Databases of proteins related to liquid-liquid phase separation

Qian Li[1], Xi Wang[1], Zhihui Dou[1], Weishan Yang[1], Beifang Huang[1], Jizhong Lou[1,2], Zhuqing Zhang[1*]

[1]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China
[2]Key Laboratory of RNA Biology, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China

Correspondence

Zhuqing Zhang, College of Life Sciences, University of Chinese Academy of Sciences, Beijing 101408, China
E-mail: zhuqingzhang@ucas.ac.cn

Acknowledgements

This work was supported by University of Chinese Academy of Sciences and National Natural Science Foundation of China [31870718, 21633001].

Conflict of Interest Statement

The authors declare that they have no competing interests.

Abstract

Liquid-liquid phase separation (LLPS) of biomolecules has been investigated intensively in recent years, which underlies the formation of membrane-less organelles (MLOs) or biomolecular condensates. It contributes to the regulation of various physiological process and related disease development. Rapidly increasing number of studies have recently focused on the biological functions, driving and regulating mechanisms of LLPS in cell. Based on the mounting data generated in the investigations, six databases (LLPSDB, PhaSePro, PhaSepDB, DrLLPS, RNAgranuleDB, HUMAN CELL MAP) have been developed, designed directly based on LLPS studies or the components identification of MLOs. These resources are invaluable for deeper understanding of cellular function of biomolecular phase separation, as well as development of phase separating protein prediction and design. In this review, we summarize the data contents, annotations and organization for each of these databases, highlight their unique features, overlaps and fundamental differences, and further discuss their suitable applications.

Keywords: Liquid-liquid phase separation; Protein; Databases; Membrane-less organelles; Condensates

1 Introduction

Biomolecules within intracellular compartments cooperate spatiotemporally in controlling the efficient and precise biochemical reactions in cell. These compartments can be roughly divided into membrane-bounded organelles and membrane-less ones with distinct structural organizations. Unlike the classic organelles bounded by bilayer lipid membranes, the membrane-less compartments have no membrane bounded, therefore are called as membrane-less organelles (MLOs) or biomolecular condensates, such as cajal body in nucleus, P-body in cytoplasm, nuage in germ cell, receptor cluster, pyrenoid matrix, *etc* [1,2]. It's widely appreciated that the formation of MLOs is driven by liquid–liquid phase separation (LLPS) of biomolecules since the first analysis of liquid droplets in Drosophila embryos [3]. LLPS is a reversible molecular process of condensing certain proteins and/or nucleic acids into dense-phase liquid condensates coexisting with dilute phase [4]. The physicochemical properties of LLPS suggest liquid condensates possess a variety of biological functions as reviewed in Simon's paper [5]. Besides of the physiological roles, LLPS can be regulated by mutations or post-translational modifications (PTMs) of proteins, which might be implicated in a range of incurable neurodegenerative diseases, such as amyotrophic lateral sclerosis (ALS) [6,7], frontotemporal dementia (FTD) [8] and Alzheimer's disease (AD) [9]. These previous studies implied that LLPS provides a new angle for researchers to inspect these diseases and various cellular processes. As a result of increased research interests on LLPS, the publications on LLPS of biomolecules have increased explosively in recent years, as the statistical plot shown in **Figure 1**.

Given many physiological and pathological functions have been discovered to be associated with the formation of MLOs, there is a pressing need to identify the driving mechanism underlying biomolecular LLPS [10]. Many proteins and nucleic acids have been found to be able to undergo LLPS both *in vivo* and *in vitro* [11-15]. Multivalent weak interactions are fundamentally deemed as the main driving force for LLPS [16,17], which are characterized as multi-site dynamic physical cross-linking among biomolecular chains via weak binding forces such as electrostatic, cation–π, π–π and

hydrophobic interactions [18,19]. Multivalent weak interactions can generally occur in proteins between multiple folded domains or between multiple interacting motifs in intrinsically disordered regions (IDRs) or between the both of them [20], as well as between proteins and RNAs/DNAs in some situations [21-23]. No matter how, intrinsically disordered proteins (IDPs) or long IDRs play essential roles in driving LLPS process [24]. They are highly flexible and lack stable 3D structures and harbor repetitive linear motifs or low-complexity regions (LCRs), thus possess great advantages to form transient multivalent weak interactions [25]. The sequence length of IDR as well as the sequence pattern, which can be modified by residues mutation, repeating certain motifs or PTMs, mediate the phase behavior of proteins [2,26]. How the various IDPs or IDRs and their modifications mediate the formation of MLOs and perform their biological functions have attracted researchers much attention recently.

Protein can phase separate on its own or with other molecules, here we call the other molecules as "partners". Partners can be the drivers of phase separation (also are referred as scaffolds), or recruited clients which preferentially participate into condensates formed by scaffolds [25]. Specifically, they could be proteins, RNA/DNA or other molecules such as ATP, *etc*. They collaboratively contribute to the formation or function of condensates [27-29]. Besides, environmental parameters such as the concentrations of protein, nucleic acid and salt, as well as the pH, pressure and temperature of system have been demonstrated to be able to regulate LLPS process [25,26]. In some situations, changes of molecular features or cellular environment may further transform liquid-like condensates into gel- or solid-like states [30-32]. These various influenced factors suggest that the phase behavior of biomolecules can be regulated through multiple aspects for normal cellular process, adaptions and dysfunctions [33].

The intensive investigations in phase separation of biomolecules provide data foundation for more comprehensive and deeper understanding of LLPS in cell biology. Around 40 MLOs have been suggested to be organized via phase separation in eukaryote, bacterial and virus [34], and several studies reviewed the components and functions of the MLOs [35-38]. Timely, a couple of databases covering the function and mechanism as well as experimental information of LLPS related proteins, have been released recently. They together provide researchers a comprehensive overview and undoubtedly serve as evaluable resources. In this review, we describe the data content, organization, annotation focuses, differences and overlap of these databases, and further discuss their applicability to experimental and computational LLPS studies.

2 Databases related to LLPS

Six LLPS-related databases, four containing proteins from direct LLPS studies and two constructed based on MLOs proteome identification are described here. The former includes LLPSDB [39], PhaSePro [40], DrLLPS [34] and PhaSepDB [41], in all or part of the deposited proteins are validated by LLPS experiments. Each database provides the basic information of recorded proteins, as well as their structural and functional annotations. The phase behavior information of proteins is also deposited in each database with more or less details. The latter includes RNAgranuleDB [42] and HUMAN CELL MAP [43], in which the proteome of specific organelles especially MLOs is curated. A general summarization of them is shown in **Table 1**.

2.1 *LLPSDB*

LLPSDB [39] is the first released database designed specifically for proteins undergoing LLPS

which were validated by experiments *in vitro*. Currently, 273 individual proteins and 1175 entries are deposited. It is the only database incorporating both natural and designed proteins. The entry in LLPSDB is defined by specific protein and/or nucleic acid constructs in system. Proteins with different type of modifications or forming condensates with different proteins or RNAs(or DNAs) belong to different entries. For example, both wild-type FUS and its cleaved low complexity region (LCR) can undergo LLPS but they belong to different entries in LLPSDB. The data can be browsed through three different classifications: protein type (natural/designed), main components type (protein(s)/protein(s)+RNA/protein(s)+DNA) or main components number (one/two/more). The detailed functional and structural information of wild type protein or designed one is recoded in protein details page. It can be accessed through browsing by "protein type" or linking back from entry page. The functional description provided in LLPSDB integrates information retrieved from UniProt [44] and literatures. IDRs and LCRs based on related databases or algorithms are visualized in protein details page. Crosslinking to other functional related databases - Uniprot [44], MobiDB [45], DisProt [46], OMIM [47], IDEAL [48], FuzzDB [49] and AmyPro [50] are provided.

A unique feature of LLPSDB is that it includes specific experimental conditions adopted in each LLPS system. The protein sequence, modifications (including cleaved, fusion, motif repeats, mutation and PTM), as well as experimental parameters such as protein and nucleic acid concentration, salt concentration, crowding agent concentration, pH, temperature, pressure etc., are clearly listed in each entry. Furthermore, the database also includes those comparative "negative" situations, where "no" phase separation was detected in the specific experimental condition in corresponding system. Meanwhile, more than 200 phase diagrams in corresponding literatures, which provide the critical phase separation condition of LLPS systems, are also recorded in LLPSDB. Although it is designed specifically for proteins undergoing LLPS in vitro, LLPSDB additionally records whether there are corresponding in vivo (or in cell) experiments in the corresponding literature for each system. However, it does not include proteins with only in vivo experiments, or only identified in MLOs but without detailed experimental conditions of LLPS.

2.2 *PhaSePro*

PhaSePro [40] is a novel database in which proteins verified to drive phase separation in vivo and/or in vitro are manually curated. It contains 121 proteins with 109 from eukaryote, 5 from bacteria and 7 from virus. In each entry, very detailed LLPS annotations of the corresponding protein are carefully summarized manually based on all currently available LLPS studies or existing databases. The protein regions that were demonstrated to drive LLPS, the partners, the molecular interaction types, the determinants of phase separation and droplet property, as well as the annotations on regulation and related disease are listed in each entry page. The functional description and experimental information of LLPS are also provided in the form of free-text, together with the supporting literature references. Besides, some general information of the protein, such as localization, species are also provided. Furthermore, the structural information, such as disordered regions, structural presentations determined by experiments, variants and PTMs are also included in a visualized form through retrieving diverse resources.

Except the detailed LLPS information, such as protein regions driving LLPS and molecular

interaction types mentioned above, another outstanding point of PhaSePro is that it introduces LLPS-specific controlled vocabularies (CVs), which are custom-built based on literatures to annotate the functional, molecular and experimental information of the protein driving LLPS. Four distinct CVs are developed in these aspects (i) 8 classes of the functional roles of membrane-less organelles/granules in the cell; (ii) 19 terms for the different molecular interaction types; (iii) 6 terms to describe the molecular determinants and mechanisms; (iv) 7 terms of experimental observations supporting the liquid state of condensates. Using CVs to standardize the annotations in this database greatly reduces the redundancy of related information, and helps to interpreting each entry.

2.3 *PhaSepDB*

PhaSepDB [41] currently contains 2914 non-redundant proteins localized in more than 30 MLOs. It includes the known 352 LLPS-associated proteins extracted from published literatures, 378 potential proteins reviewed from UniProt according to their subcellular locations, as well as 2516 proteins identified by high throughput experiments including the organelle purification, proximity labeling, immunofluorescence image-based screen and affinity purification. Therefore, in this database, those proteins localizing in specific membrane-less organelle with no direct LLPS investigations are considered as LLPS related. The data can be browsed either through different sources as described above, or through specific membrane-less body locations in a form of graphical navigation on its home page.

For each entry, PhaSepDB provides the information such as species, localization, IDR content, supporting literature, as well as functional description, cell line, and some experimental detail and notes with original sentences from literatures. It's worth noting that PhaSepDB also provides bioinformatic analysis of the sequence properties and displays each of them by an easily interpreted per-residue plot. The analysis integrates results of IDR prediction by ESpritz [51], prion-like sequence prediction by PLAAC [52], electrostatic interaction prediction by Pi–Pi [53] as well as charged/hydrophobic residue distribution analysis by CIDER [54]. It also contains post-translational modifications (PTMs) [55], secondary structure annotations, domain and compositional bias annotations. The molecular properties analysis in PhaSepDB is provided for all human proteins to help the identification of potential LLPS proteins.

2.4 *DrLLPS*

DrLLPS [34] is a gene-centered database and currently holds the largest amount of data. It totally contains 437887 proteins in 164 eukaryotes including 150 scaffold proteins, 987 regulators and 8148 potential client proteins manually curated from published literatures, and their orthologs considered as potential LLPS-associated proteins identified via a genome-wide detection by protein sequence blast. The scaffolds are defined as the drivers of LLPS; the regulators refer to proteins that have not been identified to undergo LLPS but known to be involved in regulating the stability and formation of MLOs and/or liquid droplets; the clients here mean those proteins co-complexed or co-localized with scaffolds, but not known whether they are indispensable for the formation of condensates. Data can be accessed through three categories: 40 biomolecular condensates belonged to five super-classes including *in vitro* droplet, nucleus, cytoplasm, germ cell and others;

LLPS types – scaffolds, regulators and clients; species which mainly include proteome sets of 68

animals, 50 plants and 46 fungi.

The annotations for each protein in DrLLPS include basic information such as Ensembl [56] /UniProt [44] /GeneBank [57] /RefSeq [58] accession numbers, functional description and protein/nucleotide sequences etc. DrLLPS also presents brief descriptions on protein roles in LLPS, localizations, effects of partners, experimental analysis description in vitro and/or in cell, as well as primary supporting references. Besides, It provides very comprehensive molecular feature annotations from 110 widely-used public resources for 28,024 known and potential LLPS-associated proteins in eight model species, which covering 16 aspects, including intrinsically disordered regions (IDRs), domain annotations, PTMs, genetic variations, cancer mutations, protein 3D structures, subcellular localizations, etc. Although most of the information is computationally predicted and has not been detected in LLPS experimental studies, it brings researchers substantial useful information and will assist the further related investigations.

2.5 *RNAgranuleDB and HUMAN CELL MAP*

RNAgranuleDB [42] and HUMAN CELL MAP [43] are two databases not directly related to LLPS but particularly focus on the proteome of organelles, especially of MLOs. RNAgranuleDB provides a comprehensive summary of SG (stress granule) and PB (P-body) components, and in total 4385 mammalian proteins (from human, mouse, and rat) are collected. All these proteins are manually curated from 122 peer-reviewed publications and identified by either high-throughput experiments or low-throughput approaches. They are categorized into 4 tiers weighted according to the degree of experimental supporting for the residence in SGs or PBs. Among them, proteins in tier 1 have the highest confidence to be considered as SG-PB proteins. In addition, RNAgranuleDB analyzed the potential LLPS capability of these proteins by six first-generation predictors based on sequence features associated with aggregation or phase-separation properties reviewed in ref [59]. It was expectedly found that proteins in the higher tier groups contain a larger fraction of proteins showing significant LLPS matches for all of the sequence with the predictions.

HUMAN CELL MAP is another database curating proteins not only in MLOs but also in membrane-bound organelles. It provides an extensive BioID-based proximity map of HEK293 cell, comprising 192 markers and 4145 high confidence prey proteins. In addition to some general information, it also summarizes the enrichment of expected domains and motifs as well as the GO-terms for each organelle. Although HUMAN CELL MAP is not specifically targeted at phase separation, many LLPS-related proteins, MLOs and their relationships can be found here, which will undoubtedly expand our understanding of LLPS in cells.

3 Comparison of the databases

These six databases provide valuable information on LLPS system and their components. They have overlaps, meanwhile each database is designed for specific aims and has unique features, as **Table 2** and **Figure 2** presented. We compare the databases from the following aspects: data groups and sources, annotations, suitable applications.

Data collected in these databases are overlapped but different in a certain extent, which can be mainly grouped into two classes: one for proteins undergoing or involving in LLPS which have been

validated directly by in vivo and/or in vitro experiments; the other for proteins identified or predicted to be components of known MLOs or biomolecular condensates. Currently, more than one hundred proteins have been verified to undergo or involve in LLPS directly. Four resources – LLPSDB, PhaSePro, PhaSepDB and DrLLPS – collect them. All proteins in LLPSDB have been verified to undergo (or NOT undergo) LLPS in vitro on their own or with other proteins or nucleic acids. PhaSePro focuses on proteins driving LLPS with explicit in vitro and/or in vivo experimental evidences. Except the first data group, PhaSepDB and DrLLPS also incorporate the second class of data. PhaSepDB includes the proteins localized in membrane-less compartments which are recorded in UniProt or identified by high throughput experiments. In DrLLPS, proteins in various biomolecular condensates with experimentally identification are collected and classified. Besides, based on genome-wide detection via protein sequence blast, the orthologs of both data groups in 164 eukaryotes are also deposited. RNAgranuleDB provides the current available compositions of SG and PB proteomes, and HUMAN CELL MAP is curated for protein components in both membrane-bound and membrane-less compartments from HEK293 cell. The data in the latter two databases are experimentally validated.

Although all the databases provide general information of deposited proteins, such as protein name, species, localization, function, PMID, short description from literatures etc., the annotations of experimental details, as well as molecular properties analysis in each of them are various and have their own emphases. LLPSDB provides in-depth annotations describing the validated phase behavior of the system in each entry, with exhaustive molecular modifications such as cleaving, mutation, PTMs, etc. for specific protein constructs and corresponding explicit phase separation conditions as well as phase diagrams. For sequence properties, it includes IDR and LCR predictions for each wild type protein. PhaSePro contains a broader array of functional and disease information of LLPS. It also provides the LLPS driving regions, molecular interaction types, as well as detailed LLPS experimental information in text form. The proposed LLPS-specific CVs are applied to standardize the descriptions. Structure-related annotations in PhaSePro are more abundant, including not only the predicted IDRs but also the PTMs, sequence variants and 3D structures in visualization. PhaSepDB specifically provides useful sequence analysis such as PTMs, secondary structure distribution, electrostatic interaction and hydrophobic residue distribution, displaying each by an easily interpreted per-residue plot. DrLLPS includes the most comprehensive structure-related annotations. It integrates 110 widely-used public resources to describe the protein structural and functional features from 16 aspects, with each aspect summarized by no less than two kinds of resources. For RNAgranuleDB and HUMAN CELL MAP, they both focus on the proteome of organelles, therefore their annotations lack detailed information of LLPS, but include more evidences in experimental identification.

These databases are complimentary and they together provide valuable and more comprehensive resources to facilitate research of biomolecular phase separation and cellular organization, not only in experimental aspect, but also in the development of theory and prediction algorithm. The proteins deposited in LLPSDB and PhaSePro are all verified by LLPS experiments, which constitute high-quality training set for the development of new methods to identify novel LLPS proteins. In LLPSDB, specific protein constructs with corresponding specific experimental conditions for LLPS will further help researchers to understand how the phase behavior of protein is sensitive to environment, and to design algorithms for predicting the phase separation propensity of new proteins. Recently, a predictor of LLPS protein (PSPredictor, http://www.pkumdl.cn/PSPredictor) based on

machine learning was developed [60], using the datasets in LLPSDB as training set. It achieves a fairly high prediction accuracy, and outperforms other reported prediction tools so far, which are all based on specific protein sequence features [53,59,61]. The well summarized structural, functional, as well as detailed experimental information provided in PhaSePro makes it very useful for researchers to find complete and systematic knowledge of the protein. PhaSepDB and DrLLPS include more proteins related to LLPS which were validated or likely localize in MOLs or biomolecular condensates. The extensive molecular property analysis within them could provide helpful information to understand if they might be potential proteins to undergo or regulate LLPS in future investigation. The large amount of orthologs and their annotations given in DrLLPS make it specifically useful in analyzing LLPS from evolutionary perspective. Taken together, a suitably combined application of these databases would definitely make a great advance in deeper understanding of LLPS in cell.

4 Summary

Investigations in LLPS of biomolecules or the formation of biomolecular condensates have grown fast in recent years. A number of databased have been constructed timely to curate the generated mounting data, which undoubtedly will make advances in the phase separation research of biomolecules. In this review, we summarized six recent released databases of proteins related to LLPS - LLPSDB, PhaSePro, PhaSepDB, DrLLPS, RNAgranuleDB, HUMAN CELL MAP. Although the data in them are overlapped in certain extent, the organization and annotations in each of them have their own focuses and unique features. We believed the careful summarization and careful comparison of these databases in this review will provide researchers a general perception and be help for utilizing these resources efficiently.

References

1. Uversky VN. Protein intrinsic disorder-based liquid-liquid phase transitions in biological systems: Complex coacervates and membrane-less organelles. *Adv Colloid Interface Sci.* 2017;239:97-114.
2. Mitrea DM, Kriwacki RW. Phase separation in biology; functional organization of a higher order. *Cell Commun Signal.* 2016;14:1-20.
3. Brangwynne CP, Eckmann CR, Courson DS, et al. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science.* 2009;324:1729-1732.
4. Dolgin E. What lava lamps and vinaigrette can teach us about cell biology. *Nature.* 2018;555:300-302.
5. Alberti S, Gladfelter A, Mittag T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell.* 2019;176:419-434.
6. Gui X, Luo F, Li Y, et al. Structural basis for reversible amyloids of hnRNPA1 elucidates their role in stress granule assembly. *Nat Commun.* 2019;10:2006-2017.
7. Patel A, Lee HO, Jawerth L, et al. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell.* 2015;162:1066-1077.
8. Mann JR, Gleixner AM, Mauna JC, et al. RNA Binding Antagonizes Neurotoxic Phase Transitions of TDP-43. *Neuron.* 2019;102:321-338.e328.
9. Kostylev MA, Tuttle MD, Lee S, et al. Liquid and Hydrogel Phases of PrP(C) Linked to Conformation Shifts and Triggered by Alzheimer's Amyloid-beta Oligomers. *Mol Cell.* 2018;72:426-443.e412.
10. Zhang CS, Lai L. Physiochemical Mechanisms of Biomolecular Liquid-Liquid Phase Separation. *Acta Physico-Chimica Sinica.* 2020;36:1907053-1907050.
11. Schuster BS, Reed EH, Parthasarathy R, et al. Controllable protein phase separation and modular recruitment to form responsive membraneless organelles. *Nat Commun.* 2018;9:2985.
12. Elbaum-Garfinkle S, Kim Y, Szczepaniak K, et al. The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc Natl Acad Sci U S A.* 2015;112:7189-7194.
13. Nott TJ, Petsalaki E, Farber P, et al. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol Cell.* 2015;57:936-947.
14. Monahan Z, Ryan VH, Janke AM, et al. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *EMBO J.* 2017;36:2951-2967.
15. McGurk L, Gomes E, Guo L, et al. Poly(ADP-Ribose) Prevents Pathological Phase Separation of TDP-43 by Promoting Liquid Demixing and Stress Granule Localization. *Mol Cell.* 2018;71:703-717.e709.
16. Fung HYJ, Birol M, Rhoades E. IDPs in macromolecular complexes: the roles of multivalent interactions in diverse assemblies. *Curr Opin Struct Biol.* 2018;49:36-43.
17. Shin Y, Brangwynne CP. Liquid phase condensation in cell physiology and disease. *Science.* 2017;357.
18. Lin YH, Forman-Kay JD, Chan HS. Theories for Sequence-Dependent Phase Behaviors of Biomolecular Condensates. *Biochemistry.* 2018;57:2499-2508.
19. Martin EW, Mittag T. Relationship of Sequence and Phase Separation in Protein Low-Complexity Regions. *Biochemistry.* 2018;57:2478-2487.
20. Harmon TS, Holehouse AS, Pappu RV. Differential solvation of intrinsically disordered linkers

drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins. *New Journal of Physics.* 2018;20:045002.

21. Zhou H, Song Z, Zhong S, et al. Mechanism of DNA-Induced Phase Separation for Transcriptional Repressor VRN1. *Angew Chem Int Ed Engl.* 2019;58:4858-4862.
22. Du M, Chen ZJ. DNA-induced liquid phase condensation of cGAS activates innate immune signaling. *Science.* 2018;361:704-709.
23. Drino A, Schaefer MR. RNAs, Phase Separation, and Membrane-Less Organelles: Are Post-Transcriptional Modifications Modulating Organelle Dynamics? *Bioessays.* 2018;40:e1800085.
24. Darling AL, Zaslavsky BY, Uversky VN. Intrinsic Disorder-Based Emergence in Cellular Biology: Physiological and Pathological Liquid-Liquid Phase Transitions in Cells. *Polymers.* 2019;11:990-1012.
25. Posey AE, Holehouse AS, Pappu RV. Phase Separation of Intrinsically Disordered Proteins. *Methods Enzymol.* 2018;611:1-30.
26. Ruff KM, Roberts S, Chilkoti A, Pappu RV. Advances in Understanding Stimulus-Responsive Phase Behavior of Intrinsically Disordered Protein Polymers. *J Mol Biol.* 2018;430:4619-4635.
27. Uversky VN, Kuznetsova IM, Turoverov KK, Zaslavsky B. Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates. *FEBS Lett.* 2015;589:15-22.
28. Banani SF, Rice AM, Peeples WB, et al. Compositional Control of Phase-Separated Cellular Bodies. *Cell.* 2016;166:651-663.
29. Ghosh A, Mazarakos K, Zhou HX. Three archetypical classes of macromolecular regulators of protein liquid-liquid phase separation. *Proc Natl Acad Sci U S A.* 2019;116:19474-19483.
30. Murakami T, Qamar S, Lin JQ, et al. ALS/FTD Mutation-Induced Phase Transition of FUS Liquid Droplets and Reversible Hydrogels into Irreversible Hydrogels Impairs RNP Granule Function. *Neuron.* 2015;88:678-690.
31. Dao TP, Martyniak B, Canning AJ, et al. ALS-Linked Mutations Affect UBQLN2 Oligomerization and Phase Separation in a Position- and Amino Acid-Dependent Manner. *Structure.* 2019;27:937-951.e935.
32. Ryan VH, Dignon GL, Zerze GH, et al. Mechanistic View of hnRNPA2 Low-Complexity Domain Structure, Interactions, and Phase Separation Altered by Mutation and Arginine Methylation. *Mol Cell.* 2018;69:465-479.e467.
33. Cinar H, Fetahaj Z, Cinar S, Vernon RM, Chan HS, Winter RHA. Temperature, Hydrostatic Pressure, and Osmolyte Effects on Liquid-Liquid Phase Separation in Protein Condensates: Physical Chemistry and Biological Implications. *Chemistry (Easton).* 2019;25:13049-13069.
34. Ning W, Guo Y, Lin S, et al. DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic Acids Res.* 2020;48:D288-D295.
35. Uversky VN. Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. *Curr Opin Struct Biol.* 2017;44:18-30.
36. Lee KH, Zhang P, Kim HJ, et al. C9orf72 Dipeptide Repeats Impair the Assembly, Dynamics, and Function of Membrane-Less Organelles. *Cell.* 2016;167:774-788.e717.
37. Youn JY, Dunham WH, Hong SJ, et al. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol Cell.* 2018;69:517-532.e511.
38. Jain S, Wheeler JR, Walters RW, Agrawal A, Barsic A, Parker R. ATPase-Modulated Stress

Granules Contain a Diverse Proteome and Substructure. *Cell.* 2016;164:487-498.
39. Li Q, Peng X, Li Y, et al. LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro. *Nucleic Acids Res.* 2020;48:D320-D327.
40. Meszaros B, Erdos G, Szabo B, et al. PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* 2020;48:D360-D367.
41. You K, Huang Q, Yu C, et al. PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res.* 2020;48:D354-D359.
42. Youn JY, Dyakov BJA, Zhang J, et al. Properties of Stress Granule and P-Body Proteomes. *Mol Cell.* 2019;76:286-294.
43. Go CD, Knight JDR, Rajasekharan A, et al. A proximity biotinylation map of a human cell. *bioRxiv.* 2019:796391.
44. Bateman A, Martin MJ, Orchard S, et al. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019;47:D506-D515.
45. Piovesan D, Tabaro F, Paladin L, et al. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.* 2018;46:D471-D476.
46. Piovesan D, Tabaro F, Micetic I, et al. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* 2017;45:D219-D227.
47. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* 2019;47:D1038-D1043.
48. Fukuchi S, Amemiya T, Sakamoto S, et al. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.* 2014;42:D320-D325.
49. Miskei M, Antal C, Fuxreiter M. FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Res.* 2017;45:D228-D235.
50. Varadi M, De Baets G, Vranken WF, Tompa P, Pancsa R. AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res.* 2018;46:D387-D392.
51. Walsh I, Martin AJ, Di Domenico T, Tosatto SC. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics.* 2012;28:503-509.
52. Lancaster AK, Nutter-Upham A, Lindquist S, King OD. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics.* 2014;30:2501-2502.
53. Vernon RM, Chong PA, Tsang B, et al. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife.* 2018;7:e31486.
54. Holehouse AS, Das RK, Ahad JN, Richardson MO, Pappu RV. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J.* 2017;112:16-21.
55. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* 2015;43:D512-D520.
56. Cunningham F, Achuthan P, Akanni W, et al. Ensembl 2019. *Nucleic Acids Res.* 2019;47:D745-D751.
57. Takeya M, Yamasaki F, Uzuhashi S, et al. NIASGBdb: NIAS Genebank databases for genetic resources and plant disease information. *Nucleic Acids Res.* 2011;39:D1108-D1113.
58. Pruitt KD, Brown GR, Hiatt SM, et al. RefSeq: an update on mammalian reference sequences.

Nucleic Acids Res. 2014;42:D756-D763.

59. Vernon RM, Forman-Kay JD. First-generation predictors of biological protein phase separation. *Curr Opin Struct Biol.* 2019;58:88-96.
60. Sun T, Li Q, Xu Y, Zhang Z, Lai L, Pei J. Prediction of liquid-liquid phase separation proteins using machine learning. *bioRxiv.* 2019:842336.
61. Orlando G, Raimondi D, Tabaro F, Codice F, Moreau Y, Vranken WF. Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates. *Bioinformatics.* 2019;35:4617-4623.

Tables

Table 1 Overview of six databases in the main text.

Databases	Organization	Data Contents	Data sources	Outstanding features of annotation	Availability	Ref.
LLPSDB	Entries are defined by specific protein and/or nucleic acid constructs. Classified by: ⅰ) protein type (Natural, Designed) ⅱ) components type (Protein(s), Proteins(s)+RNA, Protein(s)+DNA) ⅲ) components number (One, Two, More)	273 proteins 1175 entries	Validated by LLPS experiments *in vitro*	• Including natural and designed proteins • Provides exhaustive molecular modifications including fusion, cleaved, mutation, repeat, and PTMs that detected experimentally for specific protein constructs • Provides explicit phase separation conditions (environmental parameters) and more than 200 phase diagrams	http://bio-comp.ucas.ac.cn/llpsdb or http://bio-comp.org.cn/llpsdb	39
PhaSePro	Entries are defined by specific proteins.	121 proteins (109 from eukaryote, 5 from bacteria and 7 from virus)	Validated by LLPS experiments *in vitro* and/or *in vivo*	• Provides each protein the validated LLPS driver region(s) and molecular interaction types contributing to LLPS • Introduces LLPS-specific CVs to annotate the functional, molecular and experimental information of the protein driving LLPS • Provides a broader array of structural, functional and disease information	https://phasepro.elte.hu	40

PhaSepDB	Entries are defined by specific proteins. Classified by: ⅰ) data sources (Reviewed, UniProt Reviewed, High Throughput) ⅱ) Location and Organelle (more than 30 MLOs)	2914 proteins	• Validated by LLPS experiments • Localized in membrane-less compartments through UniPort review and high throughput	• Proteins can be browsed through specific membrane-less body locations in a form of graphical navigation on its home page • Provides bioinformatic analysis of the sequence properties such as PTMs, secondary structure distribution, the electrostatic interaction and hydrophobic residue distribution, and displays the result of each by an easily interpreted per-residue plot • Provides sequence analysis of other human proteins	http://db.phasep.pro/	41
DrLLPS	Entries are defined by specific genes. Classified by: ⅰ) condensates (*In vitro* droplet, Nucleus, Cytoplasm, Germ cell, Others) ⅱ) LLPS types (Scaffold, Regulator, Client) ⅲ) Species (Animals, Plants, Fungi)	437,887 proteins in 164 eukaryotes	• Validated by experiments of LLPS or membrane-less compartments identification • Identified computationally via the protein sequence blast	• Holds the largest amount of data • Includes the most comprehensive structure-related annotations from 110 public resources covering 16 aspects	http://llps.biocuckoo.cn/	34
RNAgranuleDB	Entries are defined by specific proteins. Classified by:	4385 proteins	Localized in stress granule and P body validated by	• All proteins are categorized into 4 tiers weighted according to the degree of support it provides for protein	http://rnagranuledb.lunenfeld.ca	42

	i) experiment design (discovery based approach, candidate based approach) ii) evidence type (cell biological, physical, genetic) iii) specific assay or dataset		experiments	residence in SGs or PBs. • Proteins are analysised by the prediction of six first-generation LLPS predictors • Lacks detailed information of LLPS		
HUMAN CELL MAP	Entries are defined by specific genes. Classified by organelles (Membrane-bound, Membrane-less)	4145 proteins	Localized in membrane-bound or membrane-less compartments validated by experiments	• Summarizes each compartment the enrichment of expected domains and motifs as well as GO-terms • Provides channels to analyze the spatiotemporal correlations between proteins in different organelles • Lacks detailed information of LLPS	https://cell-map.org/ or https://humancellmap.org/	[43]

Abbreviations: LLPS, liquid-liquid phase separation; PTMs, post-translational modifications; CVs, controlled vocabularies; MLOs, membrane-less organelles; SGs, stress granules; PBs, P-bodys.

Table 2 Overlaps between six databases of proteins related to LLPS.

	LLPSDB	PhaSePro	PhaSepDB	DrLLPS	RNAgranuleDB	HumanCellMap
LLPSDB	*273*					
PhaSePro	65	*121*				
PhaSepDB	94	82	*2957*			
DrLLPS	115	83	1520	*9281*		
RNAgranuleDB	25	16	324	359	*380*	
HumanCellMap	45	35	1056	1825	242	*4424*

The numbers of overlapped proteins between any two databases were obtained though "UniProt ID" except RNAgranuleDB. For the overlapped proteins between RNAgranuleDB and other databases, "gene name" was used for comparison.

Figure Legends

Figure 1 Number of publications about biomolecular LLPS increased in recent years. The Web of Science with default options by a keyword combination ‘protein* AND liquid* AND (phase separation* OR phase transiton* OR demixing* OR assembly* OR condensate* OR condensation* OR coacervate* OR segregate* OR segregation*) NOT (liquid chromatography)’ was used in retrieving publication records.

Figure 2 Schematic of relationship and unique features of the six databases related to LLPS.

Figures

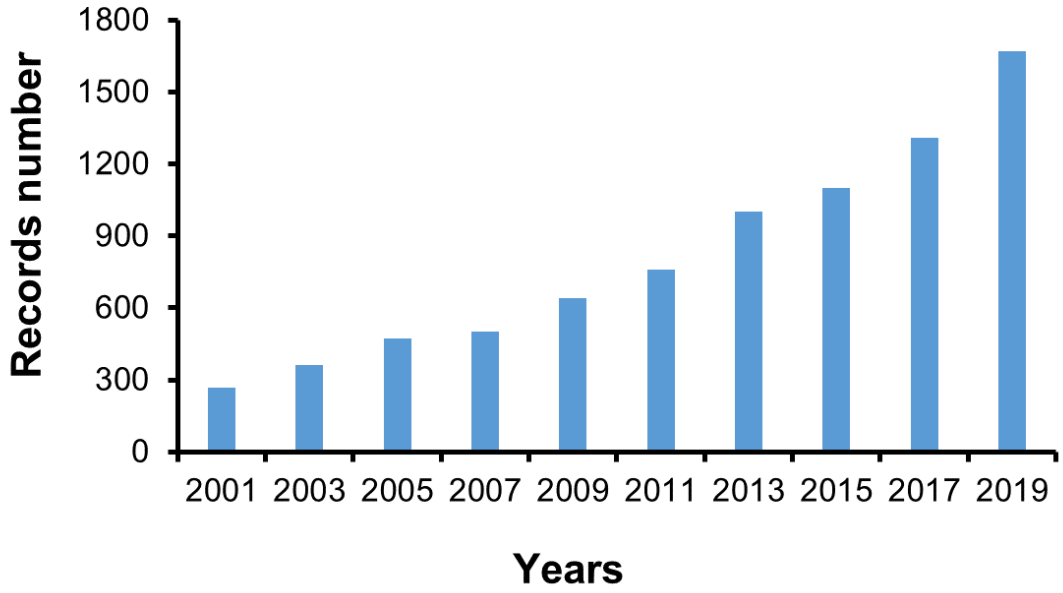
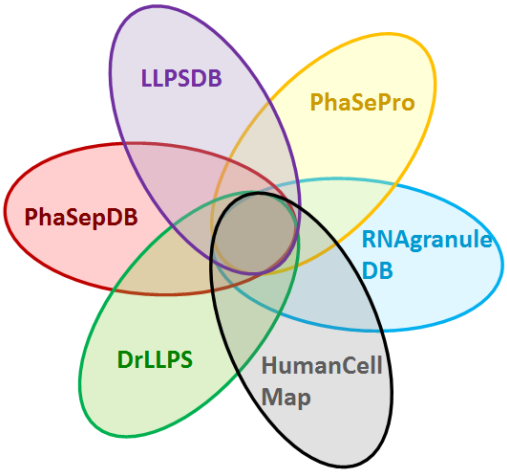


Figure 1 Number of publications about biomolecular LLPS increased in recent years. The Web of Science with default options by a keyword combination 'protein* AND liquid* AND (phase separation* OR phase transiton* OR demixing* OR assembly* OR condensate* OR condensation* OR coacervate* OR segregate* OR segregation*) NOT (liquid chromatography)' was used in retrieving publication records.



Unique Features:

LLPSDB – Includes designed proteins; validated by LLPS experiments *in vitro*; provides explicit phase separation conditions.

PhaSePro – Validated by LLPS experiments *in vitro/in vivo*; provides driving mechanism and extensive functional inforamtion.

PhaSepDB – Includes proteins localized in MLOs; provides bioinformatic analysis of the sequence properties such as secondary structure distribution, PTMs sites, etc

DrLLPS – Includes proteins identified in MLOs and orthologs obtained via the proteome-wide blast; Provides the most comprehensive structure-related annotations.

RNAgranuleDB – Focus on proteins localized in stress granule and P-body.

HumanCellMap – Provides a location map for intracellular proteins, both in membrane-bounded and membrane-less organelles.

Figure 2 Schematic of relationship and unique features of the six databases related to LLPS.