LLPS DB Annotation guide

Please read and adhere to the guidelines below. These describe how to complete all fields in the example document.

While the following should address the most important aspects, if you feel something is not clear, or certain things should be added/modified, contact Rita at pancsa.rita@ttk.mta.hu !

Important! We aim to describe 'systems', i.e. a minimal set of proteins (and potentially other biomolecules) that in themselves are able to undergo liquid-liquid phase separation (LLPS). Hence, it is often insufficient to annotate a single protein alone, if that protein absolutely requires other proteins as well for forming membraneless organelles. In these cases, you should annotate all essential protein components of the LLPS system, and you need to fill out a separate annotation document for all of them. If a single protein is enough for LLPS but it requires some other biomolecule (such as DNA, RNA, etc), these should be described in a single document, defining the non-protein components in the respective fields (fields F and N).

Some fields are required to be the same for all components of the system. These are marked in notes at the end of the field description.

A) Common name(s) (as named in the article)

The name(s) that is usually used in the articles, multiple names should be separated by commas.

B) UniProt accession

Either a single character string, or in case of an isoform, a UniProt accession followed by '-' and a number, defining the isoform number (matching the UniProt definitons).

C) Characterization of region(s) mediating LLPS

A free text, but short concise description of the protein regions that mediate the LLPS. If there are multiple regions in the same protein, list all, separating them by semicolon.

a. Indicate if the given region is a domain or a disordered region (IDR) (maybe also low sequence complexity (LC) region or prion-like domain (PLD)), indicate location (N- or C-terminal) if mentioned by the authors
b. If LC, which amino acids are enriched?
c. Indicate RNA-binding domains by name (RRM, KH, RGG…)
d. Indicate if the given domain is an oligomerization domain based on the article (write "dimerization/homooligomerization domain" in brackets after domain name

Example: "N-terminal N/Q-rich IDR", and "PB1 domain (oligomerization); UBA (polyUbi chain binding)"

D) UniProt region boundaries

Precise UniProt region boundaries of the previously described (field C) LLPS driver regions, separated by semicolon. **WARNING: the numbering in the paper might not match the UniProt numbering. Pay attention to this, especially with sequences different from the canonical isoform, in case of proteins with signal peptides, N-terminal tags, or where the initiator M is cleaved!**

Example: “100-200; 255-370”

E) LLPS region reference PMID

From which article did you take the region boundaries? Provide **a single** PubMed ID (only numeric characters).

F) Binding partners of LLPS

Provide important binding partners/competitors for LLPS. Define each partner in a new row. Number each partner - if known, follow the order of importance (“1)” is most important). Add in brackets if the partner is “strictly required for LLPS”. Known negative regulators/competitors of LLPS should also be listed here continuing the numbering of partners, but add “negative regulator”, “competes for RNA(or-whatever partner) binding” after their name in brackets. Use tags (marked in blue above) if possible, however they are not mandatory (i.e. you can provide more descriptive notes).

Example:

1) GRB2 (strictly required for LLPS)
2) SOS1 (strictly required for LLPS)

G) Determinants of phase separation and droplet properties

Provide environmental or other features which were demonstrated to affect LLPS (phosphorylation state, modification state, temperature, salt concentration, crowding agent concentration, protein concentration of …, protein density in membrane of …, pH, ionic strength, valency of …, stoichiometry of the components, presence of metal ions (…type_of_ion), molecular affinities between the components, DNA concentration, RNA concentration etc.) in a numbered manner. Substitute ‘…’ with any free text that adequately describes the feature. If you cannot find the tag needed in the above list, feel free to add your own short and concise definition. Define multiple descriptions in separate lines. Exact definitions PTMs and alternative splicing with known effects are **not** to be listed here but other dedicated fields. Here only mark if these are known to regulate LLPS.

Note: these determinants are the same for all components of the system.

Example:

1) protein density in membrane of LAT
2) valency of LAT
3) valency of GRB2

H) How is the condensate named by the authors?

A free text, taken from the paper, that describes what structures are formed (e.g. liquid droplets, condensates, aggregates, foci etc.). If there are multiple applicable descriptions, separate them with commas.

Note: these are the same for all components of the system.

I) Type of membraneless organelle

Provide the type of subcellular compartment/membraneless organelle associated with this LLPS system. For multiple organelles, separate them by semicolon.

Note: these are the same for all components of the system.

Examples: “stress granule”; “P-granule”; “inclusion body”; “nuclear body”

Note: these will be converted into controlled vocabulary GO terms. Try to use fairly conventional and often used terms if possible.

J) Literature supporting the LLPS

Provide the list of all related PMIDs, where they describe/review the LLPS of the given protein or results directly related to it. For multiple PMIDs, separate them using comma. For each PMID mark separately the type of paper in

brackets. Possible choices: research article or review (this includes actual reviews, comments, perspectives, hypothesis papers, etc.). Please mark in red the PMIDs of the LLPS-related primary research papers from which you could not finish LLPS annotation.
Note: these are the same for all components of the system.
Example: 27056844 (research article), 30951647 (review)

K) Description of the LLPS system

Provide textual description for the given LLPS minimal system with PMIDs indicated in the text. The **functional relevance** and most important features of the LLPS should be discussed.
Note: do not detail the experiments done that support the LLPS or describe its physical features. These information are detailed in fields L and M.
Note2: these descriptions are the same for all components of the system.
Example: "Many cell surface receptors and downstream signaling molecules coalesce into micrometer- or submicrometer-sized clusters upon initiation of signaling. However, the effect of this clustering on signal transduction is poorly understood. T cell receptor (TCR) signaling is a well-studied example of this general phenomenon. In the upstream module, the TCR is phosphorylated by Lck, a membrane-bound protein kinase of the Src family. TCR phosphorylation is opposed by a transmembrane phosphatase, CD45. The phosphorylated cytoplasmic domains of the TCR complex recruit and activate the cytosolic tyrosine kinase ZAP70, which then phosphorylates the transmembrane protein LAT on multiple tyrosine residues. These phosphotyrosines are binding sites for the SH2 domains of adapter protein Grb2 (or Gads), which further interacts with Pro-rich motifs within Sos1 (or SLP-76) through its SH3 domains. LAT and its binding partners coalesce into micrometer- or submicrometer-sized clusters at the plasma membrane upon TCR activation. Dephosphorylation of pLAT by high concentrations of the soluble protein tyrosine phosphatase 1B (PTP1B, 2 μM) caused the clusters to disassemble. Components of the LAT complex activate several downstream modules that mediate calcium mobilization, mitogen-activated protein kinase (MAPK) activation, and actin polymerization. Actin polymerization is initiated from and can reorganize LAT clusters. The experiments suggest that both the phosphorylation state and pY valency of LAT as well as the presence of both SH3 domains in GRB2 are important for cluster formation (PMID: 27056844)."

L) Experimental techniques applied to prove/investigate LLPS

The following is a guide on how to provide a description of the experimental procedure used to address the connection between the given protein and phase separation. Use free text to briefly summarize the main experiments that were done. Use tags and expressions from the list below, where possible. If the exact expression does not fit the text you are writing, feel free to put them in brackets, or to slightly re-phrase them. In many cases, the easiest option is to simply copy sentences and statements (e.g. subsection titles) from the paper. In these cases, again, provide tags from the list below in brackets, where possible.
The following list partitions the experimental aspects into 7 parts. Always define if the experiment was done in vitro or in vivo (aspect 1 - this is mandatory!). Aim to include terms and expressions from all 7 aspects, if possible.

Focus only on the experiments that describe the LLPS. Omit experiments that describe e.g. the cellular role of the phase separation, connection to far away downstream components of signaling pathways, etc.

<u>Aspect 1) Determine the basic type of experiment:</u>
Choose exactly one tag from either in vivo or in vitro. This is mandatory to include.
<u>Aspect 2) Alterations in expression system / cell type used [only for in vivo experiments!]:</u>
If the protein(s) was/were analyzed in their non-native cells (e.g. human proteins produced in E. coli), attach the **most specific** tag that describe this procedure:

- knock-in (the targeted replacement of a wild-type genomic DNA sequence with a different sequence)
- genetic transformation (mutant phenotype resulting from the introduction of exogenous DNA)

<u>Aspect 3) Alteration of protein abundance [only for in vivo experiments!]:</u>
If the abundance of the protein studied was altered in either direction (elevated/decreased), attach all tags that

describe how this was done:

- knock-out (removal of a particular gene (or genes), disrupting the function)
- overexpression
- RNA interference

Aspect 4) Alteration of the gene sequence / protein product:

If the gene sequence encoding the protein was altered in any way, or the protein product was altered in any way, attach all tags that describe how this was done:

- creation of a fusion protein (where the protein is fused with another protein or peptide, such as a fluorescent tag)
- mutation (a change in the DNA encoding the protein, resulting in a single missense mutation, insertion, deletion, frameshift or nonsense mutation at the protein level)
- immunofluorescent tagging of the protein (where the DNA sequence is not changed, but the protein itself is tagged with a fluorescently labeled antibody)
- fluorescent tagging/labeling of the protein by amine-reactive fluorescent dyes

Aspect 5) High level/system level alterations:

If the experiment included any modification that affects multiple protein residues, multiple proteins, or the whole cell, attach all tags that describe how this was done:

- protein phosphorylation (addition of a phospho group to one or more positions)
- protein dephosphorylation (removal of a phospho group from one or more positions)
- protein methylation (addition of a methyl group to one or more positions)
- other PTMs (either the addition or the removal of e.g. ubiquitin, SUMO, etc.)
- perturbation of the cell environment to induce phenotypic changes [only for in vivo experiments!] (modification to the surroundings or conditions in which an organism lives or operates)
- perturbation of in vitro environment [only for in vitro experiments!] (if possible, include more specific terms from the list below)
    - change in pH
    - change in temperature
    - change in redox environment
    - change in protein concentration (either of the LLPS inducing protein, or any other)
    - change in RNA concentration
    - change in DNA concentration
    - change in the concentration of a crowding agent (e.g. polyethylene glycol, ficoll, dextran, or serum albumin)
    - change in the concentration of a small molecule (any other substance that doesn't fit the above)
    - change in salt concentration

Aspect 6) The directly observed effect:

Describe what was the observed effect of the above alterations. This should be a description of how the proteins/cell organelles/condensates/etc behave in response to the experimental procedure. Attach all relevant tags:

- protein localization (where a protein is located inside the cell)
- protein co-localization (where a protein is located compared to another protein or RNA/DNA)
- particle size and count (of the phase separated granule)
- morphology (of the phase separated granule)
- physical interaction (between any two biomolecules or cell constituent)
- other change in phenotype/functional readout (any other feature of the cell or the condensate, any functional readout)
- change in optical properties (e.g. turbidity)

Aspect 7) The experimental technique used to observe the induced change:

Describe the experimental technique used. This is restricted to the actual instrumentation used in the detection, such as various types of microscopy, various biochemical assays, or other biophysical techniques. Does not describe the deduced/calculated parameters, such as viscosity or surface tension values. Attach all relevant tags:

- microscopy (any kind of imaging technique, such as light microscopy, confocal/electron/wide-filed, etc),
- fluorescence detection (if possible, include more specific terms from the list below)
    - FCS (fluorescence correlation spectroscopy)
    - SMFS (Single-Molecule Fluorescence Spectroscopy)

    - FLIP (fluorescence loss in photobleaching)
    - FRAP (fluorescence recovery after photobleaching)
    - FRET (fluorescence/Förster resonance energy transfer)
- immunodetection assay (the use of antibodies to detect biomolecules)
- enzymatic activity assay (measuring the presence/change in enzyme activity)
- protein-protein interaction detection assay (any technique that supplies direct protein-protein interaction information)
- electrophoretic mobility shift assay (solutions of protein and nucleic acid are combined, and the resulting mixtures are electrophoresed under native conditions through polyacrylamide or agarose gel to detect protein-nucleic acid interactions)
- NMR (any kind)
- X-ray crystallography (not small angle)
- SAXS (small angle X-ray scattering)
- SANS (small angle neutron scattering)
- DLS (dynamic light scattering)
- SLS (static light scattering)
- proteomics techniques (e.g. mass spectrometry, MALDI-TOF, SDS, etc.)
- imaging assay evidence [only use if you cannot attach a more specific term!]

Note: these are the same for all components of the system.
*Note2: If there are multiple papers, first describe **all** the relevant experiments done in that paper, specify the PMID after that, and move on to the next paper.*

Example (used tags/expressions from the above list are marked in blue): "Phosphorylated and fluorescently tagged LAT was shown to form liquid-like clusters interacting with GRB2 and SOS1 using microscopy imaging (TIRF) in vitro; and protein dephosphorylation led to the decrease of the particle size and count, marking the disassembly of the condensate. The liquid-like property was evidenced by FRAP. Induced mutation removing the second SH3 domain of GRB2 led to the disassembly of the condensate, demonstrating the importance of valency of the interacting partners. Similarly, stepwise induced mutations of the phosphorylated tyrosines to phenylalanines of LAT correlated with the degree of disruption of the condensate. Functional readout in in vivo studies showed that the clustering of mCitrine-fused LAT is localized to the plasma membrane, and promotes MAPK(ERK) signaling in T cells, thus the in vitro determined effects are biologically relevant in the cellular context. In vitro LAT clusters co-localized with CD45 (a physiological phosphatase of LAT) in artificial membranes, serving as a dephosphorylation assay. The in vivo morphology of liquid droplets were observed and fusion events were followed using microscopy. In vitro LAT clusters enhanced the polymerization of actin, given that the required components are available. PMID: 27056844.**"**

M) The experimental observations supporting the liquid material state of the condensates

List all experimental techniques used, separated with commas. For each technique, provide a PMID in brackets.
Select methods from the list below, use the tags marked in blue:
1 - dynamic movement/reorganization of molecules within the droplet (e.g. FRAP, NMR)
2 - dynamic exchange of molecules with surrounding solvent
3 - morphological traits (e.g. round shape, fusion of droplets, wetting)
4 - rheological traits (material properties e.g. viscosity, surface tension, molecular network mesh size)
5 - sensitivity to 1,6-hexanediol
6 - temperature-dependence
7 - reversibility of formation and dissolution (with changes in environmental conditions)
8 - other (none of the above, but supports the liquid material state)
Note: If you choose 8 (other) please describe the observation.
Note2: these are the same for all components of the system.
Examples: "morphological traits (PMID:12345678), sensitivity to 1,6-hexanediol (PMID: 12345678), reversibility of formation and dissolution (PMID:98765432)"
"dynamic movement/reorganization of molecules within the droplet (PMID:12345678), morphological traits (PMID: 12345678), other: supportive observation…blablabla (PMID: 12345678)"

N) Type of RNA(s) required/used for the LLPS

Describe if RNA is needed for the LLPS of the system, or if it was used in the study of the LLPS. Use any of the following tags:

- cellular RNA
- polyA RNA
- polyU RNA
- mRNA of ... (substitute '...' with the name of the corresponding gene)
- random RNA
- other specific RNA: ... (substitute '...' with the name/definition of the RNA)
- other type of RNA: ... (substitute '...' with the name/definition of the RNA)
- RNA not required

Note: these are the same for all components of the system.
Example: "other specific RNA: NEAT1 architectural lncRNA"

O) PTMs that affect the formation or stability of LLPS

Format: include the position(s) according to UniProt numbering, residue(s), type of PTM(s), the effect the PTM has on LLPS, source publication(s), and the enzyme responsible for the modification, if known.
Example: 403|S|phosphorylation|abolishes|PMID:12345678;PMID:87654321|CK2,CDK2|Notes:....
Mandatory information is marked in colour. **If you want to describe several PTMs, use a separate line for each of them.**
Definitions:

- **position:** either a number, or a range if the position is not well-defined (e.g. 124-200), according to UniProt numbering. If the described PTM is hyperphosphorylation/hypermethylation, always use ranges! Provide UniProt residue number(s), be careful if the article uses different numbering than UniProt due to any reason (e.g. residue numbers provided for sequence without signal peptide/starting methionine..etc) you need to convert it to UniProt residue numbering!
- **residue type:** any of the 20 standard amino acid one letter codes. If a range is defined, define all types of residues that are modified.
- **modification:** any of the following: phosphorylation, hyperphosphorylation, methylation, hypermethylation, ubiquitination, polyubiquitination, SUMOylation, PARylation, acetylation, unknown modification. If there is a well defined modification that is not included in this list, just say so and we can add it.
- **effect:** can be abolishes, weakens, promotes, enables, affects. Only use 'abolishes', if the observed liquid droplets/condensates are dissolved in the modified form! 'Affects' means that the PTM Somehow affects LLPS but it is not stated how or it is stated but it cannot be categorized as any of the available terms. If there is a well defined effect that is not included in this list, just say so and we can add it.
- **source publication(s):** PubMed IDs, separated with ';'.
- **modifying enzyme:** gene name(s), separated by comma. Optional information.
- **notes:** can be any free text, and can contain for example additional information, such as the modification happens in a cell-cycle-dependent or stress-dependent manner, or anything that you find important in relation to the given PTM based on the literature.

Example: "4-60|ST|phosphorylation|promotes|PMID:12345678;PMID:87654321|CDK2|Notes: none"

P) Disease mutations affecting LLPS

Format: define the position and the type of the mutation using the following nomenclature. Use UniProt numbering for defining positions:

- **missense mutations** (SNVs), e.g.: Y234A (a tyrosine mutated to alanine at position 234)

- **deletions**, e.g.: V25_R28del (residues 25-28 are deleted, in total 4 residues, starting with a valine and ending in an arginine). N-terminal truncations shall also be denoted as deletions of the N-terminal region.
- **insertions**, e.g.: V25_L26ins8 (8 residues are inserted between positions 25valine and 26leucine) or if short, the inserted sequence can be added, e.g.:V25_L26insAHG
- **frameshift mutations**, e.g. V25fs (the reading frame is shifted at position 25valine, so in the mutated sequence this is the first possibly modified residue)
- **nonsense mutations (C-terminal truncations)**, e.g. V25* (the valine at position 25 is mutated into a stop codon, this is the first residue that is absent from the mutated protein) [only use if this is a truncation compared to the protein region defined in the entry!!!]

In addition, define dbSNP id, the exact disease name with corresponding abbreviation in brackets, the associated OMIM ID, the effect the mutation has on LLPS, source publication(s).

Example: Y234A|dbSNP:rs1165095258|amyotrophic lateral sclerosis 6 (ALS6)|OMIM:608030|promotes| PMID:12345678|Notes:….

Mandatory information is marked in colour. If you cannot find dbSNP id for the given mutation in UniProt or Ensemble you can indicate “None” instead.

If you want to describe several mutations, use a separate line for each of them.

Help: Looking up the mutation in UniProt will most likely help you with the exact disease names, abbreviations, dbSNP and OMIM IDs.

Effect: can be abolishes, weakens, promotes, enables, affects. Only use 'abolishes', if the observed liquid droplets/condensates are dissolved in the modified form! 'Affects' means that the PTM Somehow affects LLPS but it is not stated how or it is stated but it cannot be categorized as any of the available terms. If there is a well defined effect that is not included in this list, just say so and we can add it.

Notes: Use notes if you can add extra info, for example that the mutation promotes increased rates of fibrilization/ pushes the equilibrium of LLPS towards amyloid-like, more solid assemblies/ abolishes essential interactions between LLPS partners/ mutation is on the dimerization interface and thus disrupts oligomerization …etc. or anything that you find important in relation to the given disease mutation and LLPS based on the literature.

Example:

M404T|dbSNP:rs1165095258|Paget’s disease of bone (PDB)|OMIM:167250|weakens| PMID:29507397| Note:none

G411S|dbSNP:rs1234567890|Paget’s disease of bone (PDB)|OMIM:167250|weakens|PMID:29507397|Note:none

Q) Alternative splicing affecting LLPS

Add “Not known” if there were no experiments done with isoforms. Most entries should get “Not known” in this first stage. If there is any available experimental results on isoforms use the following format:

Isoform P12345-2|abolished/weakened/affected/not affected/promoted/enabled| PMID:12345678;PMID:87654321| Notes:…

Mandatory information is marked in colour. If you want to describe the effects of several isoforms, use a separate line for each of them.

- **isoform:** Add the word Isoform and then the UniProt AC of the given isoform.
- **effect:** Compared to the canonical form the LLPS of the indicated isoform can be abolished, weakened, promoted, enabled, affected, not affected. Only use 'abolished', if the observed liquid droplets/condensates are dissolved when the isoform is investigated! 'Affected' means that the isoform shows somewhat different LLPS than the canonical form but it is not stated how or it is stated but it cannot be categorized as any of the available terms. If there is a well defined effect that is not included in this list, just say so and we can add it.
- **source publication(s)reporting on the LLPS of the isoform:** PubMed IDs, separated with ';'.

If you want to describe several isoforms, use a separate line for each of them.

Example:

Isoform Q9NQI0-2|affected|PMID:25747659

Isoform Q9NQI0-3|affected|PMID:25747659

R) Molecular interaction types contributing to LLPS

Please indicate the tags from this list below, which best describe the given LLPS mechanism according to the authors (in most cases multiple choices should be made). Add the PMID in brackets where they claim this interaction type after the tag (They may describe it in other words but their assessment on the molecular interactions contributing to LLPS should meet at least one or a few of the categories listed below).

1 – multivalent domain-motif interactions (eg. SH3 domains and proline-rich motifs)
2 – multivalent domain-PTM interactions (eg. SH2 domain - pY, UBA domain - ubiquitin, domains recognising histones with specific modifications)
3 – discrete oligomerization (via ordinary oligomerization domains; defined number of monomers/valency)
4 – linear oligomerization/self-association (undefined number of monomers/valency)
5 – coiled-coil formation
6 – helix-helix interaction driven oligomerization (e.g. formation of helical bundles)
7 – prion-like aggregation (typically Q/N rich regions)
8 – formation of amyloid-like/cross-beta/kinked/stacked beta-sheet structures
9 – protein-RNA interaction (often multivalent)
10 – protein-DNA interaction (often multivalent)
11 – simple coacervation of hydrophobic residues
12 – complex coacervation (IDRs with high net charge and large global dimensions form condensed droplets with oppositely charged polymers)
13 – electrostatic (cation-anion) interaction (typically claimed when blocks of opposite charges alternate)
14 – cation-π (cation-pi) interactions
15 – π-π (pi-pi) interactions
16 – dipole-dipole interactions
17 – RNA base pairing/RNA self-assembly
18 – weak electrostatic or hydrophobic interactions between folded domains (like Pab1 RRMs without RNA)
19 – gelation (formation of a system-spanning gel instead of condensed droplets)
20 – not known

Note: these are the same for all components of the system.
Example: "linear oligomerization/self-association (PMID:12345678); prion-like aggregation (PMID:12345678); protein-RNA interaction (PMID:23456789)"

S) Determinants and mechanisms of LLPS formation

Define if possible if the following properties are true or false for the protein/system described:
Membrane cluster
Partner-dependent
RNA-dependent
PTM required for LLPS
Domain-motif interactions involved
Discrete oligomerization involved

T) Functional class of membraneless organelle

Please indicate the tags (in blue) from this list below, which best describe the functions of the given membraneless organelle (MO) based on your perception of the related literature. Provide all tags for each entry that have sufficient support in the literature. Provide the tag describing the primary function (based on your judgement) first and may or may not add a secondary tag.

1 - activation/nucleation/signal amplification/bioreactor
(MOs that activate reactions based on high local concentration of the components)

2 - inactivation/separation/molecular shield

(MOs that inactivate reactions by sequestering some of the required components while keeping others outside)

3 - protective storage/reservoir

(MOs that form to store/protect molecules in an inactive state for a certain period of time, eg. during stress)

4 - biomolecular filter/selectivity barrier

(MOs whose primary function is the selective concentration of certain molecules)

5 - sensor

(MOs which form/dissolve on environmental changes (pH, temperature, stress etc.) to signal them to the cell)

6 - regulator of spatial patterns

(MOs which act as markers of cell polarity, e.g. help asymmetric cell divisions)

7 - memory device

(MOs whose primary function is to act as long-lasting molecular footprints of past external/internal signals)

8 - mechanical property exploitation

(MOs whose primary function is dependent on the mechanical/elastic properties of the condensate itself)

9 - not known/not clear

(if you cannot decide on the function of the given MO based on the information in the related articles)

Note: the above list is based on the functional classification of MOs proposed in the following articles:

PMID:28808090, PMD:28864230, PMID:30826453, PMID:30682370.

Note: these are the same for all components of the system. If for two entries the numbers in column B match, these should match too.

U) Corresponding author contact

Please add the email address(es) of the corresponding author(s) of the main articles belonging to your proteins, so that we can write to the community about the database when your entry is uploaded.

Note: these are the same for all components of the system.