Characterization of Liquid-liquid phase Separation in 3D Genome Organization

Tingting Li

institute of basic medical sciences

Jiaqing Xing

institute of basic medical sciences

Tao Li

institute of basic medical sciences

Teng Li

institute of basic medical sciences

Weihua Li (✉ whli@ncba.ac.cn)

institute of basic medical science <https://orcid.org/0000-0002-9547-6292>

Research article

Keywords: proteins, LLPS, RNPII, HP1a, 3D genome organization, GSEA analysis

Posted Date: September 28th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-78761/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Characterization of liquid-liquid phase separation in 3D genome organization

Tingting Li¹, Jiaqing Xing¹, Tao Li¹, Teng Li¹ and Weihua Li^{1*}

¹ State Key Laboratory of Proteomics, Institute of Basic Medical Sciences, National Center of Biomedical Analysis, Beijing, China.

Correspondence: whli@ncba.ac.cn

Abstract

Many proteins have been demonstrated to participate in 3D genome organization through liquid-liquid phase separation (LLPS) such as RNP11, HP1a. However, systematic investigation of relationships between LLPS and 3D genome organization remains lacking. Here, we predicted the intrinsic disordered regions (IDRs) and modular domains of all human proteins and performed GSEA analysis according to their proportions of IDRs. Our results showed that main biological processes involved in 3D genome organization are highly enriched with IDRs, including chromatin organization, RNA splicing and histone modification, demonstrating the key role of LLPS in regulating nuclear structure. Of the 3885 IDR-rich proteins, 1427 proteins are involved in 3D genome organization. IDR regions of these proteins have strong preference of Ser, Leu, Pro, Ala, Gly, Glu and Lys, and lack of hydrophobic amino acids such as Trp, Tyr, Phe and Met, suggesting dipolar interactions rather than aromatic-involved interactions

involved. Further motif enrichment analysis suggests that RNA recognition motif and zinc finger motif are the two most abundant repeatedly-occurred modular domains within IDR-containing proteins. Finally, we developed a Shiny APP named phasepro that interactively analyze and visualize a protein' s potential of LLPS, including IDRs, motifs, amino acid preferences and electric charges.

Introduction

The 3D eukaryotic genome structures are hierarchically organized from megabase-scale A/B compartments, to sub-megabase scale topologically associating domains (TADs), and kilobase-scale chromatin loops[1-4]. How the sophisticated nuclear structure is established and regulated remains a fundamental question in cell biology. Growing number of recent studies suggest a key role of liquid-liquid phase separation (LLPS) in nuclear organization[5-8]. Nuclear membraneless granules such as nucleoli, Cajal bodies and PML bodies are liquid condensates driven by LLPS. Liquid nuclear condensates can pull targeted genomic loci together and push out non-targeted genomic regions[5]. Moreover, transcription factors activate genes through a phase-separation mechanism[9, 10], and HP1a forms droplets suggesting a role of phase separation in heterochromatin formation[7, 8, 11]. Besides these specific studies, systematic investigations of relationships between LLPS and nuclear organization is lacking, despite several LLPS protein databases[12-14]. We thus genome-widely analyzed the roles of LLPS in 3D genome organization, uncovering biological process

relevance, amino acid preferences and modular domain occurrence. We also developed a shiny app facilitating LLPS predication and visualization

Results

GSEA analysis uncovers genome-wide regulation roles of LLPS in core biological processes.

LLPS processes have two types of multi-valence interaction mechanisms: IDR-regulated protein-protein interactions and module-regulated protein-protein interactions[15]. We first predicted all the IDR regions by calculating the iupred scores [16] and protein motifs through interproscan software [17] of all human proteins with a minimum length of 100. The results showed that in the total 19843 human proteins, 3885 proteins have IDR regions longer than 30% of the entire amino acid sequences and 2182 proteins the half of which are IDRs, indicating IDRs are very common genome-widely (Fig 1A, supplementary Table 1). Similarly, we calculated the motifs occurred more than once in the same protein, and the results showed that 4045 proteins contain repeated motifs, and the most abundant repeated motifs are zinc finger (C2H2-type) motifs (Fig 1B, supplementary Table2).

Next, we performed GSEA analysis of the ranked gene list based on IDR ratios calculated by iupred2 to enrich LLPS related biological processes (Supplementary Table 3). As Fig 1C shows, IDR-containing proteins are enriched in dozens of key biological processes including cell cycle, development, neurotransmitter, SUMOylation and etc, which are highly consistent with previous studies [18, 19].

Signaling pathways including Notch, EGFR, WNT are also significantly enriched with proteins containing large IDRs. Specially, several biological processes which have not been reported to related with LLPS such as sex differentiation, circadian clock, cell polarity are enriched, suggesting important roles of LLPS within them.

LLPS is involved in Main 3D genome organizations

We specifically investigated the GSEA-enriched biological processes participate in 3D genome organization. As expected, main chromatin-related biological processes are enriched, including chromatin organization, chromatin remodeling, histone modification, RNA transcription and RNA splicing (Fig. 1C, supplementary Table 4). The most significantly enriched GO term is “regulation of RNA splicing” , followed by “transcription by RNAP II (Fig. 2, Fig. supp1, supplementary Table 3), which have been verified by previous studies [6, 9, 10]. 3D genome directly-related biological processes including “regulation of chromatin organization” , “chromatin assembly or disassembly” , “DNA packaging” , “DNA conformation change” , “chromatin remodeling” and “chromatin silencing” occurred in the top 20 GO terms ranked by their adjusted p-values (Fig. 2, supplementary Table 4). Proteins participate in histone methylation and acetylation also showed significance in IDRs enrichment. We calculate the numbers of proteins containing large IDRs ($\geq 30\%$) in each GO term, and surprisingly found that approximately half or more proteins in each significant term contain large IDR regions. For example, GO term “regulation of chromatin organization” contains 131 proteins and 71 of them (54.2 %) are IDR-containing

proteins, which is much higher than the average level (3885/19844, 19.6 %).

IDR-enriched proteins that participate in 3D genome organization have preference of amino acids and modular domains.

As proteins participate in 3D genome organization prefer to contain IDR regions, we next investigate the sequence features of their IDRs. We first collected all the 3D-genome related proteins that have a high potential of LLPS (supplementary Table 5), and 1427 proteins were identified that have at least 30% of IDR regions. Next, we calculated the frequencies of amino acids within IDR regions. The results clearly showed that IDR-containing proteins that involved in 3D genome organization are rich of Ser, Leu, Pro, Ala, Gly, Glu and Lys, and lack of amino acids that contain large hydrophobic side chain such as Trp, Tyr, Phe and Met (Fig 3). It is reasonable to infer that multivalent interactions between IDRs are mainly dipolar interactions rather than aromatic-involved interactions such as pi-stacking or cation-pi interactions, probably due to their frequent interactions with nucleic acids.

More and more evidences suggest that LLPS are regulated by IDR and modular multivalent interactions together. So, we investigated which repeated modular domains these IDR-containing proteins have. As Fig. supp2 indicates, Zinc finger and RNA recognition motifs (RRM) are the top two repeated modular domains these IDR-containing proteins have, not only in RNA transcription and splicing, but also in chromatin organization and histone modification. Dozens of other motifs also occur at a relatively lower frequency, such as WD 40, K homology

domain and bromodomain (Fig. supp2). Take CTCF as an example, CTCF consists of two large IDR domains at N-terminal and C-terminal, which are separated by seven Zinc finger motifs (Fig. supp3, generated by phasepro shiny app), demonstrating the high potential ability of LLPS.

Developing a shiny APP that facilitate phase separation prediction.

To facilitate the analysis of protein potentials of LLPS, we developed a shiny app named phasepro that integrate IDR prediction, modular domain classification and AA properties together (Fig. 4). Through the interactive app, one only need to input the official gene symbol name and can clearly get a global picture of phase separation potential of selected protein, including IDR regions, amino acid preferences, charges, net charges and motifs (Fig. 4A). The regions and motifs that related with IDR regions is visualized and the preferences of amino acid and charges that may affect LLPS is easy to infer. We also offered sequence and IDR score download options (Fig. 4B) and useful links (Fig. 4C) facilitating further scientific investigation.

Discussion

In this work we systematically studied the regulatory role of LLPS in all biological processes including 3D genome organization. The results are consistent with previous studies that many key biological processes are significantly related with IDR-regulated LLPS such as development, cell cycle and RNA transcription. More

importantly, we found that main nuclear structure related processes are highly enriched with IDR-containing proteins, many of which have not been reported such as chromatin assembly, chromatin remodeling and histone modification. H3K9me3 is reported to interact with the repeated chromodomains of HP1a protein, which is essential for heterochromatin formation. And it is reasonable to conclude that active histone modifications such as H3K4 methylation and H3 acetylation may regulate the euchromatin activity through similar LLPS mechanism.

Our results suggest a key role of RNAs in the entire genome structure organization through LLPS. Over 70% of the mammalian genomes are actively transcribed, and transcriptions and chromatin activities are largely involved with regulatory RNAs. Multiple RNA-binding proteins (RBPs) are localized on active chromatin regions and enable RNA-based regulation of transcription. LINE1 repeat transcripts are strongly associated with L1 DNA sequences and induce LLPS of HP1 α . And our data found that RNA recognition motifs are the most abundant repeated domains within IDR-containing proteins. Taken together, RNA centered multi-valence interactions are important compartments besides IDRs in nuclear structure related LLPS.

Although LLPS has emerged as a fundamental regulation mechanism in almost all important biological processes, specific bioinformatic analysis tools focusing on phase separation is lacking. So we integrated multiple analysis tools together into one interactive Shiny App named phasepro, facilitating specific genes and

genome-wide LLPS analysis. Phasepro app can be downloaded and run within RStudio freely and it is more convenient and powerful in evaluating and visualizing the LLPS potential of interested protein.

Materials and methods

Prediction of IDR regions and modular domains

All sequences of human proteins are downloaded from the ensemble website (<http://asia.ensembl.org/info/data/ftp/index.html>) and only the longest isoforms are kept for each protein. IDR scores are calculated with iupred2a python script[16] for proteins containing more than 99 amino acids and a score larger than 0.5 is regarded as intrinsically disordered. Usually a protein is regarded with strong potential to phase separate if over 30 % percentage of its amino acid scores are larger than 0.5. Modular domains are predicted with interproscan [17] and we use the predicted motifs of SMART[20] for the following analysis.

GSEA and enrichment map analysis

GSEA[21] analysis (desktop version 3.0) was performed with pre-ranked gene list sorted by the ratios of its IDR regions, and the enrichment map gene sets of human GOBP all pathways were used for enrichment score (ES) calculation. Clustered results were visualized by cytoscape combined with EnrichmentMap APP.

3D genome related protein sequence analysis

To obtain biological processes related with both LLPS and 3D genome

organization, 'chromatin' , 'histone' , 'DNA' and 'RNA' were used as keywords to gather the significantly enriched gene sets by GSEA. Proteins within these gene sets that have a higher ratio of 30 % as IDRs are selected and module domain enrichment analysis was performed with DAVID. Amino acid distributions of these IDR regions were calculated.

Shiny app development

The phasepro shiny app was developed based on RStudio and shiny package. Briefly speaking, the amino acid properties of all human coding proteins including amino acid sequences, iupred scores, modular domains and electric charges are calculated. The intermediate code and data for reproducible research are stored at github (<https://github.com/TintingLi/LLPS>), and the entire Shiny app code can be downloaded (<https://github.com/TintingLi/phasepro>) and run within RStudio on any desktop computer or webserver.

Conflicts of interest

The authors declare that they have no competing interests.

Funding

This work was supported by the National Science Fund for Excellent Young Scholars (81922051) .

Authors' contributions

LTT designed this project and wrote the manuscript. XJQ participated in the project, LT and LWH supervised the study. All authors read and approved the final manuscript.

Figure Legends

Figure 1 Gene set enrichment analysis of proteins containing IDRs. A. Statistics of IDR-containing proteins across all human proteins. An amino acid with an iupred2 score larger than 0.5 is regarded as intrinsically disordered and the total ratio of IDR-amino acid is calculated. 30% indicates 30% of a protein's amino acids have a higher score over 0.5. B. Statistics of proteins containing repeated motifs across all human proteins. Motifs are predicted with smart software. C. Enrichment map clustering of GSEA results of all human proteins ranked by their IDR-region ratios.

Figure 2 IDR-enriched gene sets related to 3D genome organization. 3D genome related terms enriched in Fig 1C are extracted and IDR-enriched (>20%) protein numbers are calculated. For example, "regulation of RNA splicing" term consists of 121 proteins and 72 of them contain IDR regions (20% compared to the whole protein sequence).

Figure 3 IDR-enriched protein motifs and PTMs related to 3D genome organization. IDR-enriched proteins belong to specific genome-organization related biological processes are extracted and the preferences of amino acids are evaluated. It is obvious that these IDR-containing proteins have similar selection preferences across these biological processes.

Figure 4 An user interface of protein LLPS investigation by

‘phasepro’ Shiny APP. A. Global picture of a protein’ s properties related to LLPS including amino acid preference, amino acid charges, IDR scores and motifs. TP53 is used as an example. B. Downloads of protein sequences and IDR scores of selected protein. C. related web sites and google search links of selected protein.

Supplementary Figure 1 Enrichment plots of top GSEA enriched terms related to 3D genome. Enrichment plots of top four IDR-enriched terms: “Regulation of RNA splicing” , “Transcription by RNA polymerase II” , “Regulation of Chromatin organization” and “Chromatin assembly or disassembly” are displayed.

Supplementary Figure 2 Statistics of repeated motifs in IDR-enriched proteins involved in 3D genome organization. IDR-containing proteins involved in 3D genome organization were first extracted and the repeated motifs within these proteins are analyzed and visualized through boxplot.

Supplementary Figure 3 Analysis of CTCF’ s LLPS potentials through phasepro App. The ability of LLPS of CTCF was analyzed and visualized through ‘phasepro’ Shiny APP.

Reference

1. Bonev, B. and G. Cavalli, *Organization and function of the 3D genome*. Nat Rev Genet, 2016. **17**(11): p. 661-678.

2. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions*. Nature, 2012. **485**(7398): p. 376-80.
3. Rao, S.S., et al., *A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping*. Cell, 2014. **159**(7): p. 1665-80.
4. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome*. Science, 2009. **326**(5950): p. 289-93.
5. Shin, Y., et al., *Liquid Nuclear Condensates Mechanically Sense and Restructure the Genome*. Cell, 2018. **175**(6): p. 1481-+.
6. Sabari, B.R., et al., *Coactivator condensation at super-enhancers links phase separation and gene control*. Science, 2018. **361**(6400).
7. Strom, A.R., et al., *Phase separation drives heterochromatin domain formation*. Nature, 2017. **547**(7662): p. 241-+.
8. Larson, A.G., et al., *Liquid droplet formation by HP1 α suggests a role for phase separation in heterochromatin*. Nature, 2017. **547**(7662): p. 236-+.
9. Boija, A., et al., *Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains*. Cell, 2018. **175**(7): p. 1842-1855 e16.
10. Lu, H.S., et al., *Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II*. Nature, 2018. **558**(7709): p. 318-+.

11. Sanulli, S., et al., *HP1 reshapes nucleosome core to promote phase separation of heterochromatin*. Nature, 2019. **575**(7782): p. 390-394.
12. You, K., et al., *PhaSepDB: a database of liquid-liquid phase separation related proteins*. Nucleic Acids Res, 2019.
13. Meszaros, B., et al., *PhaSePro: the database of proteins driving liquid-liquid phase separation*. Nucleic Acids Res, 2019.
14. *LLPSDB: a database of proteins undergoing liquid-liquid phase separation in vitro*. Nucleic Acids Research, 2019.
15. Banani, S.F., et al., *Biomolecular condensates: organizers of cellular biochemistry*. Nature Reviews Molecular Cell Biology, 2017. **18**(5): p. 285-298.
16. Meszaros, B., G. Erdos, and Z. Dosztanyi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding*. Nucleic Acids Research, 2018. **46**(W1): p. W329-W337.
17. Mitchell, A.L., et al., *InterPro in 2019: improving coverage, classification and access to protein sequence annotations*. Nucleic Acids Res, 2019. **47**(D1): p. D351-D360.
18. Alberti, S., A. Gladfelter, and T. Mittag, *Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates*. Cell, 2019. **176**(3): p. 419-434.
19. Hahn, S., *Phase Separation, Protein Disorder, and Enhancer Function*. Cell, 2018. **175**(7): p. 1723-1725.

20. Letunic, I. and P. Bork, *20 years of the SMART protein domain annotation resource*. Nucleic Acids Research, 2018. **46**(D1): p. D493-D496.
21. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.

Fig. 1 Gene set enrichment analysis of Proteins containing IDRs

Fig. 2 IDR-enriched gene sets related to 3D genome organization

Fig.3 IDR-enriched motifs and PTMs related to 3D genome organization

Fig.4 phasepro shiny APP

A

B

C

Fig. supp1 GESA results of top enriched terms related to 3D genome

Fig.supp2 Statistics of repeated motifs in IDR-enriched proteins involved in 3D genome organization

Fig. supp3 LLPS potential analysis of CTCF

Figures

Figure 1

Gene set enrichment analysis of Proteins containing IDRs

Figure 2

IDR-enriched gene sets related to 3D genome organization

Figure 3

IDR-enriched motifs and PTMs related to 3D genome organization

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [figS1.png](#)
- [figS2.png](#)
- [suppTables.xls](#)
- [figS3.png](#)