New curation guidelines for complex curation

New identifiers:

- All references to complex must now be made to the CPX-xxx ACs.
- Lots of updates in manual: mentions in Description and Properties, inferred-from xrefs, GO xrefs
- Also new source DB “complex portal”
- Complexes as participants: still use EBI-xxxxxxx AC!

Versioning:

Every complex now has a version that’s being created and can be changed manually in the Editor. Publically, only the canonical AC is displayed but version are kept in the original files.

New version when evidence emerges

- that the complex has a different composition (adding or removing participant(s))
- that the complex has a DIFFERENT function. Updates would have to be 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.
- *When stoichiometry changes???*
- *When protein instance changes btw canonical/isoform/PRO chain???*

Complexes as participants:

- We’ve implemented the use of complexes as participants in Interactions and Complexes, e.g. CPX-1265 laminin111-nidogen complex.
- In some cases, there are several layers of complexes as participants of complexes, see CPX-1556 CMG-Pol epsilon complex.
- If 2 homo-oligomers are forming 1 supercomplex: stoichiometry based on all proteins, interaction type “direct”.
- Add parent GO CC term for the big -osomes to make them findable
- *Use square brackets when combining complexes in a supercomplex entry??? Not actually done, see CPX-1265 or CPX-1556!!!* → Just use alphanumerical order
- *Linked features between subcomplex participants, see below.*

Experimental evidence:

- Experimental evidence must be available **for the whole complex in one interaction**.

ECO:

- ECO:0005542 (>1 expt evidence required): do not use any more
- ECO:0005543, ECO:0005610 : The evidence must originate from a single interaction. Previous: mixed evidence allowed.
- Inference is made primarily on sequence, composition and functional conservation
- Use ECO:0005610 when inferring from a complex that has experimental evidence from mixed species, i.e. is annotated with ECO:0005543.
- If only one subunit is mutated or knocked out to “prove” a function use ECO:0005547 (modelled by background knowledge) in the GO annotation.

When to use PRO chain IDS:

- If the complex contains a **splice variant** 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.

Interaction DBs as expt evidences:

- DIP-IMEx now imported into IntAct DB → use EBI- ACs as they are searchable (same for evidences from MINT, UCL or Matrix DB) and annotate with qualifier=”expt-evidence”
- If evidence is in PDB or EMDB no evidence in IntAct is necessary - choose one xref and annotate with qualifier=”expt-evidence”
- EMDB xrefs: don’t include EMB- prefix.

IntEnz xrefs:

- Need to stop with a number not “full stop”.

GO terms:

- GO:0043234 protein complex now merged into parents GO:0032991 protein-containing complex (was “macromolecular complex”, now synonym)
- Annotations have been reviewed and cleaned up. Please adhere to the rules so our GPAD remains valid and annotations can all go out.
- New GPAD does not include inferred annotation that have no ref or ECO code (legacy data)

GO Annotations to individual components:

- If specific annotation, such as “X binding” or “regulator activity” are missing in GO please use Protein2GO to add such annotations directly to the Gene Products.

Systematic names for Isoforms and PRO chains

- *Just use gene symbols???*

- *Use UniProt AC with -X for isoforms???*
- *Use UniProt AC with _PRO_xxxxxxxx chain ID for PRO chains???*

Question arises as Noe could autogenerate the systematic name but we'd need a clear rule for this. She's also used the systematic name for some QC checks.

Systematic name to be abbreviated with recommended name:
Have we actually done that? → no

GO syntax checks:

- These are sent to intact-help with every monthly GO release. Include obsoleted and merged terms. These terms need manual checking by a CP curator every release. I have done a large tidy-up so the list shouldn't be too bad every month.

Features of subcomplexes:

Use cases:

- We want to annotate the binding regions of two **participants** (protein, nucleic acids or small molecules) where at least one, but possibly both, are members of complexes within an interaction or complex.
- We have feature information (e.g. mutations, PTMs, tags) for **participants** that are members of a complex where the complex is the primary interaction participant.

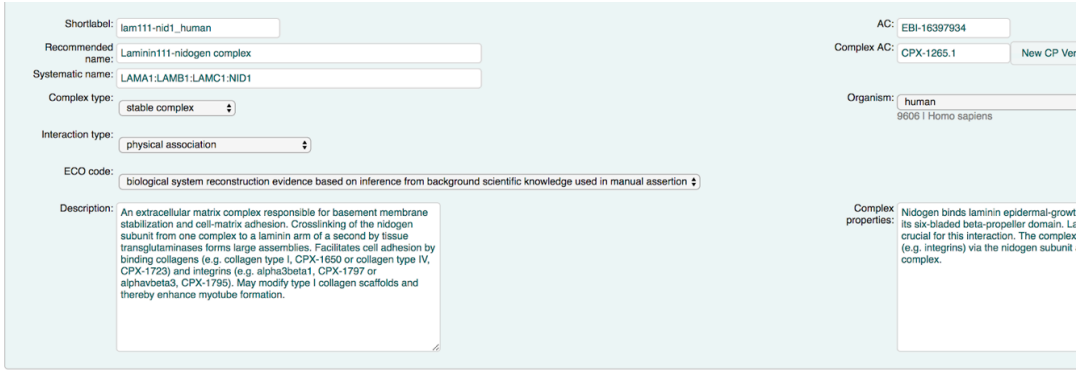
Proposed annotation guidelines:

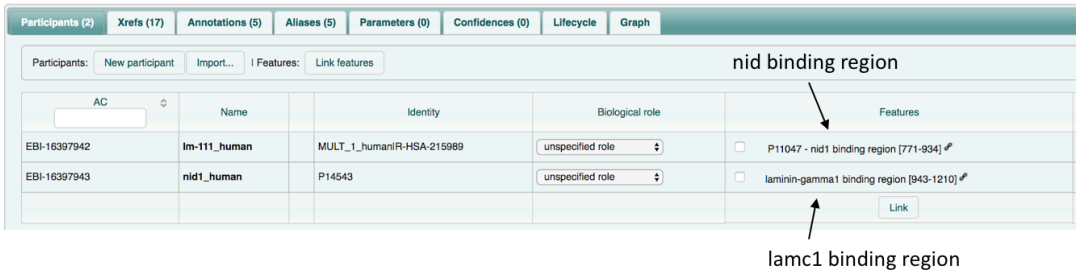
- If the participant binding a participant that is part of a complex (that is an interactor itself) is a protein, nucleic acid or small molecule the feature annotation rules remain unchanged.
- If the participant binding another participant is a complex participant itself the following annotation rules apply:
- Annotate the range at the participant level.
- Add a cross-reference to the range feature as follows: database="intact", identifier="EBI-xxxxxxx", qualifier="complex component" to identify the appropriate participant AC of the complex component. The AC can be retrieved from the complex instance in the Editor. This xref allows the editor to identify where to take the appropriate sequence ranges from and **there must be only one range** per feature.
- All shortlabels should be constructed in the usual way of "<gene_symbol> binding region", "<RNAcentral_name> binding region" or "<ChEBI_symbol> binding region". [Do we need UniProt ACs in the shortlabel?]
- Link the ranges at the interaction/complex level.

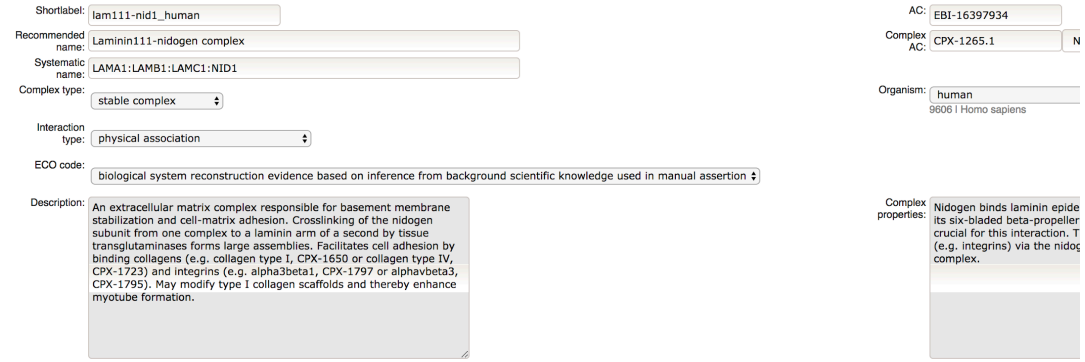
- If a binding region in a complex is formed by two pieces of sequence from two different components of the complex, two separate features need to be added and linked together in the editor.
- The xref must always refer to the lowest level component that bears it (e.g. the protein), even in complexes formed by several levels of subcomplexes.
- Examples laminin-nidogen complex (EBI-16397934/CPX-1265) and CMG-Pol epsilon complex (EBI-16706245/CPX-1556). IntAct?
- For other feature annotation to subcomplex participants use the same procedure except no linking is required. Use the appropriate shortlabel rules for the type of annotation feature.

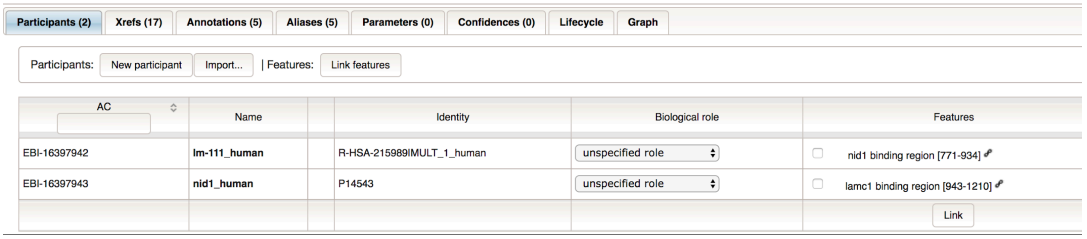
Examples below:

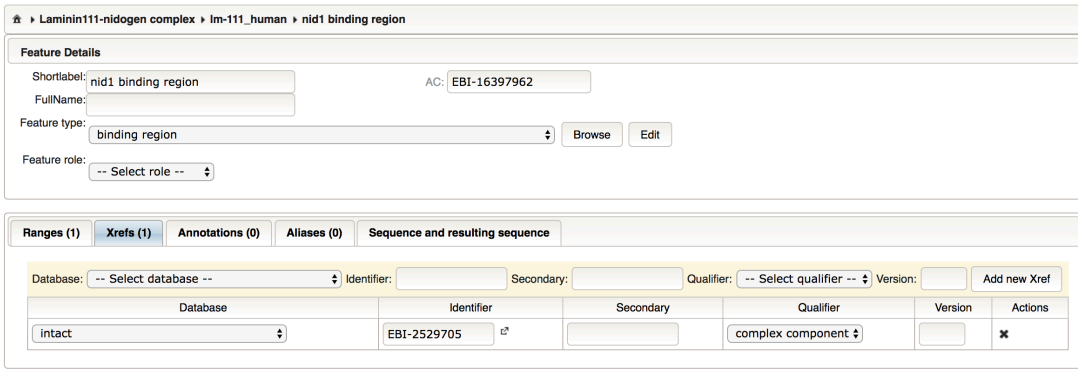
Complex Example:





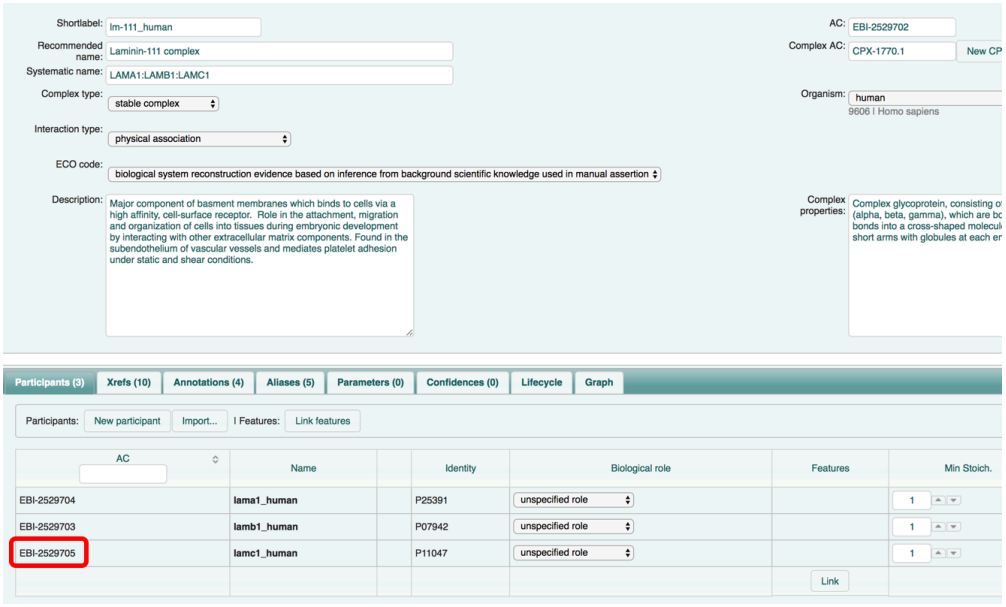






Add gene name as <secondary> to make curation/checking easier?

We have to use the internal EBI- ID for the interactor as we need to point to the participant in the laminin-111 complex:



Interaction example: