 <**ECO Code**>: used to indicate the depth of experimental evidence available for the existence of the complex. There are the following options:

2.2.7.1 "**physical interaction evidence used in manual assertion**" (ECO:0000353) indicates that evidence for the complex comes from one single experiment. Complexes annotated to this code MUST have a cross-reference to experimental data in either an IMEx database (IntAct (MI:0469) or IMEx (MI:0670)), wwPDB (MI: 0805) or EMDB (MI:0936).

2.2.7.2 "**biological system reconstruction evidence by experimental evidence from single species used in manual assertion**" (ECO:0005542) is no longer used.

2.2.7.3 "**biological system reconstruction evidence by experimental evidence from mixed species used in manual assertion**" (ECO:0005543), indicates that "inference is made primarily on functional conservation between the two systems. The sequences and number of genome-encoded components are fairly conserved but some divergence is observed. The evidence must originate from a single interaction evidence."

2.2.7.4 "**biological system reconstruction evidence based on homology evidence used in manual assertion**" (ECO:0005610), indicates that "inference may be based on paralogy or orthology of the genome-encoded components and is made on sequence, composition and functional conservation between the two systems. The sequences and number of genome-encoded components are fairly conserved but some divergence is observed. The evidence must originate from a single interaction evidence." Use this term when inferring to the ortholog of a paralogue, i. e. when neither child term is appropriate. Also use this term when inferring from a complex that has experimental evidence from mixed species, i.e. is annotated with ECO:0005543. A cross-reference has to be added to the complex with <Database>="complex portal (MI:2279)", <Identifier >= "[complex_ac]" for the complex with the experimental evidence and using <Qualifier>="inferred-from (MI:1351)". The original complex must be annotated with either ECO:0000353 or ECO:0005543.

2.2.7.5 "**biological system reconstruction evidence based on orthology evidence used in manual assertion**" (ECO:0005544), indicates that "Inference is made primarily on sequence, composition and functional conservation between the two systems. The sequences and number of genome-encoded components are fairly conserved but some divergence is observed. The evidence must originate from a single interaction evidence." Use this term when only limited experimental evidence exists for a complex in one species (e.g. mouse) but it is desirable to curate the complex which has been curated in another species (e.g. human) and orthologous gene products exist. A cross-reference has to be added to the complex with <Database>=" complex portal (MI:2279)", <Identifier >= "[complex_ac]" for the complex with the experimental evidence and using <Qualifier>="inferred-from (MI:1351)". The original complex must be annotated with either ECO:0000353 or ECO:0005543.

2.2.7.6 "**biological system reconstruction based on paralogy evidence used in manual assertion**" (ECO:0005546), indicated that "inference is made primarily on sequence, composition and functional conservation between the two systems. The sequences and number of genome-encoded components are fairly conserved but some divergence is observed. The evidence must originate from a single interaction evidence." Use this term when only limited experimental evidence exists for one complex but full experimental evidence exists for a similar complex of the same species. A cross-reference has to be added to the complex with <Database>=" complex portal (MI:2279)", <Identifier >= "[complex_ac]" for the complex with the experimental evidence and using <Qualifier>="inferred-from (MI:1351)". The original complex must be annotated with either ECO:0000353 or ECO:0005543.

2.2.7.7 "**biological system reconstruction evidence based on inference from background scientific knowledge used in manual assertion**" (ECO:0005547) indicates that "The available knowledge is usually a combination of partial or weak experimental evidence where either some components are missing or no physical interaction evidence can be found but the

system is inferred by similarity to related systems in taxonomically-disparate organisms with experimental evidence. Functional studies or ligand binding evidence are often used for the reconstruction of biological systems. It does not provide physical interaction evidence but uses proxies such a ligand binding evidences to infer the presence or absence of the complex components." E.g. use this code for complexes with only pharmacological evidence or where many interaction evidences are required to 'glue together' the whole complex.

2.2.8 <**Description**> (database field: <curated-complex>): a brief, free-text description of the function of the complex. Follow the UniProt style. For example "Required for processive DNA replication and may act as a replicative helicase during DNA synthesis. Plays a central role in S-phase genome stability." (This field is repeated in the <Annotations> tab but greyed out.)

2.2.9 <**Complex properties**>: details of physical properties of the complex. Give details about the topology, varying stoichiometry, molecular weight "MW=XXX", size (Stoke's radius=XXX A) etc. (This field is repeated in the <Annotations> tab but greyed out.)

2.3 Participants Tab

2.3.1 Import **participants** following the usual rules (see IntAct curation manual): Go to the <Import…> button and type in the UniProt, ChEBI or RNAcentral ACs, hit <Search>, tick the correct participant(s) and enter their stoichiometry (if applicable) and hit <Import selected>.

2.3.2 If the complex contains a **post-processed chain** import using the PRO ID (e.g. PRO_0000005719). Do not use the PRO chain if it is the full length protein except for the signal peptide.

2.3.3 Should one or more **isoforms (splice variants)** exist, annotate to the parent protein unless either only one isoform is known to exist in the complex or different isoforms give the complex different properties. In the latter case, a separate entry should be made for each variation with detail given in "description" or "complex-properties" as appropriate (see below for details on complex variants).

2.3.4 **Nucleic acids:** ONLY enter nucleic acid as participants when they are an obligate part of the complex, otherwise use GO terms like GO:0003677 DNA binding or children of, e.g. CPX-1943 DnaA-DNA complex or CPX-17 telomerase complex. ChEBI provides generic nucleic acid terms, if no specific sequence identifiers (e.g. Ensembl or RNAcentral IDs) can be used.

2.3.5 **Small molecules/polysaccharides:** enter all small molecules ONLY if they are integral to the complex or binding to the complex is part of its function, e.g. cofactors, electron donors/acceptors etc. (e.g. ATP, H+) (e.g. CPX-3206/EBI-9689905 Acetyl-CoA carboxylase complex). Do NOT add enzyme targets as participants. E.g. ATP may be entered as cofactor if the enzyme function is NOT primarily an ATPase (e.g. CPX-2177/EBI-9008779 gyrase_ecoli) but NOT entered for straight forward ATPases (e.g. CPX-2155/EBI-9007893 mfd-uvra_ecoli, a DNA translocase). If the complex binds small molecules annotate with appropriate GO terms, e.g. GO:0005524 ATP binding. Note: If a complex binds a range of small molecules and they all belong to the same type of parent, if contextually appropriate, add the parent ChEBI term as participant rather than making lots of variants as they won't differ in protein composition (no example available yet). Where a synthetic derivative of a natural component was used experimentally, use the natural form as complex participant.

2.3.6 **Complex as participant:** Import as any other molecule but this is the only case where you must use the internal EBI-xxxxxxx ACs for complexes.

2.3.7 **Molecule set:** If a complex exists as a number of variants of paralogous protein and the expansion of all possible combinations becomes unwieldy use a "Molecule set" as interactor. Search for an existing set (e.g. EBI-16710309) or create the set of paralogous protein as a new internal molecule.

2.3.7.1 To create a new “Molecule set”, go to <Main>, <New>, <Interactor> and select Type=“Molecule set” and click <create>.

2.3.7.2 Complete the main details as follows:

- <Shortlabel>=”genename1_genename2_species”,
- <Fullname>=”uniprot-shortlabel1_ uniprot-shortlabel2” and remove duplicated “_species” extensions,
- <Organism>: select from menu
- <Type>=”molecule set”
- Add UniProt ACs using the button <Import Interactor set member>.
- <Save>

2.3.8 <**Biological role**> should be “unspecified role (MI:0499)”, except for the catalytic component of an enzyme complex, which may have <Biological role>=”enzyme (MI:0501)”, ”donor (MI:0918)”, ”acceptor (MI:0919)” (or children of donor/acceptor), “enzyme regulator (MI:1343)” or ”cofactor (MI:0682)”.

2.3.9 <**Stoichiometry**> should be added when known. Normally this will be absolute (i.e. minimum and maximum values will be the same) but there may be cases when it can be a range of values. For homo-oligomers, enter the protein twice and add stoichiometry = 0. If you are annotating internal binding features (see 2.3.10.3), manually expand the complex so every participant has stoichiometry = 1 and add unique binding feature (e.g. Hemoglobin HbA CPX-2158/EBI-9008420).

2.3.10 <**Features**>: Any feature known to be involved in the complex formation or function should be mapped to the underlying protein sequence, as given in the source database, and cross-referenced to InterPro when possible. Binding features, where known, should also be added to nucleic acids small molecules.

2.3.10.1 <Feature full name> can be ignored.

2.3.10.2 If a **PTM** is required for complex activation, curate this as a feature with <Feature type> = “[MI term for PTM]” (e.g. “opser” for an o-phosphorylated serine), <Feature role> = “prerequisite-ptm” and the position in the range field (if known). If you know which region on another protein this PTM binds to also create the binding feature (see below). Enter further details in <Complex properties>. E.g. SMAD2-3-4 complex (CPX-1) or translocon complex (CPX-3055).

2.3.10.3 Participants known to **directly interact** within the complex should be linked by creating a <New Feature> with <Shortlabel>=”[gene_symbol / RNAcentral_ac / ChEBI_name] binding region” and <Feature type>=”binding region”. For isoforms and PRO chains just use the gene names unless more than one isoform or chain from the same canonical protein are included in the complex, in that case add the isoform or PRO chain name (e.g. Caspase-3 CPX-970/ EBI-12735071). For viral polyproteins use <Shortlabel>=”[protein name within polyprotein] binding region” (e.g. SARS-CoV-2 polymerase complex, CPX-5742).

2.3.10.4 Ignore point mutation data from crystals as interaction sites unless highlighted as specifically important; concentrate on binding regions.

2.3.10.5 If the precise **binding region is unknown** (e.g. for all small molecules) enter <range>=”?-?”.

2.3.10.6 If the **binding region is known** add the amino acid or nucleic acid residue numbers as <range>. If this binding site also matches an **InterPro domain**, add the InterPro **Domain** ID as a <Feature Xref> with <Database>=”InterPro (MI:0449)”, <Identifier >= “[InterPro_Id]”, <Qualifier>=”identity (MI:0356)” (e.g. dimerisation domain of CPX-2883 PDGF receptor alpha-beta - PDGF-BB heterotetramer).

2.3.10.7 If the binding region is likely to be the full length of the protein or the PRO chain use <range>=”?-?” unless the protein is very short and there is proof that the whole protein interacts, e.g. the 40 aa amyloid-beta chains (EBI-13943327/CPX-1069).

2.3.10.8 If the same feature range binds more than one complex participant, make one features and concatenate all gene names for all participants this participant binds, e.g. ”ProtA_ProtB binding region” (e.g. SARS-CoV-2 polymerase complex, CPX-5742).

2.3.10.9 If a binding region in a complex is formed by two pieces of sequence from two different components of the complex (= composite binding sites), two separate features need to be created and linked together in the Editor. The Shortlabel for the binding feature on the binding partner should contain the concatenated gene names of the composite binding site, e.g. ”ProtA_ProtB binding region”.

2.3.10.10 If the **stoichiometry** of a participant with a binding feature is **greater than 1** each molecule needs to be added individually with stoichiometry = 1 and annotated with their specific binding regions. [If the participant(s) are not manually expanded the feature notations will be ambiguous in the files and the participant(s) and feature(s) will be matrix-expanded in ComplexViewer providing a misleading topologies.]

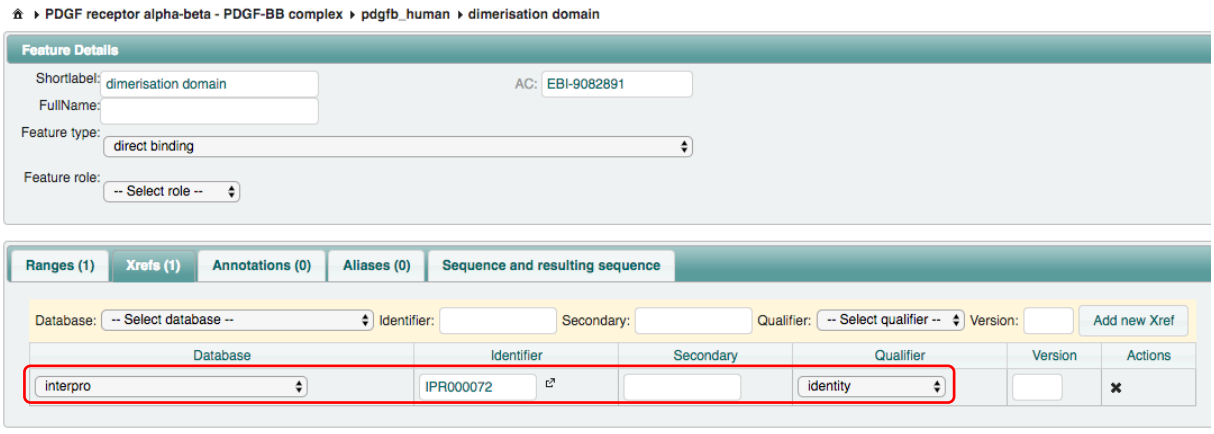
2.3.10.11 If the complex undergoes a conformational rearrangement during activation curate the internal binding features according to the active topology and provide details of the rearrangement in <complex-properties>.

2.3.10.12 At the complex level link the binding features of the two participants by checking the appropriate <Feature> boxes, then click ‘Link features’ and <Save>.

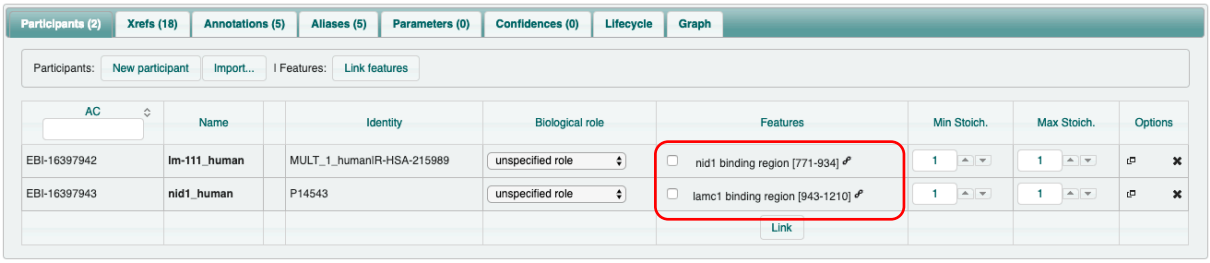
2.3.10.13 Features on complexes as participants:

- Annotate the range or PTM at the participant level in the usual way.
- In the <Refers to participant> dropdown menu chose the complex participant that this feature refers to.
- The referred participant must always refer to the lowest level component that bears it (e.g. the protein), even in complexes formed by several levels of subcomplexes. Examples laminin-nidogen complex (EBI-16397934/CPX-1265) and CMG-Pol epsilon complex (EBI-16706245/CPX-1556).
- The short label should always refer to the complex participant that the features links to.
- For binding features, link the ranges at the complex level.
- If it is unknown which complex participant binds to another protein or complex don’t create a binding feature to the complex.

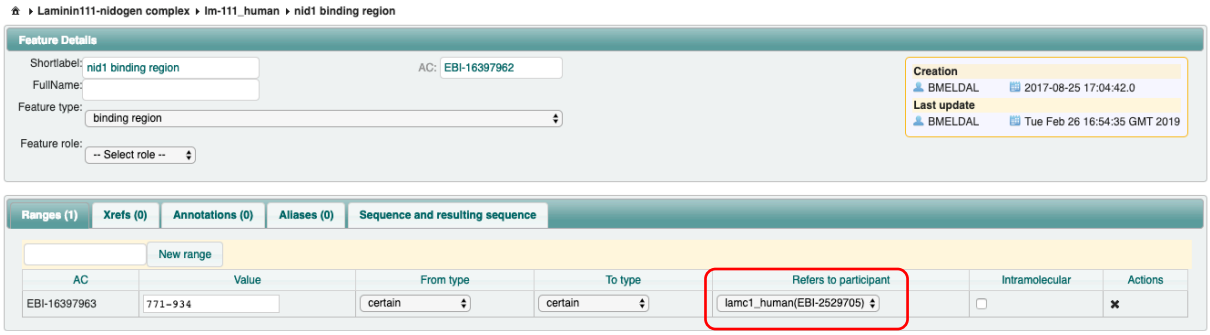
Feature with InterPro xref (Feature ID: EBI-9082891):



Complex as complex participant (Laminin-nidogen complex, EBI-16397934/CPX-1265):
Shortlabel construction and binding link of features:



Chross-reference to complex participant at participant feature level:



2.4 Annotations tab

2.4.1 <**Description**>: see Header section (not editable here)
2.4.2 <**complex-properties**>: see Header section (not editable here)
2.4.3 <**complex-assembly**>: e.g. Homodimer, Heterohexamer… Only add if stoichiometry fields are all filled in. If the complex consists of heterogeneous cleavage products of the same gene product, e.g. CPX-362 heparanase, it is a heterodimer etc. In complexes with complexes as participants the assembly refers to the total protein count.
2.4.4 <**ligand**>: enter known ligands and give references where possible, e.g. maltose (CHEBI:17306). Focus on natural molecules or common drugs, e.g. acetylcholine (CHEBI:15355) and nicotine (CHEBI:18723) in CPX-2179. Please use separate fields per ligand.
2.4.5 <**disease**>: enter the disease name with EFO AC in square brackets, followed by a brief description of the disease in which the COMPLEX is involved in. This should match the information from the EFO/HPO/Orphanet xrefs (see below). Only use this field if the whole complex (possibly in mutated form) is involved in the disease. Use separate fields per disease. E.g. CPX-2158: “Heinz body anemia [Orphanet:178330]: a form of nonspherocytic hemolytic anemia of Dacie type I.”
2.4.6 <**agonists**>: focus on natural molecules or common drugs
2.4.7 <**antagonist**>: focus on natural molecules or common drugs
2.4.8 <**comment**>: If it doesn’t fit in any other field but you feel it’s important to advise the user.
2.4.9 <**rapid curation**>: a complex can be curated without adding extensive cross-references. In that case, please add this annotation so the complex can be revisited and cross-references added.

2.5 Alias tab

2.5.1 <**complex recommended name**>: see Header section (not editable here)
2.5.2 <**complex systematic name**>: see Header section (not editable here)

2.5.3 <**complex synonym**>: enter all other possible names the complex is known by or can be described as. For enzymes, IntEnz often has a useful list of synonyms. Only enter the IntEnz 'systematic name' if it is specific to the complex, not the whole enzyme family. Use separate fields per synonym.

2.6 Xrefs tab (cross references)

2.6.1 "**[Interaction_database]**": All complexes require a link to a database that demonstrates the existence of the complex, unless it has been inferred by the curator based on background scientific knowledge (ECO:0005547). The precise annotation depends on the ECO code assigned to the complexes (see section 2.2.7). For complexes with experimental evidence (ECO:0000353 & ECO:0005543) the annotation follows the pattern <Database>="intact (MI:0469)", <Identifier >="EBI-xxxxxxxx", <Qualifier>="exp-evidence". For other source databases (e.g. MatrixDB, DIP, IMEx), replace IntAct identifiers with relevant identifiers from another database. Choose a sensible rather than exhaustive list of experiments and avoid using evidence from high-throughput experiments (unless they have been validated in a second step, such as MI:1356 [validated two hybrid]). Experimental evidence must be available for the whole complex in one interaction. An exception exists for complexes that have been crystallised but not been curated in an interaction DB: they will have xrefs to wwPDB (section 2.6.4) or EMDB (section 2.6.5) and require an appropriate ECO code (section 2.6.2) but no interaction DB xref. For complexes that have been inferred by homology (ECO:0005610, ECO:0005544 & ECO:0005546) use pattern <Database>="complex portal (MI:2279)", <complex_ac >="CPX-xxxx [of complex with experimental evidence]", <Qualifier>="inferred-from".

2.6.2 "**evidence ontology**": The ECO code is automatically imported from the header into the xref table when you save the complex. However, if you have to change the ECO assignment while curating the complex, you need to change the code in the xrefs table as well. See section 2.2.7 for details.

2.6.3 "**pubmed**": Papers which relate to the function of the complex as a whole or review articles. <Database>="pubmed (MI:0446)", <Identifier >= "[pubmed_id]", <Qualifier>="see-also (MI:0361)"

2.6.4 "**wwPDB**" when the complex has been crystallised. Add all applicable crystals from wwPDB. <Database>="wwPDB (MI:0805)", <Identifier >= "[pbd_ac]", <Qualifier>="identity (MI:0356)". If the crystal is only part of the complex, use <Qualifier>="subset (MI:2179)" but do not add cross-references to monomers. If the constructs come from different species but orthologues genes add the AC as cross-reference with <Qualifier>="identity (MI:0356)".

2.6.5 "**EMDB**" when the complex has an EM structure. Add all applicable structures from EMDB. <Database>="emdb (MI:0936)", <Identifier >="[emd_ac]", <Qualifier>="identity (MI:0356)". If the structure is only part of the complex, use <Qualifier>="subset (MI:2179)". If the constructs come from different species but orthologues genes add the AC as cross-reference with <Qualifier>="identity (MI:0356)".

2.6.6 "**Reactome**": for human ONLY. If the components of the complex are an **exact match** use <Database>="reactome (MI:0467)", <Identifier>="[R-HSA-xxxxx]", <Qualifier>="identity (MI:0356)". If the components **differ** because Reactome have included **a set of paralogous components** and we curated them as a set of variants of a complex use <Qualifier>="see-also (MI:0361)". As Reactome provide different IDs for the same complex in a different cellular compartment you may have more than 1 Reactome xref per complex. Beware, there are instances in Reactome where the complex in question does not exist in a diagram as it is part of a bigger complex but the appropriate ID can be found as a complex component.

2.6.7 "**ChEMBL**" when the complex exists in ChEMBL. <Database>="chembl (MI:0967)", <Identifier>="[CHEMBL_ac]", <Qualifier>="identity (MI:0356)".

2.6.8 "**IntEnz**" when an enzyme complex is described the appropriate EC number should be added. <Database>="intenz (MI: 0585)", <Identifier >="[intenz_id]", <Qualifier>="identity (MI:0356)". E.C. number must finish with a digit (not a full stop) so that the hyperlink on the website works.

2.6.9 "**EFO**" for complexes involved in disease. <Database>="efo (MI:1337)", <Identifier >= "[efo_id]", <Qualifier>="see-also (MI:0361)". Beware, EFO is a collection and retains original IDs from imported DBs, e.g. Orphanet, HP, etc. Use the ID supplied by EFO but <Database> = "efo" at all times. If no EFO term exists for your disease term, please request a new term via the EFO 'request' form at http://www.ebi.ac.uk/efo/ quoting as much information as possible, e.g. if it already exists in Orphanet, OMIM or the Disease Ontology, etc.

2.6.10 "**GO**": annotate cellular component (both classes, a "*subcellular location*" that is a child of "*GO:0005575 cellular_component*" and a *complex* that is child of either "*GO:0032991 protein-containing complex*" or "*GO:0099080 supramolecular complex*"), biological process and molecular function for each complex as far as it is known. <Database>="go (MI:0448)", <Identifier >= "[GO:xxxxxxx]", <Secondary>="[GO name]", <Qualifier>="[component, process or function, see below]". <Secondary> and <Qualifier> should be populated automatically from the ontology file but this sometimes fails.

2.6.10.1 <Qualifier>="**process/function**": should refer to the complex as a whole, not the component parts. You can add as many functions and processes as are applicable to the complex. This includes "X binding" annotations if compound X is NOT a complex participant. You may wish to request a new GO term via the GitHub go-ontology tracker (https://github.com/geneontology/go-ontology/issues). Please provide the following information to the Editors: term name, definition, PMID, relationships, synonyms (define whether it is exact, narrow or broad).

2.6.10.2 <Qualifier>="**component**": choose a GO term that refers to the exact complex. If the exact complex does not exist in GO use the most appropriate parent term. You may request a new complex term via the GitHub tracker (see above). Be aware that GO will not create component or species-specific terms. Also add the cellular location of the complex (e.g. plasma membrane or nucleus) using the GO 'component class'.

2.6.10.3 **<Evidence Code> and <Pubmed Identifier>**: we follow Gene Ontology Annotation rules as follows:

2.6.10.4 "**components**": can have <evidence code>="ECO:0000353[IPI] or ECO:0005543" with <reference>="[PMID]" if experimental evidence exists, <evidence code>="ECO:0005544, ECO:0005546 or ECO:0005610" with <reference>="[CPX-xxxx]" from complex with experimental evidence if complex has been inferred by homology, or <evidence code>="ECO:0005547" with <reference>="[PMID]" if complexes has been inferred from curator background knowledge (see table below).

2.6.10.5 "**function/process**": can have <evidence code>=" ECO:0000353[IPI] (for binding terms only), ECO:0000314[IDA], ECO:0000315[IMP] or ECO:0000269[EXP]" with <reference>="[PMID]" if experimental evidence exists, <evidence code>="ECO:0005544, ECO:0005546 or ECO:0005610" with <reference>="[CPX-xxxx]" from complex with experimental evidence if complex has been inferred by homology, or <evidence code>="ECO:0005547" with <reference>="[PMID]" if complexes has been inferred from curator background knowledge (see table below).

2.6.10.6 Note: If only one subunit is mutated or knocked out to "prove" a function use ECO:0005547 (modelled by background knowledge). Likewise, if the function of complex components is shown in separate experiments, even if they are done using the same methods and reagents, use ECO:0005547.

2.6.10.7 Always annotate to the most granular term in the ontology as possible. An exception occurs when there is experimental evidence only for a parent term but

circumstantial/indirect evidence for a child term. In this case annotate separately to both terms using the appropriate ECO codes. It will generate 2 lines in the GO annotation file.

Class	GO ECO	GO Ref
Component	ECO:0000353/5543	PMID
Function/Process	ECO:0000353/314/315/269/5543	PMID
Any	ECO:0005544/546/610	CPX-xxxx
Any	ECO:0005547	PMID

2.6.10.8 **Annotations to individual components:** if specific annotation to gene products (GPs), such as "X binding" or "regulator activity", are missing in GO please use Protein2GO to add such annotations directly to the GPs.

Summary Table: Xref topic and qualifiers used

Cross Ref	Qualifier
ChEMBL	identity
ECO	none
EFO	see-also
GO	Should be added by the system but can fail: should be function, process or component
IntAct/IMEx/DIP	exp-evidence
Complex Portal	Inferred-from (when used as reference complex for evidence attribution)
IntEnz	identity
PubMed	see-also
Reactome	identity or subset
wwPDB/EMDB	Identity or subset

3. Checking

Checking and correcting a complex is done in the same way as a normal curation:

3.1. When a complex is ready for checking, click on the yellow <Ready for checking> button near the top right corner. The complex will turn yellow on your dashboard and appear in the reviewer's list on their Dashboard.
3.2. If the reviewer rejects your complex, the complex will turn red on your dashboard.
3.3. The reviewer's comments occur in the yellow <To be reviewed> box above the Header fields, as a field in the Header and in the Annotation tab (greyed out).
3.4. Make your changes and leave any comments in the <Correction comments> box (next to the <Ready for checking> button). Your comments will appear above the Header fields, as a field in the Header and in the Annotation tab (greyed out).
3.5. Send the complex back to the checker by clicking on the yellow <ready for checking (again)> button. Your complex will turn yellow.
3.6. Once the checker accepts your complex it will disappear from your dashboard. You can see it if you tick the <ready for release> button above the dashboard.

4. Cloning a complex

4.1 Cloning a complex from an interaction

You can clone any interaction and directly make it a complex. Go to:

4.1.1 <Tools>, <Clone as biological complex>

4.1.2 <Save>

Saving triggers the following:

4.1.3 Clones all <Participants> with their <Features> (e.g. binding regions, stoichiometry etc.)

4.1.4 Clones the <Interaction type>

4.1.5 Clones all <Annotations>. Check and delete those that are not applicable to the complex.

4.1.6 Clones all <xrefs> except those with qualifier="identity". Check and delete those not relevant to the complex, such as the psi-mi and crystal xrefs.

4.1.7 Automatically selects <Complex type>="complex". Change if appropriate, e.g. "stable complex".

4.1.8 Automatically adds a cross-reference to the interaction evidence: <Database>="intact", <Identifier>="[Interaction_AC]", <Qualifier>="exp-evidence"

4.1.9 Automatically selects <ECO code>="physical interaction evidence"

4.1.10 Automatically adds a cross-reference <Database>="evidence ontology", <Identifier>="ECO:0000353", <Qualifier>="[none]". Change if necessary, e.g. if combining evidences or if evidence was derived from a mixed species experiment.

4.1.11 Now fill in the rest of the information!

4.2 Cloning a complex from another complex

You can clone a complex the same way you can clone an interaction. This makes your life much easier when two complexes are very similar, e.g. differ by a participant or making it for related species. Sometimes, experimental evidence does not exist in the model organism you are focusing on but there is evidence for an orthologous complex in a closely related species, e.g. evidence is in rat but you are focusing on human. We are also using this feature to clone all human complexes systematically to their mouse orthologues.

You can use the cloning function in the Editor to create the inferred complex easily:

4.2.1 Go to <Tools> <Clone Complex>

4.2.2 **NEW:** Change one participant (UniProt Ac or stoichiometry) and <save>. You must change the participant list to make the complex unique in order to save changes.

4.2.3 Fix xrefs: Cloning automatically strips all xrefs that have the <qualifier>="identity", such as references to ChEMBL, wwPDB, EMDB, Reactome and E.C. numbers but will keep the InterPro xrefs on the participant binding features. Check and delete those not relevant to this complex and add in any missing xrefs, such as E.C. numbers that don't get cloned.

4.2.4 If applicable, change the ECO code manually in the Header section and cross-references

4.2.5 If applicable , reference the original complex with <Database>="complex portal", <Identifier>="[CPX-xxxx]" and <Qualifier>="inferred-from"

4.2.6 Re-link all features

4.2.7 Make all other appropriate changes

4.3 Additional guidelines for cloning to a different species (in any order):

4.3.1 Change the participants to the inferred species:

4.3.1.1 Click on the participant <name> to go to the detailed participant view
4.3.1.2 Click on the <Import> button
4.3.1.3 Enter the UniProt AC for the protein of the inferred species and click on <Search>
4.3.1.4 Tick the correct line (usually already done by default) and click on <Import selected>
4.3.1.5 Click <Save>
4.3.1.6 If the participant had a defined binding region make sure this applies to the inferred species or correct accordingly (align the two sequences)
4.3.2 Link the binding regions (the links disappear when cloning)
4.3.3 Change the <Shortlabel> to the inferred species
4.3.4 Change the evidence ontology in the Header section and cross-references to "biological systems reconstruction evidence based on homology evidence used in manual assertion (ECO:0005610)", "biological systems reconstruction evidence based on paralogy evidence used in manual assertion (ECO:0005546)" or "biological systems reconstruction evidence based on orthology evidence used in manual assertion (ECO:0005544)"
4.3.5 Reference the original complex with <Database>="complex portal", <Identifier>="[CPX-xxxx]" and <Qualifier>="inferred-from"
4.3.6 Change all cross references and UniProt ACs in the Annotations to the inferred species
4.3.7 Delete/correct all inappropriate Annotations and xrefs, such as disease annotations

5. New Complex Versions

5.1 A new version of a complex can be created manually when evidence emerges
5.1.1. that the complex has a different composition (adding or removing participant(s))
5.1.2. that the complex has a DIFFERENT function.
5.2 Updates for function are made in line with UniProt and the GO Molecular Function would be from a different branch. Changing the annotation to a more granular term does not warrant a new version.
5.3 When creating a new version, the reason must be entered in the pop-up box. Please add PMIDs for the new evidence wherever possible.
5.4 Note: The information you enter into the pop-up box is only saved in the old entry. Optional: You can copy the comment it into a <remark-internal> annotation of the new version.

6. Obsoleting (and merging) a Complex

6.1. A complex may be obsoleted if
6.1.1. Evidence emerges that it doesn't exist at all (e.g. CPX-3043 SUMO-E2 ligase emerged to not be a complex at all as SUMO is a modification, not a participant).
6.1.2. Evidence emerges that the complex composition is different and therefore becomes the same as another entry (e.g. SUMO1 was removed from CPX-4922 activated SUMO1-E1 ligase complex making it identical to CPX-2161 SUMO activating enzyme complex)
6.2. Click <on hold> button and put the reason into the text box:
6.2.1. "obsoleted complex without replacement", or
6.2.2. "complex merged into different entry"
6.3. Add <obsolete complex> annotation and add reason into the text box, e.g. "SUMO is not a participant but a modification.", or "Merged into CPX-2161 because SUMO is not a participant but a modification."
6.4. It is possible to create a new version and then merge complexes into this, e.g. we removed SUMO1 as participant creating CPX-4747.2 E3 ligase (RANBP2), then CPX-4922 SUMO2 E3

ligase (RANBP2) was merged into CPX-4747.2 as the only difference was that this entry had SUMO2 as participant.

NB:

- <on hold> removes the entry from the release process
- <obsolete complex> makes the entry read-only. This ensures that it can no longer be altered but it can still be searched for in the editor just like an old version (see above)

7. Importing a complex from an xml file

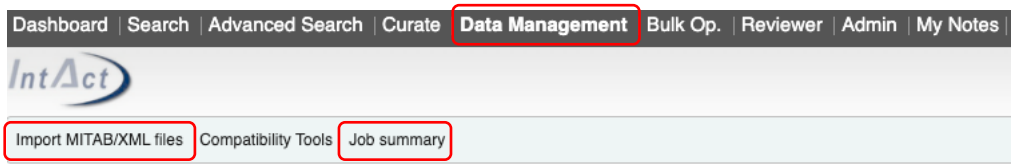
Notes:

- The importer compares the proteins in the xml file and their stoichiometry to existing entries. It ignores non-protein molecules.
- The import will fail if the entry has identical proteins and stoichiometry to an existing entry
- The import will add an <annotation topic>="caution" to the new entry if the entry has identical proteins but with different stoichiometries compared to an existing entry. The Caution adds a "warning triangle" next to the entry on the dashboard and an annotation in the "Annotations" Tab.

7.1. Check if the complex is already being curated: in the Editor, using "Advance Search" -> "Molecule", search with the UniProt AC of one of the proteins and check if there are entries in the "Complexes" column. If yes, check the complex(es) to make sure they are different. If the complex exists, check if it needs updating, if it doesn't yet exist, proceed as follows:
7.2. Go to "Data Management" in the top menu
7.3. Select "Import MITAB/Psi-XML File"
7.4. Select the second "Choose file" button and select the file from your computer
7.5. Click on "Upload biological complexes"
7.6. Go to "Data Management" – "Job Summary" and scroll down to the complex section to find your import (you can search by job number but the last job will be on the top)
7.7. Additional information about the import is sent to your email including any detailed error messages.
7.7.1 If the upload worked it will have status ="COMPLETED"
7.7.2 If the complex already exists will have status = "FAILED".
7.7.3 If the complex is identical except for the protein stoichiometry it will add a comment to the messages and a caution to the entry (see above).
7.8. If the import was successful there should be a new complex in the "All the Biological Complexes" dashboard
7.9. Click on the entry and claim ownership via the yellow button on the top right
7.10. Complete the entry manually

PDBe have created a tool that creates XML files from wwpdb files. It autofills a number of fields. Please check all autofilled fields, then manually add:

1. Complex description
2. Complex properties (optional)
3. Complex assembly
4. GO cross-references: Check and add references & ECO codes
5. IntAct cross-reference (recommended): check if there is experimental evidence in IntAct (www.ebi.ac.uk/intact) and add it as cross-reference. In this case, the IntAct cross-reference will need qualifier = "exp-evidence" and the wwpdb cross-reference qualifier = "identity"
6. Optional: add any other annotations as per normal curation.



8. Additional points

8.1. Enzyme/substrate and receptor/ligand complex that we DO curate

We will curate enzyme/substrate and receptor/ligand complexes if any of the entities consist of more than one chain and the binding of the substrate or ligand is obligate for the formation of the enzyme or receptor complex, respectively. E.g. EBI-6477643 maltose transporter, or EBI-9080360 PDGF-AA receptor alpha complex.

8.2. Complex Variants

If variant forms of a complex exist i.e. the same functional unit can exist in alternate forms with differing macromolecular composition, these should be curated as separate objects. If the variants have well-accepted names these may be used as the primary name, e.g. glutamate decarboxylase 1 complex (EBI-9293677, a homodimer), glutamate decarboxylase 2 complex (EBI-9293944, another homodimer), and glutamate decarboxylase 1/2 complex (EBI-9491125, the heterodimer of the aforementioned complex participants). If not, then use a consistent name throughout, qualified by variant 1, variant 2, e.g. TRAMP complex variant 1 (EBI-2352894), TRAMP complex variant 2 (EBI-2352906).

9. Reviewers Checklist

Spelling	Check carefully for hyphens, they must be the short type.
	Are all characters ASCII-compliant?
Shortlabel	Does it reflect the name?
Recommended name	Only one allowed.
	Have all the gene symbols been updated after cloning?
	Are hyphens used between gene symbols?
Systematic name	Only one allowed.
	Have all the gene symbols been updated after cloning?
	Does the syntax match the species conventions?
	Are colons used to list gene symbols?
	Are gene symbols in alphanumerical order? (number before letters)
Species	Same at complex and participant level? Has it been updated after cloning?
Complex type	Correctly assigned?
Interaction type	Must be “physical association”
ECO	Has the same code been used in the header and the xref table?
	Has it been updated after cloning?
Description/Properties	If complexes are quoted are they using the CPX- AC?
Stoichiometry	If all proteins have known stoichiometry has complex-assembly been added to the annotations?
Participants	Are all biological roles chosen appropriately?
	Have all binding links been added?
	Have gene symbols been used for range features?
	Have InterPro links been added to ranges?
Xrefs	Have the correct identifiers and qualifiers been used?
	Are GO lines complete and ECO codes used according to rules?
Annotations	Have disease identifiers been added to the text in square brackets?
	Have UniProt or CheEBI ACs been added to ligands, agonists and antagonists?
Aliases	Have all gene symbols been updated after cloning?
	Are hyphens used between gene symbols?