Project FK-128133 **Lead researcher: Rita Pancsa**

The interplay of alternative splicing and the assembly of membraneless organelles

Final report:

Creating the PhaSePro database:

The performed a comprehensive literature screening in 2019 for proteins known to drive liquid-liquid phase separation (LLPS) and their thorough annotation and classification was performed. After assembly of the database data, we created the user interface in collaboration with Zsuzsanna Dosztanyi's group at ELTE. The PhaSePro database of LLPS driver proteins and the accompanying manuscript was published in Nucleic Acids Research 2020 Database issue (with this we achieved Aim 1 of the project):

Please see the database here: https://phasepro.elte.hu/
…and the accompanying publication here: https://doi.org/10.1093/nar/gkz848.

Review on the emergent functions of phase-separated liquid condensates:

We have published a review article in which I wrote a larger section that introduces the functional categorization of phase-separated liquid condensates.

See review: https://www.sciencedirect.com/science/article/pii/S1570963919300433?via%3Dihub

Integration of LLPS-related data into the core data resources of our research field: DisProt and ELM:

1) DisProt: I have annotated a subset of the identified LLPS driver proteins into the DisProt database, the major resource for experimentally validated intrinsically disordered proteins (IDPs) and protein regions (IDRs). LLPS drivers have extended disordered regions and DisProt now allows assigning LLPS and the formation of liquid condensates as function to these disordered regions. These annotations are available in the last two published DisProt releases.

Please see the 2020 DisProt publication here: https://doi.org/10.1093/nar/gkz975
…and the 2022 DisProt publication here: https://doi.org/10.1093/nar/gkab1082

Commitment to annotating new disordered regions is exceptionally important because prediction methods that can address the level and locations of protein disorder in whole proteomes generally rely on the manually curated regions (serving as training sets). The new community paper published on the quality assessment of new disorder prediction methods is available here. Our group members and myself are listed among DisProt Curators who contributed to the article:

https://www.nature.com/articles/s41592-021-01117-3

2) ELM: I have also annotated a subset of the identified LLPS driver proteins into the ELM database, the central resource of eukaryotic linear motifs (ELMs). LLSP drivers often form condensates through domain-motif or domain-PTM interactions mediated by ELMs. These are already available in the recently published two ELM releases, of which the 2022 database article contains a complete large section on linear motifs in LLPS that was written by me (I am shared corresponding author of the article).

Please see the 2020 ELM publication here: https://doi.org/10.1093/nar/gkz1030
…and the 2022 ELM publication here: https://doi.org/10.1093/nar/gkab975

Review on the modes of regulation of phase-separated biomolecular condensates:

We have published a comprehensive review on the regulation of phase-separated biological condensates. In this review we have dedicated a whole larger section to regulation by alternative splicing, wherein we have comprehensively reviewed the already known cases where alternative splicing plays a role in the regulation of LLPS.

Please see the new publication here: https://febs.onlinelibrary.wiley.com/doi/full/10.1111/febs.15254

I created Figure 1 of the article which was later on selected for providing the basis of the cover of that special issue of The FEBS Journal: https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.14890

Yet unpublished work performed in the first two years related to the project:

In the meantime, the physiological AS variations were collected for the human LLPS proteins of PhaSePro and checked for tissue-specificity (WP3). Tamás Horváth performed the CIDER and DynaMine predictions and comparisons for the wild type and physiological AS variant LLPS protein regions to estimate differences in conformations and dynamics (WP4). Based on the results we have decided which cases to take for experiments (cases selected: HnRNPA1, DDX4, RBFOX1) (WP5). After we have selected the wild type- physiological AS pairs, we have obtained part of the desired constructs, produced the other ones by mutagenesis and cloned them into expression vectors. We tried to optimize their expression, but in many cases, we unfortunately failed. For some of them protein purification could be achieved (WP6).

Then, Erzsébet Fichó has obtained the cancer-associated splice variants of the human LLPS proteins from our previously published comprehensive dataset of mutation-mediated alternative splicing perturbations in 33 different cancer types. The ones which likely perturb oligomerization, protein-protein or protein-RNA interactions important for LLPS have been identified based on dedicated databases (Aim 3 achieved). Besides their impact on the interactions of the respective proteins, the conformational impacts of these splicing perturbations were also estimated (WP 4), this was intended to help us decide which cancer-associated cases to choose for experimental investigation (WP 5). Unfortunately, we could not identify enough promising cases, and there were hardly any cases where we have seen a chance for successful experimental investigation. Still, we could select a cancer-associated splice variant of the GATA3 transcription factor that is a known LLPS driver (WP 4 and 5).

Due to the COVID pandemic starting in 2020 the ordering of materials became very difficult and lengthy in our research institute, also, people could only work in the lab with time restrictions. Furthermore, the person whom we have hired for the project (Marcell Váradi) has left the lab after few months of work (decided to work for a company) and, partly due to the COVID pandemic, we could not find another person who could be fully dedicated to this project and perform the remaining experiments. Therefore, the experimental steps could not be carried out with the efficiency we have originally planned.

Review introducing the available computational resources for studying LLPS proteins:

In collaboration with Dr. Bálint Mészáros from EMBL Heidelberg and Prof. Wim Vranken from VIB/VUB Brussels we have wrote a review article on the available computational resources for identifying and describing proteins driving liquid-liquid phase separation. In this user guide/review article we comprehensively introduce our LLPS resource (PhaSePro) and other LLPS databases, and the recently published associated prediction methods. The review article was published in the prestigious Briefings in Bioinformatics journal (published by Oxford Academic), see here:
https://academic.oup.com/bib/article/22/5/bbaa408/6124912

Collaboration on describing the LLPS behaviour of E. coli SSB and the two human SSB proteins:

In collaboration with the Motor Enzymology Research Group of Dr Mihály Kovács at ELTE, we have published an interesting case of phase separation, that of the bacterial single-stranded DNA-binding protein (SSB). Please see the new publication here: https://www.pnas.org/content/117/42/26206.long

Although this bacterial LLPS protein is not affected by splicing and therefore its LLPS is not regulated by alternative splicing, we currently investigate the phase separation behaviour of its human orthologs. Both human orthologs (hSSB1 and hSSB2) show LLPS (unpublished results) and they are subjects of alternative splicing. Moreover, both the region responsible for oligomerization and the LLPS-prone disordered regions differ in the alternatively spliced isoforms, so the two proteins are promising candidates for alternative splicing-mediated LLPS regulation. We already see that the two human SSBs show different redox state-dependent LLPS propensities. While hSSB1 only undergoes LLPS in the oxidized state, hSSB2 also undergoes LLPS in the presence of reducing agents and forms amorphous aggregates in the oxidized state. These results will be published soon.

Cellular concentrations and dosage sensitivity of LLPS driver proteins:

In collaboration with our sister group (Peter Tompa's other lab in Brussels) we have studied the concentration conditions of the hitherto experimentally investigated LLPS driver proteins. We described multiple reasons why the relatively high protein concentrations applied in in vitro LLPS experiments could indeed represent physiologically relevant concentration conditions even if the protein concentrations measured by quantitative proteomics are considerably lower (when protein counts are averaged out to whole cell volumes). We also showed a statistically significant relationship between gene dosage sensitivity and the ability of the resulting proteins to drive LLPS. The described aspects and datasets are important for making a clear distinction between real physiological LLPS drivers and those proteins that were artificially induced to undergo LLPS by applying them in very high concentrations in the respective LLPS trial experiments.

This analysis is particularly relevant for the proposed project because the different alternatively spliced isoforms of proteins are usually available in varying copy numbers/concentrations in cells. Therefore, it is not only differences in their sequence, but also differences in their availability that could determine their LLPS behaviours. It is also noteworthy that isoforms could undergo LLPS completely independently, for instance in different subcellular locations if they have different localization patterns. However, their LLPS could be additive as well (if they can co-phase separate) or one could negatively affect the LLPS of the other. Therefore, not only their absolute, but also their relative availabilities could influence their LLPS behaviour. See publication here: https://www.mdpi.com/1422-0067/22/6/3017/htm

Method development for the study of aggregation-prone LLPS driver proteins:

In collaboration with our sister group (Peter Tompa's other lab in Brussels) we recently demonstrated that at carefully selected pH values proteins such as the low-complexity domain of hnRNPA2, TDP-43, and NUP98, or the stress protein ERD14, can be kept in solution (avoiding their aggregation) and their LLPS can then be induced by a jump to native pH. This approach represents a generic method for studying the kinetics of LLPS under near native conditions that can be easily controlled, providing a platform for the characterization of physiologically relevant LLPS behaviour of diverse proteins.

This method is important for our project because we are trying to apply the described principles when analysing the LLPS behaviours of our purified (wild type or alternatively spliced) protein isoforms.

See the publication here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7815728/

Charged regions in LLPS-associated proteins:

In collaboration with the group of Dr. Zoltán Gáspári at the Pázmány Péter Catholic University we have demonstrated that proteins harbouring regions with specific charged residue patterns are significantly associated with LLPS. In particular, regions with repetitive arrays of alternating charges show the strongest association, whereas segments with generally high charge density and single α-helices show detectable but weaker connections (see publication: https://doi.org/10.1002/1873-3468.14294). In this analysis we used a relatively large dataset of LLPS-associated proteins, to ensure that our statistics are strong and meaningful. However, in that larger dataset of LLPS-associated proteins, the ability to undergo LLPS is associated to proteins, not to their regions, so we could not check if the identified charged regions reside within the regions that drive LLPS. We are currently investigating to what extent these charged regions overlap with LLPS driver regions and if splicing could affect them.

The RNA-dependent phase separation mechanism of G3BP1 and 2, and its implications in amyotrophic lateral sclerosis (ALS):

In collaboration with Peter Tompa's group in Brussels, we study the LLPS of the human G3BP1 and G3BP2 proteins that are the major drivers of the formation of stress granules (SGs). Although these proteins are not produced by the alternative splicing of gene products transcribed from the same gene, but are encoded on different genes, they show high sequence and structural similarities apart from some smaller insertions/deletions and are therefore somewhat similar to regular splice variants. Being a key SG protein, the intramolecular interactions and switch-like behaviour of G3BP1 have already been elucidated by others. In its closed conformation two of its oppositely charged IDRs interact and the protein is unable to undergo LLPS. RNA opens it up and induces LLPS and the formation of SGs. G3BP1 was expressed and purified in the lab in Brussels and they have shown that interaction with mRNA moderately increases its LLPS propensity, while electrostatic interaction with arginine-rich dipeptide repeats through its negatively charged first IDR largely increases the LLPS propensity of G3BP1 and makes it RNA-independent. Dipeptide repeats also show much stronger binding to the protein than mRNA. I made a complementary bioinformatics analysis to this impactful study that shows that many of the hitherto discovered cellular targets or arginine-rich dipeptide repeats show similar sequence features to G3BP1. They tend to have high predicted LLPS propensity and harbour positively and negatively charged IDRs and at least one RNA-binding domain, and therefore seem to function through a similar switch-like behaviour as seen for G3BP1. The manuscript reporting on the above results is already deposited to the bioRxiv preprint server (here: https://doi.org/10.1101/2023.03.31.535023), it will be submitted to a prestigious journal soon. The protocol for G3BP2 expression and purification is currently optimized, after successful purification, comparisons of the physiological and pathological LLPS mechanisms of the two proteins will become possible that would be very relevant for the project.

Other observations resulting from the proposed project:

After assembling the PhaSePro database (see https://doi.org/10.1093/nar/gkz848) and later on by following the LLPS literature and assembling a confident driver dataset (see https://doi.org/10.3390/ijms22063017), we have summarized the data on the protein regions confirmed to drive LLPS through in vitro experiments (only in vitro experiments can highlight the minimal requirements of LLPS). As this dataset is gradually growing we see that LLPS driver regions are mostly encoded by constitutive exons. In cases where they are alternatively spliced, usually the whole region is spliced out (all the key phase-separating modules are lost) and therefore we can be quite sure that LLPS is completely abolished in the splice variants, or only a few residues are affected by splicing in a longer, low-complexity LLPS driver region, which is probably not enough to considerably affect LLPS. So, the splicing of LLPS driver regions seems to work in an all or nothing manner. Therefore, there aren't many cases, where a reduced/increased/somehow fine-tuned LLPS is expected based on the region differences and the conformational effects of those differences between the isoforms. Due to these reasons, and the fact that quite some alternatively spliced isoforms have been experimentally studied by other groups in the last few years (see our review article here: https://doi.org/10.1111/febs.15254), it was not an easy task to select isoform pairs for experimental studies that are both interesting from the point of view of differing LLPS mechanisms and not studied yet.

The rise of AlphaFold 2 and its effect on the current project:

AlphaFold (AF) is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of a deep learning algorithm that allows highly accurate sequence-based protein structure prediction. With the rise of AlphaFold and the appearance of AlphaFold 2 structure predictions in public protein databases, like UniProt and the PDBe Knowledge Base (PDBe-KB), some steps of WP4 were re-considered. The conformational differences of the alternatively spliced isoform pairs could be better assessed by AlphaFold 2 than MD simulations or charge patterns, so we downloaded the AF structures of the canonical proteins and performed AF predictions on the relevant isoform sequences for which those were not available yet. We then assessed the conformational differences between the isoforms.

Furthermore, in collaboration with the group of Dr. Wim Vranken in Brussels, we performed an analysis, where we assess the ability of AlphaFold 2 to estimate the conformations of intrinsically disordered and structurally ambiguous protein regions (see publication: https://doi.org/10.3389/fmolb.2022.959956). This is important for the current project, because the alternatively spliced exons often encode disordered protein regions and the splice sites also tend to reside within disordered regions (as shown by our group many years ago in this study: https://doi.org/10.1093/nar/gkq843). Meanwhile, many other research groups in the IDP field have also suggested that AF2 predictions are more effective in identifying disordered regions than classical disorder prediction methods.

Ongoing experimental work in our lab:

For some of the wild type - physiological AS pairs we could gain purified proteins in low amounts, however, mostly not enough for the biophysical experiments planned to address their LLPS behaviour. Since HnRNPA1 and DDX4 have already been studied by other groups quite thoroughly, we mainly focus on RBFOX1 now. Unfortunately, we could not yet gain such data for both counterparts of the wild type-splice variant pair, but we did not give up on that yet, we are still trying to obtain sufficient amounts of the proteins.

For the selected wild type – cancer-associated GATA3 AS pair, Rawan Abukhairan in our lab obtained part of the desired constructs, produced the other ones by mutagenesis and cloned them into expression vectors in the second half of 2022. Expression of the constructs is still under optimization, unfortunately we faced many difficulties with them.

Ongoing collaborative work related to the project:

Peter Tompa’s group in Brussels studied androgen receptor (AR), a nuclear hormone receptor that regulates the transcription of numerous developmental genes. Misregulation of AR is a major cause of prostate cancer (PC). Interestingly, while full-length AR can undergo LLPS in a cellular model of PC, the oncogenic splice variant of AR, AR-v7 that lacks the ligand binding domain (LBD) is incapable of LLPS under similar circumstances. Using full-length AR and its truncation mutants, they have analysed which AR region is responsible for LLPS. They found that the DNA-binding domain (DBD) is capable of RNA binding and undergoes RNA-dependent LLPS (See their publication: PMID:33938068).

Importantly, although the AR LBD does not seem to have a direct role in driving LLPS (the DBD is sufficient in itself), its absence seems to hinder LLPS, maybe due to an intramolecular interaction. Our new results obtained in collaboration suggest that the long N-terminal disordered transactivation domain could inhibit AR-v7 phase separation maybe through interactions with the DBD that are somehow repressed in the full-length protein. In an ongoing project, we are performing experiments on the two isoforms and test several circumstances and interaction partners that could potentially promote the LLPS of the AR-v7 isoform in cells.