dSCOPE: a software to detect sequences critical for liquid-liquid phase separation

Shihua Li[1, 2, †], Kai Yu[1, †], Qingfeng Zhang[1, †], Zekun Liu[1], Jia Liu[1], Huai-Qiang Ju[1], Zhixiang Zuo[1], Xiaoxing Li[3], Zhenlong Wang[2], Han Cheng[2], Ze-Xian Liu[1, *]

[1] State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

[2] School of Life Sciences, Zhengzhou University, Zhengzhou 450001, China

[3] Institute of Precision Medicine, First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510080, China

[†]These authors contributed equally to this work.

*To whom correspondence should be addressed.

Dr. Ze-Xian Liu, Tel/Fax: +86 20 8734 2522; Email: liuzx@sysucc.org.cn

ABSTRACT

Membrane based cells are the fundamental structure and function units of organisms, while evidences were increasing that liquid-liquid phase separation (LLPS) is associated with the formation of membraneless organelles, such as P-bodies, nucleoli and stress granules. Many studies have been undertaken to explore the functions of protein phase separation, but these studies lacked an effective tool to identify the sequence segments that critical for LLPS (SCOPEs). In this study, we presented a novel software called dSCOPE (http://dscope.omicsbio.info) to predict the SCOPEs. To develop the predictor, we curated experimentally identified sequence segments that can drive LLPS from published literature. Then sliding sequence window based physiological, biochemical, structural and coding features were integrated by random forest algorithm to perform prediction. Through rigorous evaluation, dSCOPE was demonstrated to achieve satisfactory performance. Furthermore, large-scale analysis of human proteome based on dSCOPE showed that the predicted SCOPEs enriched various protein post-translational modifications and cancer mutations, and the proteins which contain predicted SCOPEs enriched critical cellular signaling pathways. Taken together, dSCOPE precisely predicted the protein sequence segments critical for LLPS, with various helpful information visualized in the webserver to facilitate LLPS related research.

INTRODUCTION

Due to the existence of different organelles, eukaryotic cells are divided into different functional domains (1). A growing number of studies have indicated that protein phase separation governs the formation of membraneless organelles, such as nucleoli, stress granules and P bodies, in cells (2-4). In addition, subcellular structures, such as heterochromatin, are also formed by phase separation and have similar potential interactions and physical properties (5,6). When proteins undergo phase separation, they condense into a dense phase that is usually similar to droplets, and the dense phase coexists with the dilute phase (7,8). Phase separation plays critical roles in many biological processes, such as signal transduction, RNA metabolism, and autophagy (8-11). Therefore, to understand the regulation and molecular mechanisms governing phase separation, it is urgently important to identify sequences that are critical for phase separation.

With widespread attention being devoted to phase separation, many databases related to phase separation have been developed. For example, Ning *et al.* (12) collected proteins that are involved in LLPS and constructed an LLPS-associated protein database named DrLLPS, which contains 7,993 scaffolds, 72,300 regulators and 357,594 clients in 164 eukaryotic species. PhaSepDB integrates 2,914 phase separation-associated proteins, and it provides such information as publication source, sequence features and immunofluorescence images of phase separation-associated proteins (13). At present, there are few experimentally verified phase-separated proteins, and the LLPSDB contains only 273 experimentally verified phase-separated proteins (14). The standard of the experimental method to identify the phase-separated structure is that it can form a spherical structure and can be fused. Fluorescence recovery after photobleaching (FRAP) is often performed on droplets as an assessment of their liquidity (15-17). Due to the time-consuming and labor-intensive nature of this method, there is an urgent need to develop an *in silico* method to accurately predict the crucial sequence that drives phase separation, which can facilitate the experimental discovery of phase separation and its mechanism.

In recent years, considerable progress has been made in dissecting the sequence features of proteins that can be phase-separated under physiological conditions. These studies indicate that only certain protein sequences have the ability to undergo phase separation under the proper conditions in living cells. Determining the molecular properties of proteins is crucial to understanding their potential phase behavior. For example, many proteins involved in LLPS have been shown to contain prion-like regions and intrinsically disordered regions. Therefore, the prion-like region prediction tool PLAAC (18) and disorder prediction tools, such as IUPred (19), PONDR-FIT (20) and MobiDB (21), are often used to predict phase separation regions in current research. In addition, hydrophobicity and charged residues are also considered to affect electrostatic interactions, which is further related to phase behavior (22,23). Recently, Vernon *et al.* also compared the prediction performance of these features and pointed out that combining multiple features may facilitate the development of a more accurate prediction tool (24). However, to the best of our knowledge, a tool that can predict the phase separation region by combining multiple features has not been presented to date.

In this study, we built an effective tool for phase separation region detection. First, we manually searched the PubMed literature database and collected all experimentally verified sequences related to phase separation. The final dataset contained 121 phase separation regions in 80 proteins. We intercepted protein sequences through a sliding window of 15 amino acids, and the peptides in the sequences critical for phase separation were defined as positive peptides. Next, eight physiological and biochemical feature scores, including disorder, prion-like, polar, relative surface accessibility (RSA), charge, hydropathy, exposure and low complexity, and four sequence structure features, including amino acid composition (AAC), composition of k-spaced amino acid pairs (CKSAAP), position-specific scoring matrix (PSSM) and binary encoding profiles (BE), were extracted and modeled with a random forest algorithm. The TOPT package was utilized to adjust the hyperparameters and to optimize the prediction model. By 4-, 6-, 8-, and 10-fold cross-validation in the training dataset, dSCOPE showed excellent robustness and satisfactory performance. In addition, by utilizing dSCOPE, we comprehensively analyzed the relationship between SCOPEs and tumor mutations, as well as posttranslational modifications, which may provide helpful information for the diagnosis and treatment of some diseases.

MATERIALS AND METHODS

Data collection and preparation

To collect the experimentally identified sequences that can drive protein phase separation, we employed the keyword “phase separation” to retrieve the published literature from PubMed (http://www.ncbi.nlm.nih.gov/pubmed), and eventually, we collected 121 sequences from 80 proteins that were essential for phase separation. The sequence of each protein was retrieved from the UniProt database (25). We treated all regions that play a crucial role in phase separation as positive sequences, and other fragments of the same protein were treated as the negative sequences. For each dataset, we generated 15-length peptides through the sliding window with a step size of 8. Among these peptides, the peptides in human proteins were treated as the training dataset, while the peptides in yeast proteins were treated as the testing dataset. In total, we obtained 1,737 positive peptides and 3,125 negative peptides for the training dataset and 379 positive peptides and 1,075 negative peptides for the testing dataset.

Feature extraction

Physicochemical properties

In recent years, researchers have made considerable progress in analyzing the sequence characteristics of proteins that can undergo phase separation under physiological conditions. The common feature of LLPS proteins is the presence of an intrinsic disordered region (IDR) with multiple interacting motifs (16,26). Charge pattern, amino acid composition, and solubility also affect phase separation (27). We calculated the disorder scores of the proteins by IUPred (19), the per-residue prion-like scores were obtained from PLAAC (18), exposure and surface accessibility analysis were performed by NetSurfP (28), the hydrophobicity was based on the theory of Kyte, J *et al*. (29), as well as charge from Fauchere *et al.* (30), and we used StatSEG (https://github.com/jszym/StatSEG) to obtain the low-complexity region scores. In addition, we also considered the polarity of amino acids.

Composition of k-spaced amino acid pairs (CKSAAP)

Similar to Zhao *et al*. (31), we used the composition ratio of residue pairs of k intervals in the protein sequence fragments in the sequence to establish a mathematical model and extract feature vectors. In other words, if a peptide consists of 20 kinds of amino acids, each amino acid and its next adjacent amino acid form a pair of extracted amino acids, that is, the separation distance between these two amino acids is k = 0 amino acids, then there are 400 possible amino acid pairs (e.g., AA, AC, AD, and so on). According to the probability of these residue pairs appearing in this protein sequence, a 441-dimensional feature vector is generated. With the increase in the k value, although the accuracy and sensitivity of the prediction model increases, the calculation time and cost of the random forest model training also increases notably. In this regard, only the CKSAAP coding with k values equal to 0, 1, 2, and 3 are considered; therefore, the total dimension of the feature vector is 400 × 4 = 1,600.

Position-specific scoring matrix (PSSM)

PSSM is a common feature extraction method in biological sequence analysis, also known as the position weight matrix (32,33). This matrix has 20 × M elements, where M is the length of the target sequence. The occurrence frequency of different amino acids at each position in the matrix was calculated, and the details are as follows:

$$\begin{cases} f_1 = P(AA_1) \\ f_2 = P(AA_2) \\ \vdots \quad \vdots \\ f_{15} = P(AA_{15}) \\ f_{16} = N(AA_1) \\ f_{17} = N(AA_2) \\ \vdots \quad \vdots \\ f_{30} = N(AA_{15}) \end{cases} \tag{1}$$

In Eq. 1, the peptides consist of 15 amino acids; P(X1) represents the occurrence frequency of amino acid AA1 at position 1 in the positive group, while N(X1) denotes the occurrence frequency of amino acid AA1 at position 1 in the negative group. Therefore, each peptide can be represented by a position weight amino acid composition vector with dimensions of 30.

Amino acid composition (AAC)

AAC is an elementary feature and describes the frequency of occurrence of each amino acid in the sequence (34). The dimension of AAC is 20 in this work.

Binary encoding profiles (BE)

Binary encoding is similar to the binary language of computers (35). We converted each sequence into a combination of 20-dimensional vectors. For example, if a sequence is ARDCQEHIGNLKMFPSTWYV, then amino acid A corresponds to (10000000000000000000), and amino acid V corresponds to (00000000000000000001). In this work, the vector size is 300.

Machine learning classifiers

To predict the sequence that is important to phase separation, the random forest algorithm is introduced into our prediction. For a given protein sequence, a short peptide consisting of 15 amino acids in length is intercepted through a sliding window. Each adjacent short peptide needs to be separated by 8 amino acid residues, and short peptides less than 15 in length are removed. Protein fragments are encoded by eight physical and chemical features scores and four feature extraction methods. Next, we tested the performance of five ML classifiers, including logistic regression, random forest, LDA, AdaBoost, and KNN, then adopted the final algorithm for prediction based on the performance. Moreover, to generate the optimal performance, the TOPT (https://github.com/EpistasisLab/tpot) package was integrated to optimize the hyperparameters.

Performance evaluation

We used the following key measurements to evaluate the performance of the model: specificity (*Sp*), sensitivity (*Sn*), accuracy (*Ac*), and receiver operating characteristic (ROC) curves were drawn, and AUC (area under ROC) values were calculated. The measurements were defined as follows:

$$\mathrm{Sn} = \frac{TP}{TP+FN}$$

$$\mathrm{Sp} = \frac{TN}{TN+FP}$$

$$\mathrm{Ac} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$$

In this work, 5- and 10-fold cross-validation was performed. In addition, an independent testing set was used to prove the advantages of our model compared to the current commonly used tools.

Implement of the webserver

We constructed the dSCOPE web server in Python, PHP, JavaScript and HTML, which can be freely accessed at http://dscope.omicsbio.info. NetSurfP was adopted to predict the surface accessibility and secondary structure information for each sequence, the query protein disorder information was calculated by IUPred (19), the per-residue prion-like scores were obtained from PLAAC (21), and the charge pattern from Fauchere *et al.* (30), amino acid composition, and solubility also affect phase separation (27). Additionally, the hydrophobicity was determined based on the theory of Kyte, J *et al*. (29), and subcellular location was obtained from UniProt.

Molecular condensate enrichment analysis

We used dSCOPE to predict 3,633 human proteins related to phase separation retrieved from DrLLPS and selected 998 proteins that contain the potential SCOPEs. Then, we performed molecular condensate enrichment analysis with the 998 proteins, and Fisher's exact test was used to test the significance. We further used WocEA to display the result in the word cloud (36).

Kinase and TF enrichment analysis

Kinase- and phosphorylation-related information was collected and curated from multiple phosphorylation databases verified by a number of experiments, including dbPTM (37), BIOGRID (38), PHOSIDA (39), phosphoELM (40), PhosphositePlus (41), and RegPhos2.0 (42), which included 192,111 phosphorylation sites. Next, we used Fisher's exact test to perform kinase enrichment analysis on potential phase-separated proteins. In addition, transcription factor (TF) enrichment analysis used ChEA3 (43).

Pathway enrichment analysis

To better understand the potential function of phase separation proteins, we used the R ClusterProfiler package for Gene Ontology (GO) function annotation (44), and the enrichment analysis of these proteins in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways was performed using KOBAS (45). The visible network was constructed using Cytoscape (46).

Mutation analysis of phase separation-related regions

We downloaded GDC TCGA somatic mutations from 11 cancer types (BLCA, BRCA, CESC, COAD, HNSC, LIHC, LUAD, LUSC, SKCM, STAD, UCEC) from the Xena Browser (https://xenabrowser.net/datapages/), removed redundant mutations, and retained only missense mutations for further analysis. We generated peptide windows composed of 15 amino acids. Each region in a protein would generate two peptides: one was extracted from the origin sequence, and the other was from the new sequence obtained after mutation. Finally, we predicted the probability for these sequence windows before and after mutation based on dSCOPE. Based on the prediction result of dSCOPE, we further divided the mutation into two types according to whether the SCOPEs were affected.

RESULTS

Construction of the computational model to predict the crucial region of LLPS

We used the keyword "phase separation" to manually collect experimentally confirmed phase-separated proteins from the literature (Figure 1), and we obtained the training and testing datasets as mentioned above. We eventually obtained 1,737 positive sequences and 3,125 negative sequences for *Homo sapiens* as the training dataset and 379 positive sequences and 1,075 negative sequences for *Saccharomyces cerevisiae* as the testing dataset. Then, we developed dSCOPE software for detecting sequences critical for phase separation-related proteins based on random forest and TPOT algorithm. Eight physicochemical property (PP) scores of the sequences were extracted, including disorder, exposure, polar, low complexity region, charge, prion-like region, surface accessibility, and hydropathy. In addition, the sequence features, including AAC, CKSAAP, PSSM, and BE, were also considered (Figure 1). Furthermore, we used Python, PHP, JavaScript and HTML to construct the dSCOPE online server, which can be accessed through http://dscope.omicsbio.info. Finally, a series of further analyses were performed using dSCOPE, including molecular condensates, posttranslational modification (PTM) analysis of potential SCOPEs, GO functional annotation, KEGG pathway, kinase, and transcription factor enrichment (Figure 1).

Evaluating the performance of dSCOPE

We compared the eight reported features related to LLPS between SCOPEs and non-SCOPEs (Figure 2A). This comparison indicated that these features are significantly different in SCOPEs and non-SCOPEs. The SCOPE shows more disorder, the sequence net charge approaches zero, the complexity is low, and the hydrophilic polar amino acids are enriched. This result is consistent with the findings of a previous report (47). In the analysis of amino acid frequency, glycine (G), asparagine (N), proline (P), glutamine (Q), serine (S) and tyrosine (Y) appeared more frequently in the key regulatory sequence of phase separation (Figure 2B), and most of these amino acids are polar amino acids.

Based on the model architecture described in the method, we compared five machine learning algorithms through 5-fold cross-validation. As shown in Figure 2C, the random forest algorithm outperformed the other ML algorithms, and it has higher robustness and faster calculation speed. In addition, we intercepted different lengths and used 5-fold cross-validation to evaluate its performance. When the length increased, the AUC value gradually increased, but when the length reached 16, it decreased instead (Figure 2D). Therefore, we chose a length of 15 to ensure prediction performance. Then, we developed dSCOPE software for detecting sequences critical for phase separation-related proteins based on random forest algorithm, and the hyperparameters were optimized by the TPOT package. 4-, 6-, 8- and 10-fold cross-validation were employed to evaluate the prediction performance of dSCOPE. The ROC curves were shown and AUCs were calculated (Figure 2E). As shown in Figure 2E, the AUC values were 0.8204 (4-fold), 0.8129 (6-fold), 0.8238 (8-fold), and 0.8213 (10-fold). The results of the various validations were very similar to one another, which indicated that dSCOPE is a stable and robust predictor.

To verify the superiority of the dSCOPE model, we compared it with two widely used phase-separated region detection tools, IUPred (19) and PLAAC (18), even though they were not developed for phase separation prediction. We evaluated the prediction performances of these tools using the testing dataset. The AUC values for dSCOPE were 0.8463, while those for IUPred and PLAAC were 0.6854 and 0.6130 (Figure 2F), respectively. In summary, our prediction tool achieved good prediction performance.

Development and application of dSCOPE

We developed a freely available website for phase separation prediction to facilitate scientific research. The dSCOPE was implemented in Python, PHP, JavaScript and HTML, and the prediction page is shown in Figure 3A. Users should paste the FASTA format sequences of their proteins of interest into the text box and then choose an organism and a threshold to obtain the result. We also provided some examples to show the usage of dSCOPE; meanwhile, users only need to click the “Example” button to see the default protein sequences and their predicted results. Users can also use UniProt ID, gene name or protein name in the search interface to directly determine the predicted results of human protein (Figure 3B). On the results page, the prediction information was organized into three sections, including “dSCOPE prediction results”, “Information of the protein” and “Sequence and structural characteristics of the protein” (Figure 3C-D). The “dSCOPE prediction results” was a detailed description of the prediction results (Figure 3D). If the users wanted to search for predicted results regarding a reviewed human protein by UniProt ID, gene name or protein name, the protein details annotated by UniProt were displayed in "Information of the protein" (Figure 3C). The "Sequence and structural characteristics of the protein" included the protein predicted potential phase separation regions, the dSCOPE score for each residue, the sequence and structure properties, and the protein subcellular location information (Figure 3D). Supporting multiple protein sequence predictions, users can select the protein prediction information to display by clicking on the selection box at the top. In conclusion, dSCOPE appears to be a comprehensive web server for protein phase separation prediction to facilitate related research.

To better characterize the application of dSCOPE, we predicted 3,633 human proteins related to phase separation retrieved from DrLLPS and selected 998 proteins containing the potential SCOPE. The prediction results showed that proteins potentially containing sequences that are critical for phase separation are significantly enriched in the P-bodies, stress granules, and nuclear speckles (Figure 2G). Previous studies have also shown that phase-separated proteins primarily exist in membraneless organelles, such as P-bodies and stress granules.

Functional analysis of potential phase separation proteins

In phase-separated proteins, the sequences critical for phase separation possess numerous disordered regions, and these regions are susceptible to various PTMs (48). To further investigate whether the protein containing SCOPE extracted from dSCOPE has the potential for phase separation, we predicted 20,380 reviewed human proteins, and under the condition of FDR<1%, 4,637 proteins containing potential sequences essential for phase separation were selected. Therefore, we employed the data of PLMD (http://plmd.biocuckoo.org/) and EPSD (http://epsd.biocuckoo.cn/) to analyze and understand the relationship between SCOPE and PTMs. Lysine modification usually was evenly distributed between protein sequences but was slightly enriched in the potential SCOPE (E-ratio = 1.40, *p-value* = 5.04 x 10^{-12}, Fisher's exact test) (Figure 4A), and the phosphorylation sites were also enriched in the SCOPE (E-ratio = 1.25, *p-value* = 1.17 x 10^{-123}, Fisher's exact test) (Figure 4B).

Not only that, we also performed GO functional annotation. As shown in Figure 4C, the proteins are highly enriched in chromatin and various protein complexes. In addition, many proteins may be involved in transcription and may be related to gene expression and regulation. KEGG pathway enrichment analysis revealed that potential phase-separated proteins are significantly enriched in several pathways related to proliferation and apoptosis, such as the Hippo signaling pathway, the Notch signaling pathway and signaling pathways regulating pluripotency of stem cells (Figure 4D). It can be observed from previous studies that these pathways contain various posttranslational modification processes, especially phosphorylation (49-51).

Among the 4,637 proteins with potential SCOPE, 3,099 proteins have phosphorylation sites. In addition, 841 proteins have phosphorylation sites for known kinases. We used Fisher's exact test to perform kinase enrichment analysis on phosphoproteins containing known kinases. As shown in Figure 4E, we found that MAPK1, SRC, CDK1, and CDK2 were significantly enriched. SRC has been shown to play a key role in the phase separation of the FUS and tau proteins (52,53). In addition, phase-separated proteins have been extensively confirmed to be involved in gene expression and regulation, and identifying transcription factors that lead to changes in gene expression is an important step in elucidating gene regulation networks (54). Therefore, we performed a transcription factor enrichment analysis using ChEA3 (43). The enrichment results are shown in Figure 4F. The top 5 TFs were SPEN, SRCAP, PRR12, HOXB3, and CIC. In these enriched TFs, CIC usually cooperates with ATXN1 to play a role in the development of the central nervous system, and ATXN1 is a proven phase separation protein(55). In addition, SPEN also regulates the notch signaling pathway (56).

DISCUSSION

In recent years, research on the mechanism of phase separation of biomolecules has increased rapidly. Accumulating evidence demonstrates that LLPS underlies the formation of membraneless organelles (2). Furthermore, LLPS is also related to various neurodegenerative diseases, such as amyotrophic lateral sclerosis (ALS) (57,58), Alzheimer's disease (AD) (59) and frontotemporal dementia (FTD) (60). However, the prediction of protein LLPS through computational methods remains a challenge. PLAAC developed with hidden Markov is a tool that is often used to predict phase separation, but its initial development was to identify prion-like sequences (18). Vernon *et al.* compared the prediction performance of six sequence-based algorithms and one empirical approach that has recently been employed to predict protein phase separation (24). It was observed that considering multiple features may help researchers obtain more accurate predictions. Hence, we developed dSCOPE in consideration of various features.

Most sequences critical for phase separation contain IDRs, which cause them lack secondary structures and therefore be particularly susceptible to PTMs, especially phosphorylation and lysine modification (48,61). This property was also demonstrated by the analysis of lysine and phosphorylation modifications. And in order to explore the relationship between cancer mutations and phase separation, we conducted a mutation analysis on the collected phase-separated proteins and proteins predicted by dSCOPE and found that mutations are more likely to occur in non-SCOPE phase-separated proteins (Supplementary Figure 1A-B).

Previous studies have shown that phase separation plays an important role in chromosome structure (62). This role is consistent with our analysis of potential phase separation proteins. In addition, KEGG enrichment analysis of potential phase-separated proteins showed that pathways related to proliferation, apoptosis and signal transduction processes had significant enrichment. Increasing experimental evidence shows that phase separation is involved in the process of cell autophagy (63). The functional preference of these predicted proteins is consistent with the characteristics of experimentally confirmed phase-separated proteins, indicating that our tool is reliable, and these analyses may provide experimental researchers with some new ideas.

dSCOPE is the first tool to combine multiple features to predict the sequences that are critical for phase separation. Various feature scores, including disorder, prion-like, polar, RSA, charge, hydropathy, exposure and low complexity, are examined. Although dSCOPE has achieved good prediction performance, there are many aspects that warrant improvement. It is well-known that a larger training dataset produces more accurate predictive performance. In the future, experimentally identified sequences with phase separation proteins will be continuously collected from the literature and integrated into the predictive model when available. Furthermore, more machine learning algorithms need to be considered, such as deep neural networks (DNNs) and recurrent neural networks (RNNs), which may also improve the current prediction performance.

DATA AVAILABILITY

The dSCOPE server is freely available at http://dscope.omicsbio.info.

ABBREVIATIONS

LLPS, liquid-liquid phase separation; SCOPE, sequence critical for phase separation; ML, machine learning; RNN, recurrent neural network; DNN, deep neural network; LDA, Linear Discriminate Analysis; KNN, K-Nearest Neighbor; PP, physicochemical properties; CKSAAP, composition of k-spaced amino acid pairs; BE, binary encoding profiles; PSSM, position-specific scoring matrix; AAC, amino acid composition; FDR, false discovery rate; Sp, specificity; Sn, sensitivity; Ac, accuracy; ROC, receiver operating characteristic; AUC, area under ROC curve; RSA, relative surface accessibility; TF, transcription factor; mK, modified lysine residue; umK, unmodified lysine residue; pSTY, phosphorylated STY residue; upSTY, unphosphorylated STY residue.

FUNDING

This work was supported by grants from the Natural Science Foundation of China (91953123 to ZX.L., 31601067 to H.C.); Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2017ZT07S096 to ZX.L.); Pearl River S&T Nova Program of Guangzhou (201906010088 to ZX.L.); Tip-Top Scientific and Technical Innovative Youth Talents of Guangdong Special Support Program (2019TQ05Y351 to ZX.L.); and Key program for Department of Science and Technology of Qinghai province (2017-ZJ-Y13 to H.C.).

CONFLICT OF INTEREST

The authors declare no competing financial interests.

REFERENCES

1. Zhang, B. and Herman, P.K. (2019) It is all about the process(ing): P-body granules and the regulation of signal transduction. *Current genetics*.
2. Hyman, A.A., Weber, C.A. and Julicher, F. (2014) Liquid-liquid phase separation in biology. *Annual review of cell and developmental biology*, **30**, 39-58.
3. Banani, S.F., Lee, H.O., Hyman, A.A. and Rosen, M.K. (2017) Biomolecular condensates: organizers of cellular biochemistry. *Nature reviews. Molecular cell biology*, **18**, 285-298.
4. Boeynaems, S., Alberti, S., Fawzi, N.L., Mittag, T., Polymenidou, M., Rousseau, F., Schymkowitz, J., Shorter, J., Wolozin, B., Van Den Bosch, L. *et al.* (2018) Protein Phase Separation: A New Phase in Cell Biology. *Trends in cell biology*, **28**, 420-435.
5. Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S. and Narlikar, G.J. (2017) Liquid droplet formation by HP1alpha suggests a role for phase separation in heterochromatin. *Nature*, **547**, 236-240.
6. Strom, A.R., Emelyanov, A.V., Mir, M., Fyodorov, D.V., Darzacq, X. and Karpen, G.H. (2017) Phase separation drives heterochromatin domain formation. *Nature*, **547**, 241-245.
7. Dolgin, E. (2018) What lava lamps and vinaigrette can teach us about cell biology. *Nature*, **555**, 300-302.
8. Feng, Z., Chen, X., Zeng, M. and Zhang, M. (2018) Phase separation as a mechanism for assembling dynamic postsynaptic density signalling complexes. *Current opinion in neurobiology*, **57**, 1-8.
9. Zhang, G., Wang, Z., Du, Z. and Zhang, H. (2018) mTOR Regulates Phase Separation of PGL Granules to Modulate Their Autophagic Degradation. *Cell*, **174**, 1492-1506.e1422.
10. Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A.S., Yu, T., Marie-Nelly, H., McSwiggen, D.T., Kokic, G., Dailey, G.M., Cramer, P. *et al.* (2018) RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nature structural & molecular biology*, **25**, 833-840.
11. Sun, D., Wu, R., Li, P. and Yu, L. (2019) Phase Separation in Regulation of Aggrephagy. *Journal of molecular biology*.
12. Ning, W., Guo, Y., Lin, S., Mei, B., Wu, Y., Jiang, P., Tan, X., Zhang, W., Chen, G., Peng, D. *et al.* (2019) DrLLPS: a data resource of liquid-liquid phase separation in eukaryotes. *Nucleic acids research*, gkz1027.
13. You, K., Huang, Q., Yu, C., Shen, B., Sevilla, C., Shi, M., Hermjakob, H., Chen, Y. and Li, T. (2019) PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic acids research*, gkz847.
14. Li, Q., Peng, X., Li, Y., Tang, W., Zhu, J.a., Huang, J., Qi, Y. and Zhang, Z. (2019) LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro. *Nucleic acids research*.
15. Mitrea, D.M., Chandra, B., Ferrolino, M.C., Gibbs, E.B., Tolbert, M., White, M.R. and Kriwacki, R.W. (2018) Methods for Physical Characterization of Phase-Separated Bodies and Membrane-less Organelles. *Journal of molecular biology*, **430**, 4773-4805.
16. Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C.C., Eckmann, C.R., Myong, S. and Brangwynne, C.P. (2015) The disordered P granule protein LAF-1 drives phase separation into

droplets with tunable viscosity and dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 7189-7194.

17. Zhang, H., Elbaum-Garfinkle, S., Langdon, E.M., Taylor, N., Occhipinti, P., Bridges, A.A., Brangwynne, C.P. and Gladfelter, A.S. (2015) RNA Controls PolyQ Protein Phase Transitions. *Molecular cell*, **60**, 220-230.
18. Lancaster, A.K., Nutter-Upham, A., Lindquist, S. and King, O.D. (2014) PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics (Oxford, England)*, **30**, 2501-2502.
19. Dosztányi, Z., Csizmok, V., Tompa, P. and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics (Oxford, England)*, **21**, 3433-3434.
20. Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K. and Uversky, V.N. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochimica et biophysica acta*, **1804**, 996-1010.
21. Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Mičetić, I., Camilloni, C., Davey, N., Dosztányi, Z., Mészáros, B., Monzon, A.M. *et al.* (2017) MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic acids research*, **46**, D471-D476.
22. Wang, J., Choi, J.M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D. *et al.* (2018) A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell*, **174**, 688-699.e616.
23. Aumiller, W.M., Jr. and Keating, C.D. (2016) Phosphorylation-mediated RNA/peptide complex coacervation as a model for intracellular liquid organelles. *Nature chemistry*, **8**, 129-137.
24. Vernon, R.M. and Forman-Kay, J.D. (2019) First-generation predictors of biological protein phase separation. *Current opinion in structural biology*, **58**, 88-96.
25. UniProt, C. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, **47**, D506-D515.
26. Smith, J., Calidas, D., Schmidt, H., Lu, T., Rasoloson, D. and Seydoux, G. (2016) Spatial patterning of P granules by RNA-induced phase separation of the intrinsically-disordered protein MEG-3. *eLife*, **5**.
27. Alberti, S. (2017) Phase separation in biology. *Current biology : CB*, **27**, R1097-r1102.
28. Klausen, M.S., Jespersen, M.C., Nielsen, H., Jensen, K.K., Jurtz, V.I., Sønderby, C.K., Sommer, M.O.A., Winther, O., Nielsen, M., Petersen, B. *et al.* (2019) NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins*, **87**, 520-527.
29. Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, **157**, 105-132.
30. Fauchère, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology. *International journal of peptide and protein research*, **32**, 269-278.
31. Zhao, X., Zhang W Fau - Xu, X., Xu X Fau - Ma, Z., Ma Z Fau - Yin, M. and Yin, M. Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs.

32. Ou, Y.-Y., Gromiha, M.M., Chen, S.-A. and Suwa, M. (2008) TMBETADISC-RBF: Discrimination of β-barrel membrane proteins using RBF networks and PSSM profiles. *Computational biology and chemistry*, **32**, 227-231.

33. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, **292**, 195-202.

34. Shi, J., Zhang, S., Liang, Y. and Pan, Q. (2006) In Rajapakse, J. C., Wong, L. and Acharya, R. (eds.), *Pattern Recognition in Bioinformatics*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 105-114.

35. Song, J., Tan H Fau - Shen, H., Shen H Fau - Mahmood, K., Mahmood K Fau - Boyd, S.E., Boyd Se Fau - Webb, G.I., Webb Gi Fau - Akutsu, T., Akutsu T Fau - Whisstock, J.C. and Whisstock, J.C. Cascleave: towards more accurate prediction of caspase substrate cleavage sites.

36. Ning, W., Lin, S., Zhou, J., Guo, Y., Zhang, Y., Peng, D., Deng, W. and Xue, Y. (2018) WocEA: The visualization of functional enrichment results in word clouds. *Journal of genetics and genomics = Yi chuan xue bao*, **45**, 415-417.

37. Huang, K.-Y., Lee, T.-Y., Kao, H.-J., Ma, C.-T., Lee, C.-C., Lin, T.-H., Chang, W.-C. and Huang, H.-D. (2018) dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. *Nucleic acids research*, **47**, D298-D308.

38. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic acids research*, **34**, D535-539.

39. Gnad, F., Gunawardena, J. and Mann, M. (2011) PHOSIDA 2011: the posttranslational modification database. *Nucleic acids research*, **39**, D253-260.

40. Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J. and Diella, F. (2011) Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic acids research*, **39**, D261-D267.

41. Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V. and Sullivan, M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research*, **40**, D261-D270.

42. Huang, K.Y., Wu, H.Y., Chen, Y.J., Lu, C.T., Su, M.G., Hsieh, Y.C., Tsai, C.M., Lin, K.I., Huang, H.D., Lee, T.Y. *et al.* (2014) RegPhos 2.0: an updated resource to explore protein kinase-substrate phosphorylation networks in mammals. *Database : the journal of biological databases and curation*, **2014**, bau034.

43. Keenan, A.B., Torre, D., Lachmann, A., Leong, A.K., Wojciechowicz, M.L., Utti, V., Jagodnik, K.M., Kropiwnicki, E., Wang, Z. and Ma'ayan, A. (2019) ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic acids research*, **47**, W212-W224.

44. Yu, G., Wang, L.-G., Han, Y. and He, Q.-Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, **16**, 284-287.

45. Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.Y. and Wei, L. (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic acids research*, **39**, W316-322.

46. Ono, K., Demchak, B. and Ideker, T. (2014) Cytoscape tools for the web age: D3.js and Cytoscape.js exporters. *F1000Research*, **3**, 143-143.

47. Alberti, S., Gladfelter, A. and Mittag, T. (2019) Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell*, **176**, 419-434.

48. Owen, I. and Shewmaker, F. (2019) The Role of Post-Translational Modifications in the Phase Transitions of Intrinsically Disordered Proteins. *International journal of molecular sciences*, **20**, 5501.

49. Fortini, M.E. (2009) Notch Signaling: The Core Pathway and Its Posttranslational Regulation. *Developmental cell*, **16**, 633-647.

50. Meng, Z., Moroishi, T. and Guan, K.L. (2016) Mechanisms of Hippo pathway regulation. *Genes & development*, **30**, 1-17.

51. Lu, Y., Wu, T., Gutman, O., Lu, H., Zhou, Q., Henis, Y.A.-O. and Luo, K.A.-O. Phase separation of TAZ compartmentalizes the transcription machinery to promote gene expression.

52. Ambadipudi, S., Biernat, J., Riedel, D., Mandelkow, E. and Zweckstetter, M. (2017) Liquid-liquid phase separation of the microtubule-binding repeats of the Alzheimer-related protein Tau. *Nature communications*, **8**, 275.

53. Wegmann, S., Eftekharzadeh, B., Tepper, K., Zoltowska, K.M., Bennett, R.E., Dujardin, S., Laskowski, P.R., MacKenzie, D., Kamath, T., Commins, C. *et al.* (2018) Tau protein liquid-liquid phase separation can initiate tau aggregation. *The EMBO journal*, **37**.

54. Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M. *et al.* (2018) Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*, **175**, 1842-1855.e1816.

55. Zhang, S., Hinde, E., Parkyn Schneider, M., Jans, D.A. and Bogoyevitch, M.A. (2020) Nuclear bodies formed by polyQ-ataxin-1 protein are liquid RNA/protein droplets with tunable dynamics. *Scientific reports*, **10**, 1557.

56. Oswald, F., Winkler, M., Cao, Y., Astrahantseff, K., Bourteele, S., Knöchel, W. and Borggrefe, T. (2005) RBP-Jκ/SHARP Recruits CtIP/CtBP Corepressors To Silence Notch Target Genes. *Molecular and cellular biology*, **25**, 10379.

57. Gui, X., Luo, F., Li, Y., Zhou, H., Qin, Z., Liu, Z., Gu, J., Xie, M., Zhao, K., Dai, B. *et al.* (2019) Structural basis for reversible amyloids of hnRNPA1 elucidates their role in stress granule assembly. *Nature communications*, **10**, 2006.

58. Patel, A., Lee, H.O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M.Y., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T.M. *et al.* (2015) A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell*, **162**, 1066-1077.

59. Kostylev, M.A., Tuttle, M.D., Lee, S., Klein, L.E., Takahashi, H., Cox, T.O., Gunther, E.C., Zilm, K.W. and Strittmatter, S.M. (2018) Liquid and Hydrogel Phases of PrP(C) Linked to Conformation Shifts and Triggered by Alzheimer's Amyloid-beta Oligomers. *Molecular cell*, **72**, 426-443.e412.

60. Mann, J.R., Gleixner, A.M., Mauna, J.C., Gomes, E., DeChellis-Marks, M.R., Needham, P.G., Copley, K.E., Hurtle, B., Portz, B., Pyles, N.J. *et al.* (2019) RNA Binding Antagonizes Neurotoxic Phase Transitions of TDP-43. *Neuron*, **102**, 321-338.e328.

61. Dyson, H.J. Expanding the proteome: disordered and alternatively folded proteins.

62. Gibson, B.A., Doolittle, L.K., Schneider, M.W.G., Jensen, L.E., Gamarra, N., Henry, L., Gerlich, D.W., Redding, S. and Rosen, M.K. (2019) Organization of Chromatin by Intrinsic and Regulated Phase Separation. *Cell*, **179**, 470-484.e421.

63. Zaffagnini, G., Savova, A., Danieli, A., Romanov, J., Tremel, S., Ebner, M., Peterbauer, T., Sztacho, M., Trapannone, R., Tarafder, A.K. *et al*. (2018) p62 filaments capture and present ubiquitinated cargos for autophagy. *The EMBO journal*, **37**.

FIGURE LEGENDS

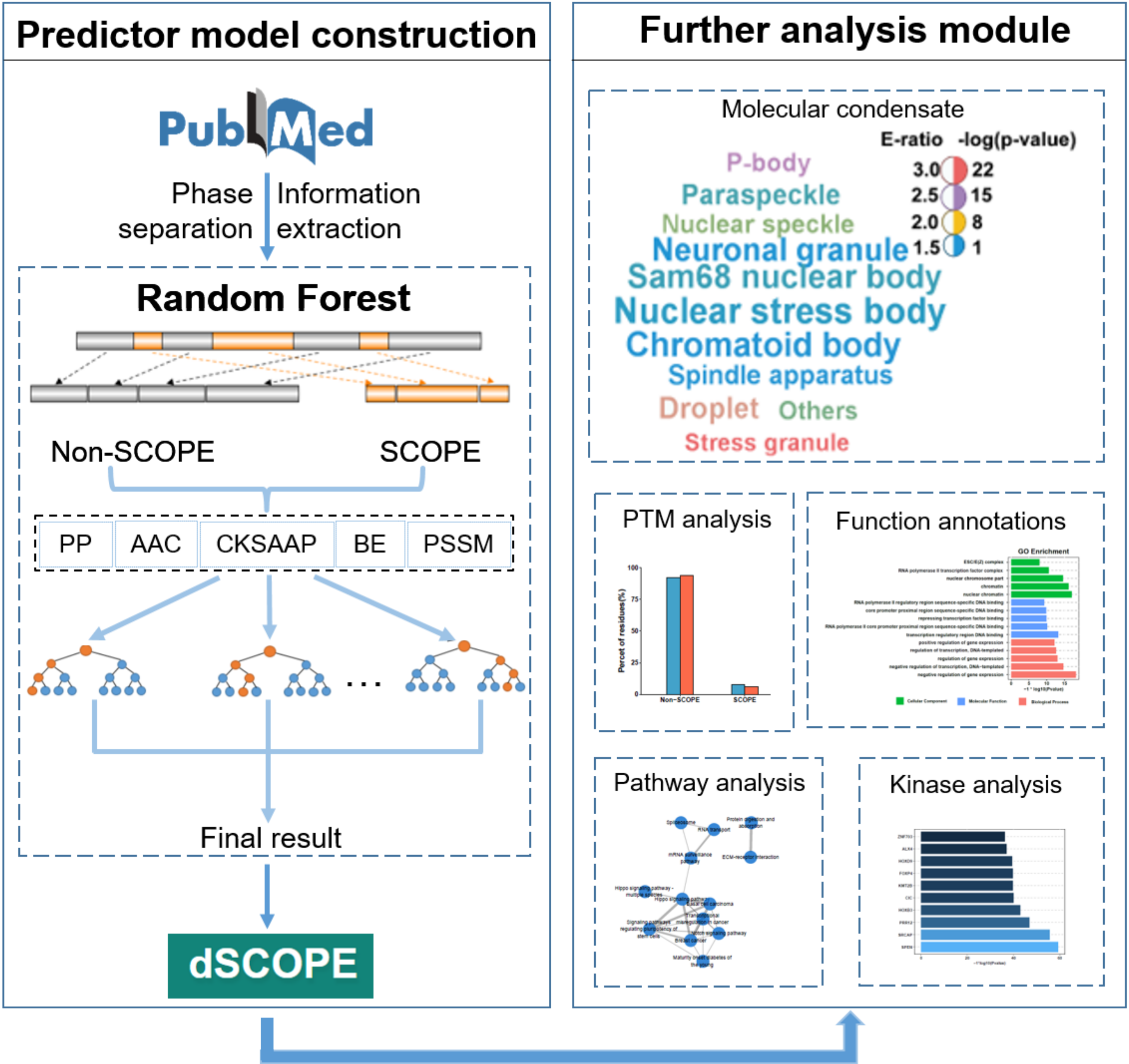
Figure 1. Overview of the model. The key regulatory sequence of phase separation was extracted from PubMed, and then dSCOPE was constructed by using random forest development. Downstream analysis reveals the characteristics of phase separation.

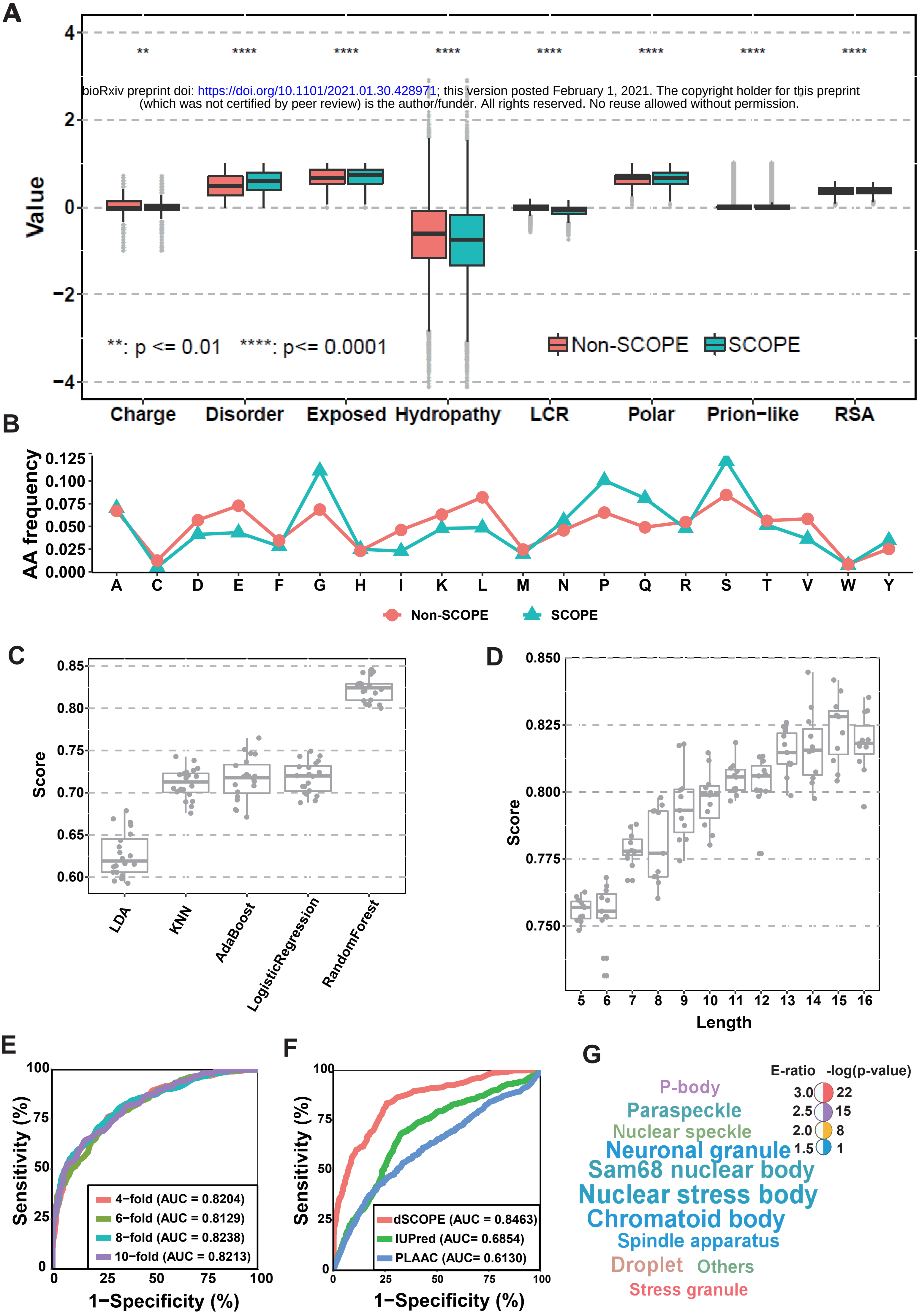
Figure 2. Features of the phase separation region and performance of dSCOPE. (A) Features of the phase separation regions. (B). The frequency of amino acids in different regions. (C) AUC values of different lengths. (D) Use 5-fold cross validation to compare the performance of machine learning algorithms; the median values are 0.8242 (Random forest), 0.7199 (Logistic regression), 0.7177 (AdaBoost), 0.7127 (KNN) and 0.6191 (LDA). (E) The 4-, 6-, 8-, and 10-fold cross validation results in the training dataset. (F) Comparison of the models with other tools. (G) DrLLPS enrichment analysis.

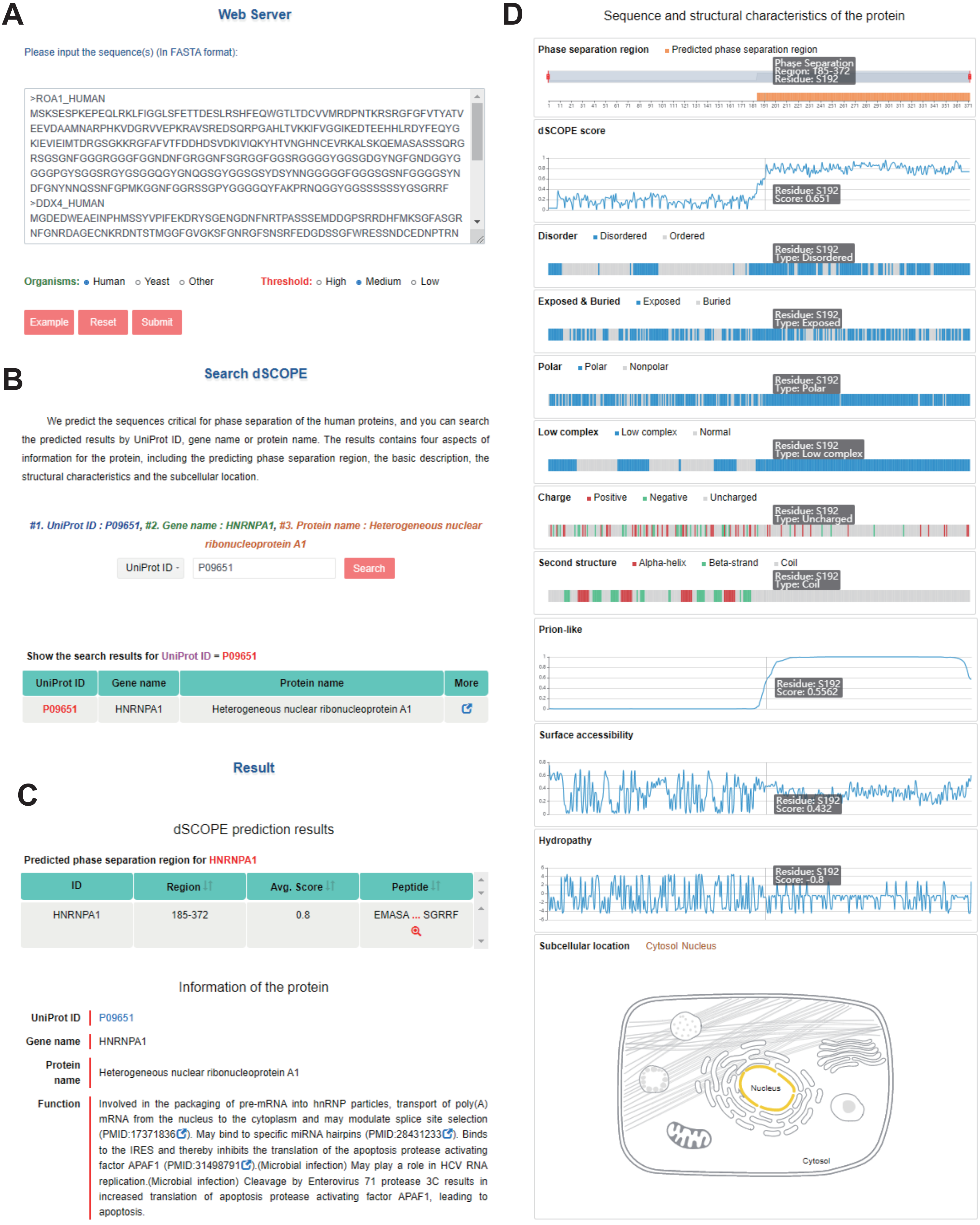
Figure 3. Web server of dSCOPE. (A) The prediction page. (B) The search page. (C) Potential phase separation sequences. (D) The sequence and structure properties of the query protein and subcellular location.

Figure 4. Functional analysis of SCOPEs and potential phase separation proteins. (A) Lysine modification enrichment. mK, modified lysine residue; umK, unmodified lysine residue. (B) Phosphorylation enrichment. pSTY, phosphorylated STY residue; upSTY, unphosphorylated STY residue. (C) GO enrichment. (D) KEGG pathway interaction network. (E) Enrichment of upstream kinases for phosphoproteins in potential phase separation proteins. (F) TF enrichment analysis.

Supplementary Figure S1. Enrichment analysis of tumor mutations. (A) Tumor mutations in SCOPEs confirmed by experiment. (B) Tumor mutations in potential SCOPEs.


Predictor model construction
PubMed
Phase separation
Information extraction
Random Forest
Non-SCOPE
SCOPE
PP
AAC
CKSAAP
BE
PSSM
...
Final result
dSCOPE
Further analysis module
Molecular condensate
E-ratio
-log(p-value)
3.0
22
2.5
15
2.0
8
1.5
1
P-body
Paraspeckle
Nuclear speckle
Neuronal granule
Sam68 nuclear body
Nuclear stress body
Chromatoid body
Spindle apparatus
Droplet
Others
Stress granule
PTM analysis
Percet of residues(%)
100
75
50
25
0
Non-SCOPE
SCOPE
Function annotations
GO Enrichment
ESC/E(Z) complex
RNA polymerase II transcription factor complex
nuclear chromosome part
chromatin
nuclear chromatin
RNA polymerase II regulatory region sequence-specific DNA binding
core promoter proximal region sequence-specific DNA binding
repressing transcription factor binding
RNA polymerase II core promoter proximal region sequence-specific DNA binding
transcription regulatory region DNA binding
positive regulation of gene expression
regulation of transcription, DNA-templated
regulation of gene expression
negative regulation of transcription, DNA-templated
negative regulation of gene expression
0
5
10
15
-1 * log10(Pvalue)
Cellular Component
Molecular Function
Biological Process
Pathway analysis
Spliceosome
RNA transport
Protein digestion and absorption
mRNA surveillance pathway
ECM-receptor interaction
Hippo signaling pathway - multiple species
Hippo signaling pathway
Basal cell carcinoma
Transcriptional misregulation in cancer
Signaling pathways regulating pluripotency of stem cells
Notch signaling pathway
Breast cancer
Maturity onset diabetes of the young
Kinase analysis
ZNF703
ALX4
HOXD9
FOXP4
KMT2B
CIC
HOXB3
PRR12
SRCAP
SPEN
0
20
40
60
-1*log10(Pvalue)


A
Value
4
2
0
−2
−4
**
****
****
****
****
****
****
****
**: p <= 0.01 ****: p<= 0.0001
Non-SCOPE
SCOPE
Charge
Disorder
Exposed
Hydropathy
LCR
Polar
Prion−like
RSA
B
AA frequency
0.125
0.100
0.075
0.050
0.025
0.000
A C D E F G H I K L M N P Q R S T V W Y
Non-SCOPE
SCOPE
C
Score
0.85
0.80
0.75
0.70
0.65
0.60
LDA
KNN
AdaBoost
LogisticRegression
RandomForest
D
Score
0.850
0.825
0.800
0.775
0.750
5 6 7 8 9 10 11 12 13 14 15 16
Length
E
Sensitivity (%)
100
75
50
25
0
0 25 50 75 100
1−Specificity (%)
4−fold (AUC = 0.8204)
6−fold (AUC = 0.8129)
8−fold (AUC = 0.8238)
10−fold (AUC = 0.8213)
F
Sensitivity (%)
100
75
50
25
0
0 25 50 75 100
1−Specificity (%)
dSCOPE (AUC = 0.8463)
IUPred (AUC= 0.6854)
PLAAC (AUC= 0.6130)
G
E-ratio
-log(p-value)
3.0
22
2.5
15
2.0
8
1.5
1
P-body
Paraspeckle
Nuclear speckle
Neuronal granule
Sam68 nuclear body
Nuclear stress body
Chromatoid body
Spindle apparatus
Droplet
Others
Stress granule



Please input the sequence(s) (In FASTA format):

```
>ROA1_HUMAN
MSKSESPKEPEQLRKLFIGGLSFETTDESLRSHFEQWGTLTDCVVMRDPNTKRSRGFGFVTYATV
EEVDAAMNARPHKVDGRVVEPKRAVSREDSQRPGAHLTVKKIFVGGIKEDTEEHHLRDYFEQYG
KIEVIEIMTDRGSGKKRGFAFVTFDDHDSVDKIVIQKYHTVNGHNCEVRKALSKQEMASASSSQRG
RSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGGYG
GGGPGYSGGSRGYGSGGQGYGNQGSGYGGSGSYDSYNNGGGGGFGGGSGSNFGGGGSYN
DFGNYNNQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
>DDX4_HUMAN
MGDEDWEAEINPHMSSYVPIFEKDRYSGENGDNFNRTPASSSEMDDGPSRRDHFMKSGFASGR
NFGNRDAGECNKRDNTSTMGGFGVGKSFGNRGFSNSRFEDGDSSGFWRESSNDCEDNPTRN
```

Organisms: ● Human ○ Yeast ○ Other

Threshold: ○ High ● Medium ○ Low

We predict the sequences critical for phase separation of the human proteins, and you can search the predicted results by UniProt ID, gene name or protein name. The results contains four aspects of information for the protein, including the predicting phase separation region, the basic description, the structural characteristics and the subcellular location.

#1. UniProt ID : P09651, #2. Gene name : HNRNPA1, #3. Protein name : Heterogeneous nuclear ribonucleoprotein A1

UniProt ID: P09651

Show the search results for UniProt ID = P09651

UniProt ID	Gene name	Protein name	More
P09651	HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1	

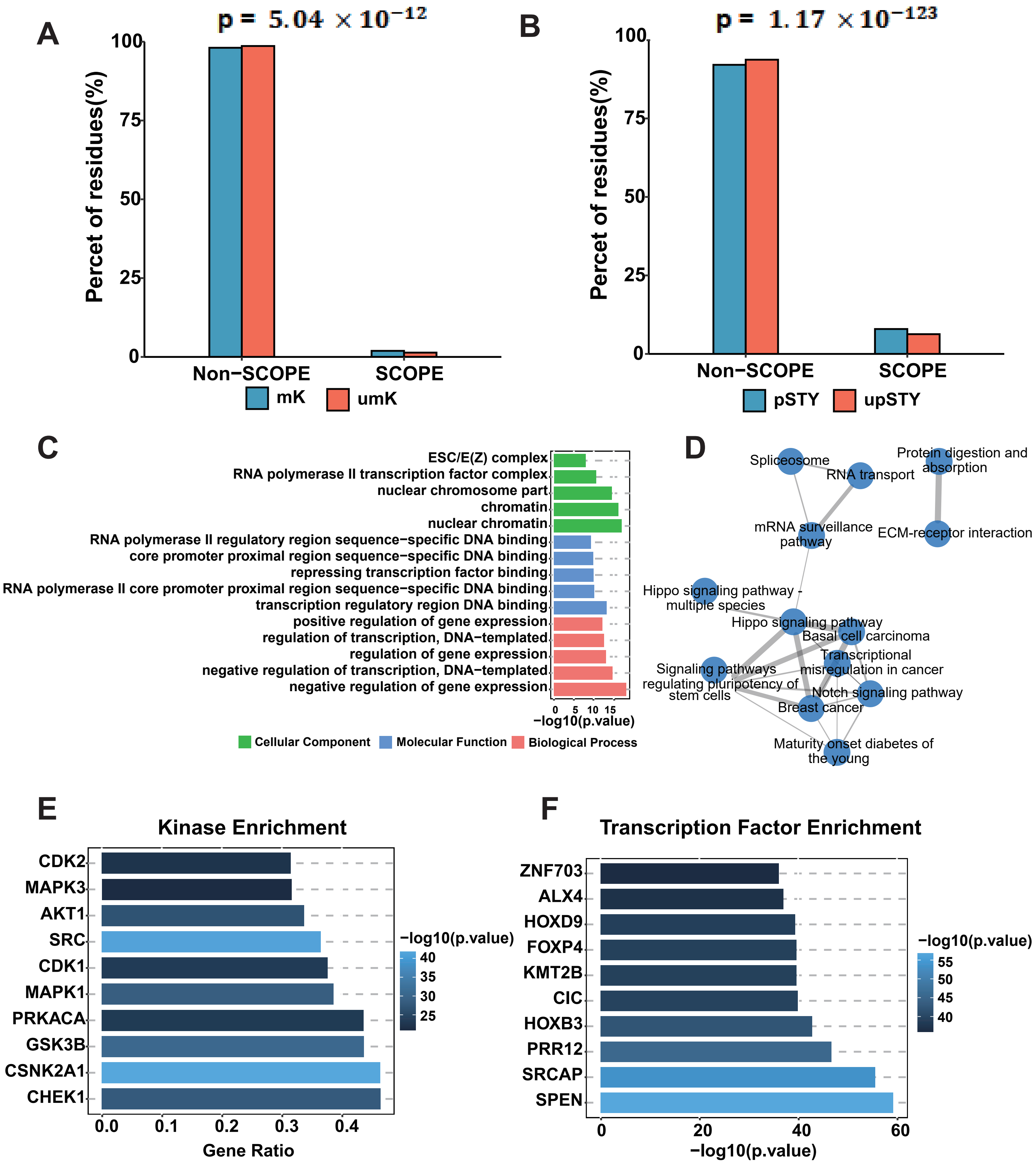
dSCOPE prediction results

Predicted phase separation region for HNRNPA1

ID	Region	Avg. Score	Peptide
HNRNPA1	185-372	0.8	EMASA ... SGRRF

Information of the protein

UniProt ID	P09651
Gene name	HNRNPA1
Protein name	Heterogeneous nuclear ribonucleoprotein A1
Function	Involved in the packaging of pre-mRNA into hnRNP particles, transport of poly(A) mRNA from the nucleus to the cytoplasm and may modulate splice site selection (PMID:17371836). May bind to specific miRNA hairpins (PMID:28431233). Binds to the IRES and thereby inhibits the translation of the apoptosis protease activating factor APAF1 (PMID:31498791).(Microbial infection) May play a role in HCV RNA replication.(Microbial infection) Cleavage by Enterovirus 71 protease 3C results in increased translation of apoptosis protease activating factor APAF1, leading to apoptosis.


A
p = 5.04 × 10⁻¹²
Percent of residues(%)
100
75
50
25
0
Non−SCOPE
SCOPE
mK
umK
B
p = 1.17 × 10⁻¹²³
Percent of residues(%)
100
75
50
25
0
Non−SCOPE
SCOPE
pSTY
upSTY
C
ESC/E(Z) complex
RNA polymerase II transcription factor complex
nuclear chromosome part
chromatin
nuclear chromatin
RNA polymerase II regulatory region sequence−specific DNA binding
core promoter proximal region sequence−specific DNA binding
repressing transcription factor binding
RNA polymerase II core promoter proximal region sequence−specific DNA binding
transcription regulatory region DNA binding
positive regulation of gene expression
regulation of transcription, DNA−templated
regulation of gene expression
negative regulation of transcription, DNA−templated
negative regulation of gene expression
0 5 10 15
−log10(p.value)
Cellular Component
Molecular Function
Biological Process
D
Spliceosome
RNA transport
Protein digestion and absorption
mRNA surveillance pathway
ECM-receptor interaction
Hippo signaling pathway - multiple species
Hippo signaling pathway
Basal cell carcinoma
Transcriptional misregulation in cancer
Signaling pathways regulating pluripotency of stem cells
Notch signaling pathway
Breast cancer
Maturity onset diabetes of the young
E
Kinase Enrichment
CDK2
MAPK3
AKT1
SRC
CDK1
MAPK1
PRKACA
GSK3B
CSNK2A1
CHEK1
0.0
0.1
0.2
0.3
0.4
Gene Ratio
−log10(p.value)
40
35
30
25
F
Transcription Factor Enrichment
ZNF703
ALX4
HOXD9
FOXP4
KMT2B
CIC
HOXB3
PRR12
SRCAP
SPEN
0
20
40
60
−log10(p.value)
−log10(p.value)
55
50
45
40