Enrichment of charge-absent regions in phase separated proteins

Sonia T. Nicolaou[1,2], Chandra S. Verma[2,3,4] and Jim Warwicker[1,*]

[1]School of Biological Sciences, Faculty of Biology, Medicine and Health, Manchester Institute of Biotechnology, University of Manchester, Manchester M1 7DN, UK [2]Bioinformatics Institute, Agency for Science, Technology, and Research (A*STAR) [3]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551 [4]Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543

* Corresponding author

jim.warwicker@manchester.ac.uk (JW)

Abstract

Many studies focus on the relationship between protein charge and liquid-liquid phase separation (LLPS), generally finding that a large degree of charge neutralisation is involved for condensate formation. Here, sequences within human proteins that lack the charge-bearing residues Asp, Glu, Lys, and Arg (termed charge-absent) are analysed alongside annotation for involvement in LLPS. Scaffold proteins, central to condensate formation, on average possess longer charge-absent regions than those not key for LLPS. Charge-absent regions tend to have relatively high hydropathy scores. Overall, they are enriched in Ala, Gly, Pro, and Ser with more specific groupings evident when the subset is clustered by amino acid composition. For several proteins, segments with charge-absent regions have been identified as modulators of LLPS. It is hypothesised that for at least some of the charge-absent regions, a lack of charged group desolvation energy, together with a relatively hydrophobic sequence composition, may facilitate condensation through homomeric interactions. If this is the case, it should be relatively easy to modulate through incorporation of charge through engineering, potentially including pH-sensing.

Introduction

Cells organise their densely packed chemical space into compartments (organelles) as a way of regulating complex biochemical reactions. Typically, cellular components are encapsulated in membranes. Alternatively, these cellular components form membraneless organelles (MLOs), through liquid-liquid phase separation (LLPS), such as stress granules, Cajal bodies, P-bodies, PML bodies, and nuclear speckles. Recent studies have linked formation of MLOs to intrinsically disordered proteins (IDPs) [1], proteins that lack distinct secondary or tertiary structures and instead exist as dynamic conformational ensembles [2-6]. IDPs are composed of amino acid (AA) sequences with lower complexity since they are enriched in "disorder promoting" residues such as Ala, Arg, Gly, Gln, Ser, Glu, Lys and Pro and are depleted in hydrophobic and aromatic residues [4, 7, 8]. Low complexity domains (LCDs) are sometimes involved in LLPS [9-11]. Some of these low complexity sequences are called prion-like domains (PLDs) due to sequence similarities with yeast prion proteins [12]. The high net charge and low mean hydrophobicity of IDP sequences are an important contributor to their lack of compact structure, due to charge-charge repulsion [13]. The charge-hydrophobicity relationship of IDPs is clearly demonstrated in the charge-hydropathy plot, also known as the Uversky plot [5, 14]. Several disorder prediction servers utilise this principle to describe IDPs [15-17]. A diagram-of-states to classify the predicted conformational properties of IDPs based on their charge-hydropathy relationship was initially created by Mao and co-workers and was later improved by Das and Pappu [18, 19]. The two-dimensional (2D) diagram-of-states has fraction of positively charged residues on the x-axis (f_+) and fraction of negatively charged residues (f_-) on the y-axis and is divided into 5 regions based on their fraction of charged residues (FCR) and net charge per residue (NCPR) scores [19, 20].

IDPs adopt an ensemble of conformations, and due to their flexible nature, exposed AA sequence, and binding promiscuity, have an increased transient interaction network thought to enhance their phase separation properties [21-24]. Multivalent interactions between IDPs, IDPs and folded proteins, IDPs and nucleic acids, and nucleic acids-nucleic acids, including hydrogen bond, electrostatic, cation-π, π-π, dipole-dipole and hydrophobic interactions are the driving force of LLPS [25], along with modulations in the solution/environment of the system such as temperature, hydrostatic pressure, pH and ionic strength [26-28]. Under physiologic conditions, some IDPs undergo LLPS as a way of regulating important biological processes. However, changes in the sequence dependent phase behaviour of IDPs such as mutations and post-translational modifications (PTMs), can sometimes lead to the formation of pathological aggregates [21, 29]. Protein aggregation is the hallmark of many neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD), Alzheimer's disease (AD), and Parkinson's (PD). TAR DNA-binding protein 43 (TDP-43), fused in sarcoma (FUS), tau, and α-synuclein are examples of IDPs found in patients with ALS, FTD, AD, and PD, respectively [29-34]. Under physiologic conditions, these proteins exist in a dispersed state and can undergo reversible transition to phase separated liquid droplets. However, under stress they form irreversible pathological aggregates.

While recent studies focused on the importance of charge in IDPs and its implications in LLPs, IDPs with regions of little or no charge also exist in the human proteome, such as in the carboxy-terminus of TDP-43. TDP-43 binds to DNA and RNA through two RNA recognition motifs (RRM1 and RRM2) [35]. It also contains a disordered, Gly-rich, low complexity (PLD) carboxy-terminus. This disordered terminal region contains a short, conserved region (CR) between residues 316-346, and can promote phase separation of TDP-43 into stress granules and cytosolic aggregates [12, 36]. The CR lacks charged AAs, but is enriched in polar ones (Gln, Asn, Ser and Gly) and contains hydrophobic residues Tyr, Phe and ten evolutionarily conserved Met residues [12, 37]. The CR forms an α-helical structure that is stabilised by intermolecular contacts between CR helices of adjacent TDP-43 molecules and is associated with condensation [36, 38].

The function of IDPs can generally be established from their amino acid sequence, often enriched in charged and polar residues. However, intrinsically disordered regions with very little or no residues expected to be charged at neutral pH pose the question of why such regions exist in the human proteome and if they are functionally significant. One in five AAs are charged (Asp, Glu, Lys, Arg) and therefore it is unlikely that long polypeptide segments without charge would exist by chance. Studies by Bremer et al. suggest that electroneutral NCPR values promote LLPS, whereas larger NCPR values (unbalanced charges) suppress LLPS [39]. The diagram-of-states [19] uses f_+, f_-, FCR and NCPR to describe the conformational preferences of IDPs. The current study focuses on the (0,0) point of the diagram to investigate regions of proteins with zero fraction of both positively and negatively charged residues (termed charge-absent regions). It is found that the critical scaffold proteins in phase separated condensates are enriched in charge-absent regions, and that these regions are as hydrophobic as structured proteins despite mostly being disordered and depleted in larger hydrophobic amino acids.

Methods

Dataset

The human proteome, with one protein sequence per gene (proteome ID UP000005640, 20588 sequences) was acquired from the UniProt database [40]. Proteins with transmembrane (TM) domains were excluded from the dataset since the study is focused on aqueous soluble proteins. Proteins were assumed to contain TM segments if their hydropathy score was greater than 1.6 on the Kyte & Doolittle [41] scale, for any 21 AA segment [42], which would be representative of a typical TM helix. After this filter, 623 proteins containing 51 AA stretches of zero charge (charge-absent, no K/R/D/E residues) were identified in the human proteome.

Prediction of intrinsically disordered segments was made with IUpred2A [43, 44]. Any single amino acid with score > 0.5 is predicted to be disordered. The AA composition of each 51 AA charge-absent segment was found using the web tool iFeature [45]. Proteins involved in LLPS were obtained from the DrLLPS database [46] and were matched with the subset of proteins containing charge-absent regions of at least 51 amino acids. The proteins were then divided into 3 categories: scaffold, client, and regulator.

The dataset for PrLDs used was acquired from March et al. [47]. Gene Ontology (GO) analysis for a given set of genes was performed using Princeton GO term finder [48-50], and Revigo [51] was used to visualise the results. An additional dataset for benchmarking hydrophobicity of the charge-absent subset was obtained from previous work [52], with human protein structures from the Protein Data Bank (PDB) [53], non-redundant at 25% sequence identity.

Clustering

The AA compositions of the 623 51 AA charge-absent sequences (one per protein) were normalised by calculating their z-scores i.e. the number of standard deviations away from the mean (over the complete subset) for the composition of each AA type in each 51 AA window. Agglomerative hierarchical clustering of the normalised AA compositions of the sequences was performed using the Ward variance minimization algorithm, and the Euclidean distance function as the distance metric between vectors. During agglomerative hierarchical clustering each object (protein sequence) constitutes its own cluster, and then the clusters closest together (clusters with the shortest Euclidean distance between them) are successively merged [54-56]. The resulting output is a dendrogram with the Euclidean distance between clusters on one axis and the clusters on the other. SciPy was used to create the dendrogram and colour-coded each group of nodes whose linkage was less than 70% of the maximum linkage. The dendrogram was then visually inspected based on the SciPy colour-coding and 11 clusters were identified [57].

Results and discussion

Scaffold LLPS proteins are enriched in charge-absent regions

Previous charge distribution analysis showed that IDRs of proteins involved in LLPS are biased towards neutral net charge [17]. Human proteins in our dataset were divided into scaffold, client, and regulator proteins, or unannotated (null), based on DrLLPS, an online database that lists proteins involved in LLPS [46]. Scaffold proteins are directly involved in LLPS, client proteins are recruited into MLOs but are not essential for LLPS, and regulators modulate LLPS but are not localised in the droplets [58]. The null subset is that part of the human proteome not annotated in DrLLPS, and without predicted TM segments. Cumulative plots of the maximum charge-absent lengths were made for proteins with the associations scaffold, regulator, client in DrLLPS, as well as other human proteins (lacking a predicted TM segment and labelled as null in Fig 1). Interestingly the regulator, client, and null plots almost overlay, whilst that of proteins labelled as scaffold is enriched in longer charge-absent regions. The maximum separation, in the cumulative plots, between scaffold proteins and the other subsets is around a charge-absent window of 50 amino acids (Fig 1). It is believed that net charge is an important regulator of IDR conformation and global dimensions in solution [5, 17, 19, 20, 59, 60] and this observation is consistent with a model in which condensation of some human scaffold proteins is aided by relatively long intrinsically disordered regions bearing zero charge at neutral pH. For more detailed study, a length of contiguous charge-absent sequence was set to 51 amino acids. This is a balance of seeking a value toward the upper end charge-absent region length and maintaining a difference between scaffold and other subsets (Fig 1).

Fig 1. **Scaffold proteins are enriched in longer charge-absent regions.** Subsets of human proteins, according to annotation in DrLLPS (client N=80, regulator N=278, client N=2651), are displayed in cumulative distributions of maximum length of charge-absent region in each protein. The null subset covers human proteins without DrLLPS annotation (and lacking a predicted TM segment, N=10778). The plots for regulator, client, and null subsets almost overlay.

The first charge-absent segment, for each of the 623 proteins with at least one 51 AA charge-absent segment was submitted to the IUpred2A disorder predictor [43, 44]. The average IUPred2A score for the 623 51 charge-absent segments was 0.532, with 52.5% of amino acids having disorder prediction (scores >= 0.5). Moreover, 482 out of the 623 51 regions had at least one AA with score >= 0.5. The majority of the charge-absent segments were predicted as IDRs, and amongst those that were not, are structurally condensed proteins such as Nup98 and keratins.

Further, to identify that scaffold proteins are enriched in charge-absent regions (Fig 1), we focus on the 51 AA charge-absent segments in proteins without a predicted TM segment. Comparing such proteins with the total number of annotated human proteins in DrLLPS, occurrence with 51 AA charge-absent regions is 26% (scaffold), 8% (regulator), and 4% (client), supporting potential roles in phase separation. Gene Ontology analysis coupled to visualisation with Revigo [51] revealed that proteins containing charge-absent segments are enriched in transcriptional processes and in particular DNA-templated transcription with nuclear localisation (S1 and S2 Figs).

Charge-absent regions in the human proteome are as hydrophobic as structured proteins

To provide context on the degree of hydrophobicity (Kyte-Doolittle hydropathy, KD, here plotted on a 0 to 1 scale, KD-scaled) of the charge-absent regions, comparison with structured proteins was made. For this purpose, the KD-scaled values were also calculated for each protein in a dataset of 1627 human proteins from PDB structures (non-redundant at 25%

sequence identity) [52]. The distribution of KD-scaled values for the charge-absent window distribution is more hydrophobic, on average, than that of the structured protein set (Fig 2). This implies propensity for a condensed phase, perhaps even approaching that of folded proteins, at least in terms of average packing in what is presumably a set of interchanging conformations. These results are biased by selection for the absence of charged amino acids, and it is instructive to also examine the compositions of other amino acids, relative to the structured protein set (Fig 3). Overall, the charge-absent regions are depleted in AAs with aromatic and large hydrophobic sidechains, and are enriched in smaller AAs, in particular Ala, Gly, Pro and Ser, which combine to make up 54% of the AA composition of these 51 residue segments (Fig 3). Calculation of the Shannon entropy for complexity of AA composition [42] revealed that the charge-absent segments are predominantly low complexity regions (LCRs, average entropy 2.9) compared with structured proteins (average 4.0), although the absence of charged AAs will contribute to this result.

Fig 2. **Charge-absent windows have high Kyte-Doolittle hydropathy.** Scaled Kyte-Doolittle hydropathy values are shown in histograms for the 623 subset of charge-absent windows of more than 51 AA length (the first occurring such window in a protein of the human proteome), in purple, and for the human 1627 PDB set (orange).

Fig 3. **Charge-absent regions are enriched in small amino acids, with a balance of small over large hydrophobics.** Percentage AA compositions are compared for the set of charge-absent windows (more than 51 AA, the first such occurrence in each human protein, without predicted TM), in purple, and the 1627 PDB set of structured human proteins (orange).

Charge-absent regions clustered by amino acid composition

Having established that charge-absent regions in the human proteome are enriched in proteins associated with LLPS and that their hydropathy values are consistent with condensation, we next examined clustering by sequence composition and mapping to specific functions, where these are characterised. The AA composition of each 51-residue segment was calculated using the web tool iFeature [45]. IDRs are enriched in LCRs which in some cases promote LLPS. Different combinations of amino acids were examined, alongside potential links according to function and, more specifically, possible involvement in phase separation. The charge-absent sequences were clustered by amino acid composition (not sequence homology) z-score as described in the methods section. Eleven clusters were identified at the specified cut-off (Fig 4).

Fig 4. **Clustermap of 623 charge-absent 51 AA protein sequences. The sequence segments were normalised by z-score and were clustered by AA composition.** Subsets of human proteins are colour coded according to annotation in DrLLPS with Scaffold proteins shown in gold, regulator proteins shown in green and client proteins shown in blue. The null subset covers human proteins without DrLLPS annotation and is coloured in red. The clustermap was also colour-coded by cluster.

Prion-like domains are enriched in uncharged polar amino acids as well as Gly and are often found in RNA binding proteins involved in neurodegenerative diseases caused by protein aggregation such as ALS [61]. Within a list of human proteins with PLDs [47], 68% contained at least one 51 segment absent of charged residues. Proteins with PLDs were found throughout the 11 clusters, apart from Cluster 7. Generally, PLDs have been reported to drive phase separation and, alternatively, to modulate protein phase behaviour [62]. Table 1 gives examples of proteins in each of the clusters, along with amino acid enrichment associated with the cluster, and numbers that have been associated with PLDs. References discussing biological background are given for each example. Some examples are highlighted in the following text, with specific context for charge-absent regions. From combining annotation of a protein as involved in LLPS, with at least one relatively long (> 50 AA) charge-absent region, it does not follow that a charge-absent region necessarily has a central role in phase separation. Even where further characterisation has been reported, this may not be the complete picture. However, there are several examples where charge-absent regions have been directly implicated in LLPS.

Table 1. Examples of proteins containing at least one 51 amino acid charge absent region from each cluster. In some of the clusters these proteins have been implicated in phase separation either as scaffolds, clients, or regulators.

Cluster	AA enrichment	Cluster Size	Proteins with Prion-like domains (PLDs)	Examples
C0	S, L	200	37	YAP1[63], UBQLN2[64, 65], RBM14[66]
C1	T, V	55	7	BRD3[67], BRD4[68, 69], YY1[70]
C2	M, N	94	41	TDP-43[12, 36, 38, 71], CREBP[72], p300[72], Nup54, Nup58, Nup62[73], Nup98[74, 75], Nup153
C3	Q	41	20	TBP[76, 77], ZNF207[78]
C4	G	28	12	Loricrin[79]
C5	H	28	3	Nufip2[80]
C6	W	35	6	TIA1[81], POLR2A[82], LGALS3[83]
C7	C	12	None	None
C8	Y	39	22	FUS[10, 11, 84, 85], EWSR1[84]
C9	A	38	3	HOXA13[76]

C10	P	53	10	WASL[86]

Bromodomain containing protein 4 (BRD4) is part of cluster 1 and is listed as a scaffold protein in DrLLPS. The N-terminus of BRD4 includes two acetyl-lysine binding bromodomains, separated by an intrinsically disordered region that includes the identified 51 AA segment without charge (232-283) [69]. BRD4 is involved in gene transcription, DNA replication and repair and a short isoform (BRD4S) forms condensates within the cell nuclei [68, 69]. LLPS of BRD4 in the nucleus is mediated by binding of its IDRs and bromodomains to acetylated chromatin and DNA [69], and is diminished upon phosphorylation.

Nucleoporins (Nups) are a family of around 30 intrinsically disordered proteins that make up the nuclear pore complex (NPC). They are in cluster 2 and are known for being rich in Phe and Gly residues, rather than in Met and Asn that is a more common feature of cluster 2 (Table 1). This also indicates that although aromatic AAs are somewhat depleted across charge-absent regions, there are specific families that buck this trend. Nup54, Nup58, Nup62, Nup98 and Nup153 fall in cluster 2. Nups contain multiple Phe/Gly-rich repeats of different kinds throughout their sequence [74] that can promote phase separation through inter-repeat interactions [75, 87], and control the transportation of cargo in and out of the nucleus [74]. Since the Phe/Gly-rich regions typically coincide with charge-absent segments, a connection between phase separation and lack of charge can be inferred.

Cluster 2 also includes transcription factors such as TDP-43, CREB-binding protein (CREBBP or CBP), p300, and Forkhead box proteins 1 and 2 (FOXA1, FOXA2). TDP-43 has a methioninerich charge-absent region (centred on 294-345) that overlaps with the conserved helical region (316-346) known to promote LLPS of TDP-43 [12, 36].

Within the glutamine-rich proteins contained in cluster 3, TATA-box binding protein (TBP) is a transcription factor regulating RNA Pol II activity, with a Gln-rich charge-absent region (including 58 to 109), coincident with an IDR that is reported to drive phase separation. Repeat expansions of the polyQ tail can alter the phase separation capacity of TBP, which in turn can contribute to disease pathogenesis [76, 77]. TBP regulates RNA Pol II transcription of eukaryotes. Moreover, T-cell restriction intracellular antigen 1 (TIA1) is an RNA binding protein residing in cluster 6 that includes overlap of a charge-absent region and a Pro-rich LCD that mediates LLPS, and in which Pro to Leu mutations alter droplet morphology and are disease-associated [81].

It is known that phase separation can also be mediated by combination of charge-absent and other regions, as in the case of FUS, an RNA binding protein in cluster 8. The charge-absent region is located in a PLD rich in Gly, Ser, and Tyr [88, 89]. Tyrosines in the PLD can assist phase separation through cation-π interactions with Arg residues from neighbouring RNA-binding domain of FUS [84]. In addition, Gln residues in FUS also contribute to phase separation, through hydrogen bonding interactions [85]. Furthermore, EWS RNA binding protein 1 (EWSR1) is part of the FUS family of proteins and has similar characteristics to FUS, with a charge-absent region coincident with a PLD. Similar to FUS, EWSR1 phase separates when the PLD interacts with the RBD through cation-π interactions [84]. Collectively, these examples demonstrate that charge-absent regions play a role in phase separation either through homomeric or heteromeric interactions, possibly with charged regions (for example with cation-π interactions).

Charge modulation of charge-absent regions

It is hypothesised that charge-absent regions (or at least a subset) may be associated with phase separation, by virtue of the absence of desolvation penalty. In this case, the core of some condensates would be devoid of net charge through absence of interacting charges, as opposed for example to those where RNA and RNA-binding proteins combine. They would therefore involve either homomeric or heteromeric interactions, but in either case partner proteins should also contain charge-absent regions. As noted previously, it is also possible that charge-absent regions could supply aromatic groups for partnering with basic residues in cation-π interactions. If charge-absent segments do encourage condensation, then two factors that could alter charge and thus modulate phase separation are phosphorylation and His protonation (at mild acidic pH).

For the 623 human protein subset, the numbers of phosphorylation sites recorded in UniProt that locate to any region within or without a charge-absent 51 AA segment were calculated. Of 2898 total phosphorylation sites from UniProt, only 9 (0.3%) are within 51 AA long charge-absent regions, leaving 99.7% without. For reference, the split of all AAs between within and without charge-absent segments is 4% and 96%, respectively. It is apparent that phosphorylation is depleted in charge-absent regions, despite the propensity in general for phosphorylation within IDRs [90]. Combined Ser and Thr residues are split at 6% within charge-absent regions and 94% without, similar to overall AA numbers, so there is no lack of phosphorylation targets in charge-absent regions. There may however be a lack of defined motifs for protein kinases, since these often involve the charged amino acids that are absent [91]. Whatever the reason, it appears that there exists limited scope for modulation of charge-absent region function by phosphorylation.

Although the normal His sidechain pKa is around 6.3, it is possible that mild acidic pH could lead to protonation (full or partial) and thus exert some influence on charge-absent regions. Percentages of His within and without charge-absent regions (51 AA windows) of the 623 protein subset are 4% and 96%, respectively, matching the overall amino acid split. It is therefore possible that His protonation could be a modulating factor for charge-absent region function, depending on the existence of a mild acidic pH environment. Although the overall His content in charge-absent regions is not enriched, motifs of > 5 consecutive His are more prevalent (11 in 623, versus 70 in the entire human proteome). His tracts with metal ion coordination ability are present in some human transcription factors [92], and separately metal ion-induced condensation has been demonstrated for proteins with engineered hexa-His tags [93]. It is not clear at this stage whether metal ion-dependent association is important for the function of some proteins in the charge-absent region subset, or if it were, how that might couple to the central property (absence of charged residues). One issue is the location of poly-His tracts, for example in BRD4, a hexa-His motif lies at one end of a charge-absent region, bordering a highly charged neighbouring segment. More generally His-rich domains (HRDs) have been associated with targeting of proteins and, as part of LCDs, LLPS activity [94]. An HRD in the YY1 transcription factor, within a region of 33 AAs lacking charged residues, contributes to LLPS and gene expression [70]. The HRD of YY1 is bounded by regions rich in negative charge, raising the possibility of synergistic charge interactions between the neighbouring domains.

Conclusion

This study focused on protein regions absent of charge, predicted to be largely disordered, and representing a single point in the diagram of fractional positive versus fractional negative charge [19]. Human proteins annotated as scaffold proteins in protein condensates are enriched in charge-absent regions. It is unclear whether charge-absent regions directly drive phase separation, as might be hypothesised from the reduced desolvation energy required, or whether they interact with other regions to promote LLPS. We speculate that charge-absent regions have a propensity to phase separate into relatively compact, but presumably dynamic, structures due to an amino acid composition that in many cases has Kyte-Doolittle hydropathy comparable to that of folded and structured proteins, and without the energetic cost for partially desolvating charged groups. Charge-absent regions are enriched in small hydrophobic AAs and depleted in some hydrophobic AAs with larger sidechains. Amino acid repeats are relatively common in charge-absent regions, compared with charge-containing windows (S3 Fig). We propose that the lack of charge is not solely a result of enrichment of certain amino acids and repeats but is (at least in some cases) key for function. This is supported by the well-characterised example from our charge-absent region dataset, TDP-43. Homomeric interactions of a partly helical segment, that overlaps the charge-absent region, modulate phase separation. Importantly, single site mutations that either alter helical propensity or add charge are associated with disease [36, 38]. It is suggested that charge-absent regions can contribute to the formation of protein condensates by virtue simply of a lack of desolvation penalty for groups bearing net charge. Many differently charged biological systems are known to undergo LLPS, including those with positively and negatively charged domains from within one molecule or between molecules. Condensation with charge-absent regions, where it occurs, would therefore represent just a small subset of phase separating systems, possibly with particular properties in terms of packing (close since there is no charge desolvation) and interchanging conformations (since they are enriched for AAs with small hydrophobic sidechains). Predicting whether a protein will undergo phase separation from sequence is challenging due to the complexity associated with LLPS such as recruitment of other molecules, including ions, proteins, and nucleic acids. While many studies focus on the implications of charge for LLPS, this study reports on regions bearing no charge. The hypothesised role for some charge-absent regions can be investigated by experiments that monitor condensation when charge is added, through mutation, via His protonation and pH-dependence, or with phosphorylation.

Acknowledgments

The authors would like to acknowledge the University of Manchester and Agency for Science, Technology and Research (A*STAR) for funding as well as the National Supercomputing Centre (NSCC), Singapore for computing facilities.

Funding

This Computational work in JW's group is supported by grant BB/V006592/1 from the UK Biotechnology and Biological Sciences Research Council

References

1. Dignon GL, Best RB, Mittal J. Biomolecular Phase Separation: From Molecular Driving Forces to Macroscopic Properties. Annu Rev Phys Chem. 2020;71:53-75. Epub 2020/04/22. doi: 10.1146/annurev-physchem-071819-113553. PubMed PMID: 32312191; PubMed Central PMCID: PMCPMC7469089.

2. Wright PE, Dyson HJ. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. Journal of Molecular Biology. 1999;293(2):321-31. doi: 10.1006/jmbi.1999.3110.

3. Tompa P. Intrinsically unstructured proteins. Trends in Biochemical Sciences: Elsevier Current Trends; 2002. p. 527-33.

4. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. Journal of Molecular Graphics and Modelling. 2001;19(1):26-59. doi: 10.1016/S1093-3263(00)00138-8.

5. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins: Structure, Function, and Genetics. 2000;41(3):415-27. doi: 10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7.

6. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nature Reviews Molecular Cell Biology: Nature Publishing Group; 2005. p. 197-208.

7. Williams RM, Obradovi Z, Mathura V, Braun W, Garner EC, Young J, et al. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2001:89-100.

8. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins: Structure, Function and Genetics. 2001;42(1):38-48. doi: 10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3.

9. Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell. 2015;163(1):123-33. Epub 2015/09/26. doi: 10.1016/j.cell.2015.09.015. PubMed PMID: 26406374; PubMed Central PMCID: PMCPMC5149108.

10. Kato M, Han TW, Xie S, Shi K, Du X, Wu LC, et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. Cell. 2012;149(4):753-67.

11. Murray DT, Kato M, Lin Y, Thurber KR, Hung I, McKnight SL, et al. Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. Cell. 2017;171(3):615-27 e16. Epub 2017/09/26. doi: 10.1016/j.cell.2017.08.048. PubMed PMID: 28942918; PubMed Central PMCID: PMCPMC5650524.

12. Tziortzouda P, Van Den Bosch L, Hirth F. Triad of TDP43 control in neurodegeneration: autoregulation, localization and aggregation. Nat Rev Neurosci. 2021;22(4):197-208. Epub 2021/03/04. doi: 10.1038/s41583-021-00431-1. PubMed PMID: 33654312.

13. Habchi J, Tompa P, Longhi S, Uversky VN. Introducing protein intrinsic disorder. Chemical Reviews2014. p. 6561-88.

14. Uversky VN. Natively unfolded proteins: A point where biology waits for physics. Protein Science. 2002;11(4):739-56. doi: 10.1110/ps.4210102.

15. Holehouse AS, Das RK, Ahad JN, Richardson MOG, Pappu RV. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. 2017. doi: 10.1016/j.bpj.2016.11.3200.

16. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, et al. FoldIndex©: A simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics. 2005;21(16):3435-8. doi: 10.1093/bioinformatics/bti537.

17. Nicolaou ST, Hebditch M, Jonathan OJ, Verma CS, Warwicker J. PhosIDP: a web tool to visualize the location of phosphorylation sites in disordered regions. Sci Rep. 2021;11(1):9930. Epub 2021/05/13. doi: 10.1038/s41598-021-88992-0. PubMed PMID: 33976270; PubMed Central PMCID: PMCPMC8113260.

18. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(18):8183-8. doi: 10.1073/pnas.0911107107.

19. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proceedings of the National Academy of Sciences. 2013;110(33):13392-7. doi: 10.1073/pnas.1304749110.

20. Das RK, Ruff KM, Pappu RV. Relating sequence encoded information to form and function of intrinsically disordered proteins. Current Opinion in Structural Biology: Elsevier Current Trends; 2015. p. 102-12.

21. Statt A, Casademunt H, Brangwynne CP, Panagiotopoulos AZ. Model for disordered proteins with strongly sequence-dependent liquid phase behavior. The Journal of Chemical Physics. 2020;152(7):075101-. doi: 10.1063/1.5141095.

22. Zhou HX, Nguemaha V, Mazarakos K, Qin S. Why Do Disordered and Structured Proteins Behave Differently in Phase Separation? Trends Biochem Sci. 2018;43(7):499-516. Epub 2018/05/03. doi: 10.1016/j.tibs.2018.03.007. PubMed PMID: 29716768; PubMed Central PMCID: PMCPMC6014895.

23. Bah A, Forman-Kay JD. Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. Journal of Biological Chemistry. 2016;291(13):6696-705. doi: 10.1074/JBC.R115.695056.

24. Mollica L, Bessa LM, Hanoulle X, Jensen MR, Blackledge M, Schneider R. Binding mechanisms of intrinsically disordered proteins: Theory, simulation, and experiment. Frontiers in Molecular Biosciences: Frontiers Media SA; 2016. p. 52-.

25. Luo YY, Wu JJ, Li YM. Regulation of liquid–liquid phase separation with focus on post-translational modifications. Chemical Communications. 2021;57(98):13275-87. doi: 10.1039/D1CC05266G.

26. Vernon RM, Chong PA, Tsang B, Kim TH, Bah A, Farber P, et al. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. eLife. 2018;7. doi: 10.7554/eLife.31486.

27. Dignon GL, Zheng W, Best RB, Kim YC, Mittal J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(40):9929-34. doi: 10.1073/pnas.1804177115.

28. Lin YH, Brady JP, Chan HS, Ghosh K. A unified analytical theory of heteropolymers for sequence-specific phase behaviors of polyelectrolytes and polyampholytes. The Journal of Chemical Physics. 2020;152(4):045102-. doi: 10.1063/1.5139661.

29. Wang B, Zhang L, Dai T, Qin Z, Lu H, Zhang L, et al. Liquid–liquid phase separation in human health and diseases. Signal Transduction and Targeted Therapy 2021 6:1. 2021;6(1):1-16. doi: 10.1038/s41392-021-00678-1.

30. Hofweber M, Dormann D. Friend or foe-Post-translational modifications as regulators of phase separation and RNP granule dynamics. Journal of Biological Chemistry: American Society for Biochemistry and Molecular Biology Inc.; 2019. p. 7137-50.

31. Zbinden A, Perez-Berlanga M, De Rossi P, Polymenidou M. Phase Separation and Neurodegenerative Diseases: A Disturbance in the Force. Dev Cell. 2020;55(1):45-68. Epub 2020/10/14. doi: 10.1016/j.devcel.2020.09.014. PubMed PMID: 33049211.

32. Schmidt HB, Barreau A, Rohatgi R. Phase separation-deficient TDP43 remains functional in splicing. Nat Commun. 2019;10(1):4890. Epub 2019/10/28. doi: 10.1038/s41467-019-12740-2. PubMed PMID: 31653829; PubMed Central PMCID: PMCPMC6814767.

33. Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, et al. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. Cell. 2015;162(5):1066-77. Epub 2015/09/01. doi: 10.1016/j.cell.2015.07.047. PubMed PMID: 26317470.

34. Ray S, Singh N, Kumar R, Patel K, Pandey S, Datta D, et al. alpha-Synuclein aggregation nucleates through liquid-liquid phase separation. Nat Chem. 2020;12(8):705-16. Epub 2020/06/10. doi: 10.1038/s41557-020-0465-9. PubMed PMID: 32514159.

35. Buratti E, Baralle FE. Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR exon 9. J Biol Chem. 2001;276(39):36337-43. Epub 2001/07/27. doi: 10.1074/jbc.M104236200. PubMed PMID: 11470789.

36. Hallegger M, Chakrabarti AM, Lee FCY, Lee BL, Amalietti AG, Odeh HM, et al. TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. Cell. 2021;184(18):4680-96.e22. doi: 10.1016/J.CELL.2021.07.018.

37. Lin Y, Zhou X, Kato M, Liu D, Ghaemmaghami S, Tu BP, et al. Redox-mediated regulation of an evolutionarily conserved cross-beta structure formed by the TDP43 low complexity domain. Proc Natl Acad Sci U S A. 2020;117(46):28727-34. Epub 2020/11/05. doi: 10.1073/pnas.2012216117. PubMed PMID: 33144500; PubMed Central PMCID: PMCPMC7682574.

38. Conicella AE, Zerze GH, Mittal J, Fawzi NL. ALS Mutations Disrupt Phase Separation Mediated by alpha-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. Structure. 2016;24(9):1537-49. Epub 2016/08/23. doi: 10.1016/j.str.2016.07.007. PubMed PMID: 27545621; PubMed Central PMCID: PMCPMC5014597.

39. Bremer A, Farag M, Borcherds WM, Peran I, Martin EW, Pappu RV, et al. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nat Chem. 2022;14(2):196-207. Epub 2021/12/22. doi: 10.1038/s41557-021-00840-w. PubMed PMID: 34931046; PubMed Central PMCID: PMCPMC8818026.

40. Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Research. 2019;47(D1):D506-D15. doi: 10.1093/nar/gky1049.

41. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology. 1982;157(1):105-32. doi: 10.1016/0022-2836(82)90515-0.

42. Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. Protein-Sol: A web tool for predicting protein solubility from sequence. Bioinformatics. 2017;33(19):3098-100. doi: 10.1093/bioinformatics/btx345.

43. Erdos G, Dosztanyi Z. Analyzing Protein Disorder with IUPred2A. Curr Protoc Bioinformatics. 2020;70(1):e99. Epub 2020/04/03. doi: 10.1002/cpbi.99. PubMed PMID: 32237272.
44. Meszaros B, Erdos G, Dosztanyi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018;46(W1):W329-W37. Epub 2018/06/04. doi: 10.1093/nar/gky384. PubMed PMID: 29860432; PubMed Central PMCID: PMCPMC6030935.
45. Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018;34(14):2499-502. Epub 2018/03/13. doi: 10.1093/bioinformatics/bty140. PubMed PMID: 29528364; PubMed Central PMCID: PMCPMC6658705.
46. Ning W, Guo Y, Lin S, Mei B, Wu Y, Jiang P, et al. DrLLPS: a data resource of liquid–liquid phase separation in eukaryotes. Nucleic Acids Research. 2020;48(D1):D288-D95. doi: 10.1093/NAR/GKZ1027.
47. March ZM, King OD, Shorter J. Prion-like domains as epigenetic regulators, scaffolds for subcellular organization, and drivers of neurodegenerative disease. Brain Res. 2016;1647:9-18. Epub 2016/03/22. doi: 10.1016/j.brainres.2016.02.037. PubMed PMID: 26996412; PubMed Central PMCID: PMCPMC5003744.
48. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Nature Genetics: Nature Publishing Group; 2000. p. 25-9.
49. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004;20(18):3710-5. doi: 10.1093/bioinformatics/bth456.
50. Carbon S, Douglass E, Dunn N, Good B, Harris NL, Lewis SE, et al. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research. 2019;47(D1):D330-D8. doi: 10.1093/nar/gky1055.
51. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS ONE. 2011;6(7):e21800-e. doi: 10.1371/journal.pone.0021800.
52. Fowler NJ, Blanford CF, de Visser SP, Warwicker J. Features of reactive cysteines discovered through computation: from kinase inhibition to enrichment around protein degrons. Sci Rep. 2017;7(1):16338. Epub 2017/11/29. doi: 10.1038/s41598-017-15997-z. PubMed PMID: 29180682; PubMed Central PMCID: PMCPMC5703995.
53. Berman HM. The Protein Data Bank. Nucleic Acids Research. 2000;28(1):235-42. doi: 10.1093/nar/28.1.235.
54. Rokach L, Maimon O. Clustering Methods. Data mining and knowledge discovery handbook. Boston, MA: Springer; 2005. p. 321-52.
55. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010;31:651-66.
56. Jain AKM, M. N.; Flynn, P. J. . Data clustering: a review. ACM Computing Surveys. 1999;31(3):264-323.
57. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261-72. Epub 2020/02/06. doi: 10.1038/s41592-019-0686-2. PubMed PMID: 32015543; PubMed Central PMCID: PMCPMC7056644.

58. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. Nature Reviews Molecular Cell Biology. 2017;18(5):285-98. doi: 10.1038/nrm.2017.7.

59. Jin F, Grater F. How multisite phosphorylation impacts the conformations of intrinsically disordered proteins. PLoS Computational Biology. 2021;17(5 May). doi: 10.1371/journal.pcbi.1008939.

60. Sawle L, Ghosh K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. Journal of Chemical Physics. 2015;143(8):085101-. doi: 10.1063/1.4929391.

61. Couthouis J, Hart MP, Shorter J, DeJesus-Hernandez M, Erion R, Oristano R, et al. A yeast functional screen predicts new candidate ALS disease genes. Proc Natl Acad Sci U S A. 2011;108(52):20881-90. Epub 2011/11/09. doi: 10.1073/pnas.1109434108. PubMed PMID: 22065782; PubMed Central PMCID: PMCPMC3248518.

62. Franzmann TM, Alberti S. Protein Phase Separation as a Stress Survival Strategy. Cold Spring Harb Perspect Biol. 2019;11(6). Epub 2019/01/09. doi: 10.1101/cshperspect.a034058. PubMed PMID: 30617047; PubMed Central PMCID: PMCPMC6546044.

63. Cai D, Feliciano D, Dong P, Flores E, Gruebele M, Porat-Shliom N, et al. Phase separation of YAP reorganizes genome topology for long-term YAP target gene expression. Nat Cell Biol. 2019;21(12):1578-89. Epub 2019/12/04. doi: 10.1038/s41556-019-0433-z. PubMed PMID: 31792379; PubMed Central PMCID: PMCPMC8259329.

64. Alexander EJ, Ghanbari Niaki A, Zhang T, Sarkar J, Liu Y, Nirujogi RS, et al. Ubiquilin 2 modulates ALS/FTD-linked FUS-RNA complex dynamics and stress granule formation. Proc Natl Acad Sci U S A. 2018;115(49):E11485-E94. Epub 2018/11/18. doi: 10.1073/pnas.1811997115. PubMed PMID: 30442662; PubMed Central PMCID: PMCPMC6298105.

65. Dao TP, Kolaitis RM, Kim HJ, O'Donovan K, Martyniak B, Colicino E, et al. Ubiquitin Modulates Liquid-Liquid Phase Separation of UBQLN2 via Disruption of Multivalent Interactions. Mol Cell. 2018;69(6):965-78 e6. Epub 2018/03/13. doi: 10.1016/j.molcel.2018.02.004. PubMed PMID: 29526694; PubMed Central PMCID: PMCPMC6181577.

66. Hennig S, Kong G, Mannen T, Sadowska A, Kobelke S, Blythe A, et al. Prion-like domains in RNA binding proteins are essential for building subnuclear paraspeckles. J Cell Biol. 2015;210(4):529-39. Epub 2015/08/19. doi: 10.1083/jcb.201504117. PubMed PMID: 26283796; PubMed Central PMCID: PMCPMC4539981.

67. Daneshvar K, Ardehali MB, Klein IA, Hsieh FK, Kratkiewicz AJ, Mahpour A, et al. lncRNA DIGIT and BRD3 protein form phase-separated condensates to regulate endoderm differentiation. Nat Cell Biol. 2020;22(10):1211-22. Epub 2020/09/09. doi: 10.1038/s41556-020-0572-2. PubMed PMID: 32895492; PubMed Central PMCID: PMCPMC8008247.

68. Wu SY, Chiang CM. The double bromodomain-containing chromatin adaptor Brd4 and transcriptional regulation. J Biol Chem. 2007;282(18):13141-5. Epub 2007/03/03. doi: 10.1074/jbc.R700001200. PubMed PMID: 17329240.

69. Han X, Yu D, Gu R, Jia Y, Wang Q, Jaganathan A, et al. Roles of the BRD4 short isoform in phase separation and active gene transcription. Nat Struct Mol Biol. 2020;27(4):333-41. Epub 2020/03/24. doi: 10.1038/s41594-020-0394-8. PubMed PMID: 32203489.

70. Wang W, Qiao S, Li G, Cheng J, Yang C, Zhong C, et al. A histidine cluster determines YY1-compartmentalized coactivators and chromatin elements in phase-separated enhancer

clusters. Nucleic Acids Res. 2022;50(9):4917-37. Epub 2022/04/08. doi: 10.1093/nar/gkac233. PubMed PMID: 35390165; PubMed Central PMCID: PMCPMC9122595.

71. Li HR, Chiang WC, Chou PC, Wang WJ, Huang JR. TAR DNA-binding protein 43 (TDP-43) liquid-liquid phase separation is mediated by just a few aromatic residues. J Biol Chem. 2018;293(16):6090-8. Epub 2018/03/08. doi: 10.1074/jbc.AC117.001037. PubMed PMID: 29511089; PubMed Central PMCID: PMCPMC5912450.

72. Ma L, Gao Z, Wu J, Zhong B, Xie Y, Huang W, et al. Co-condensation between transcription factor and coactivator p300 modulates transcriptional bursting kinetics. Molecular Cell. 2021;81(8):1682-97.e7. doi: https://doi.org/10.1016/j.molcel.2021.01.031.

73. Gleixner AM, Verdone BM, Otte CG, Anderson EN, Ramesh N, Shapiro OR, et al. NUP62 localizes to ALS/FTLD pathological assemblies and contributes to TDP-43 insolubility. Nature Communications. 2022;13(1):3380. doi: 10.1038/s41467-022-31098-6.

74. Nag N, Sasidharan S, Uversky VN, Saudagar P, Tripathi T. Phase separation of FG-nucleoporins in nuclear pore complexes. Biochim Biophys Acta Mol Cell Res. 2022;1869(4):119205. Epub 2022/01/08. doi: 10.1016/j.bbamcr.2021.119205. PubMed PMID: 34995711.

75. Dormann D. FG-nucleoporins caught in the act of liquid-liquid phase separation. J Cell Biol. 2020;219(1). Epub 2019/12/14. doi: 10.1083/jcb.201910211. PubMed PMID: 31834369; PubMed Central PMCID: PMCPMC7039191.

76. Basu S, Mackowiak SD, Niskanen H, Knezevic D, Asimi V, Grosswendt S, et al. Unblending of Transcriptional Condensates in Human Repeat Expansion Disease. Cell. 2020;181(5):1062-79 e30. Epub 2020/05/11. doi: 10.1016/j.cell.2020.04.018. PubMed PMID: 32386547; PubMed Central PMCID: PMCPMC7261253.

77. Palacio M, Taatjes DJ. Merging Established Mechanisms with New Insights: Condensates, Hubs, and the Regulation of RNA Polymerase II Transcription. J Mol Biol. 2022;434(1):167216. Epub 2021/09/03. doi: 10.1016/j.jmb.2021.167216. PubMed PMID: 34474085; PubMed Central PMCID: PMCPMC8748285.

78. Jiang H, Wang S, Huang Y, He X, Cui H, Zhu X, et al. Phase transition of spindle-associated protein regulate spindle apparatus assembly. Cell. 2015;163(1):108-22. Epub 2015/09/22. doi: 10.1016/j.cell.2015.08.010. PubMed PMID: 26388440; PubMed Central PMCID: PMCPMC4607269.

79. Avecilla ARC, Quiroz FG. Cracking the Skin Barrier: Liquid-Liquid Phase Separation Shines under the Skin. JID Innovations. 2021;1(3):100036. doi: https://doi.org/10.1016/j.xjidi.2021.100036.

80. Dury AY, El Fatimy R, Tremblay S, Rose TM, Côté J, De Koninck P, et al. Nuclear Fragile X Mental Retardation Protein is localized to Cajal bodies. PLoS Genet. 2013;9(10):e1003890. Epub 2013/11/10. doi: 10.1371/journal.pgen.1003890. PubMed PMID: 24204304; PubMed Central PMCID: PMCPMC3814324.

81. Ding X, Gu S, Xue S, Luo SZ. Disease-associated mutations affect TIA1 phase separation and aggregation in a proline-dependent manner. Brain Res. 2021;1768:147589. Epub 2021/07/27. doi: 10.1016/j.brainres.2021.147589. PubMed PMID: 34310938.

82. Boehning M, Dugast-Darzacq C, Rankovic M, Hansen AS, Yu T, Marie-Nelly H, et al. RNA polymerase II clustering through carboxy-terminal domain phase separation. Nature structural &amp; molecular biology. 2018;25(9):833-40. doi: 10.1038/s41594-018-0112-y. PubMed PMID: 30127355.

83. Lin Y-H, Qiu D-C, Chang W-H, Yeh Y-Q, Jeng US, Liu F-T, et al. The intrinsically disordered N-terminal domain of galectin-3 dynamically mediates multisite self-association

of the protein through fuzzy interactions. The Journal of biological chemistry. 2017;292(43):17845-56. doi: 10.1074/jbc.m117.802793. PubMed PMID: 28893908.

84. Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M, et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell. 2018;174(3):688-. doi: 10.1016/J.CELL.2018.06.006.

85. Murthy AC, Dignon GL, Kan Y, Zerze GH, Parekh SH, Mittal J, et al. Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. Nat Struct Mol Biol. 2019;26(7):637-48. Epub 2019/07/05. doi: 10.1038/s41594-019-0250-x. PubMed PMID: 31270472; PubMed Central PMCID: PMCPMC6613800.

86. Banjade S, Wu Q, Mittal A, Peeples WB, Pappu RV, Rosen MK. Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(47):E6426-35. doi: 10.1073/pnas.1508778112. PubMed PMID: 26553976.

87. Frey S, Richter RP, Gorlich D. FG-rich repeats of nuclear pore proteins form a three-dimensional meshwork with hydrogel-like properties. Science. 2006;314(5800):815-7. Epub 2006/11/04. doi: 10.1126/science.1132516. PubMed PMID: 17082456.

88. Kwon I, Kato M, Xiang S, Wu L, Theodoropoulos P, Mirzaei H, et al. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. Cell. 2013;155(5):1049-60. Epub 2013/11/26. doi: 10.1016/j.cell.2013.10.033. PubMed PMID: 24267890; PubMed Central PMCID: PMCPMC4010232.

89. Schwartz JC, Wang X, Podell ER, Cech TR. RNA seeds higher-order assembly of FUS protein. Cell Rep. 2013;5(4):918-25. Epub 2013/11/26. doi: 10.1016/j.celrep.2013.11.017. PubMed PMID: 24268778; PubMed Central PMCID: PMCPMC3925748.

90. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, et al. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research. 2004;32(3):1037-49. doi: 10.1093/nar/gkh253.

91. Sugiyama N, Imamura H, Ishihama Y. Large-scale Discovery of Substrates of the Human Kinome. Sci Rep. 2019;9(1):10503. Epub 2019/07/22. doi: 10.1038/s41598-019-46385-4. PubMed PMID: 31324866; PubMed Central PMCID: PMCPMC6642169.

92. Hecel A, Wątły J, Rowińska-Żyrek M, Świątek-Kozłowska J, Kozłowski H. Histidine tracts in human transcription factors: insight into metal ion coordination ability. JBIC Journal of Biological Inorganic Chemistry. 2018;23(1):81-90. doi: 10.1007/s00775-017-1512-x.

93. Hong K, Song D, Jung Y. Behavior control of membrane-less protein liquid condensates with metal ion-induced phase separation. Nature Communications. 2020;11(1):5554. doi: 10.1038/s41467-020-19391-8.

94. Lu H, Yu D, Hansen AS, Ganguly S, Liu R, Heckert A, et al. Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. Nature. 2018;558(7709):318-23. Epub 2018/06/01. doi: 10.1038/s41586-018-0174-3. PubMed PMID: 29849146; PubMed Central PMCID: PMCPMC6475116.

Supporting information

S1 Fig. Treemap summarising GO analysis based on cellular component. The analysis was carried out for the subset of proteins containing charge-absent regions.

S2 Fig. Treemap summarising GO analysis based on biological process. The analysis was carried out for the subset of proteins containing charge-absent regions.

S3 Fig. Repeats of certain amino acids are enriched in charge-absent regions. All 51 AA windows in the human proteome (excluding proteins predicted to contain a TM segment) are classified as charge-absent (no-charge, blue) or not (with-charge, orange). The percentage of windows containing a tri-AA motif (excluding Lys, Arg, Asp, Glu) is calculated for the two subsets (no-charge, with-charge).

Figure 1.

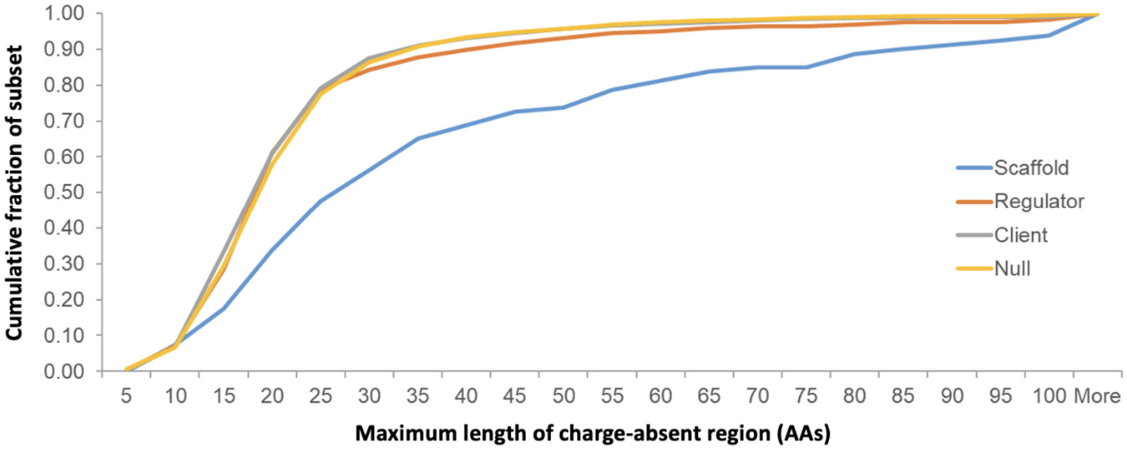

Cumulative fraction of subset
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
More
Maximum length of charge-absent region (AAs)
Scaffold
Regulator
Client
Null

Figure 2.

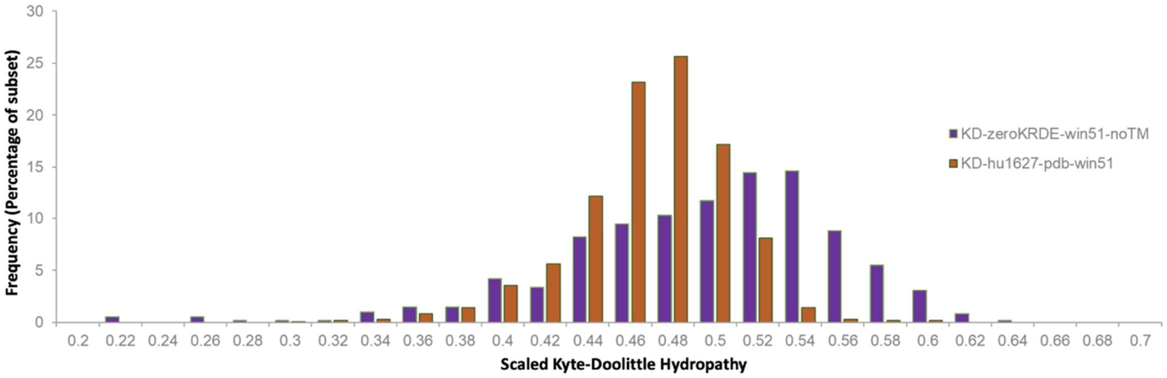

Frequency (Percentage of subset)
30
25
20
15
10
5
0
0.2
0.22
0.24
0.26
0.28
0.3
0.32
0.34
0.36
0.38
0.4
0.42
0.44
0.46
0.48
0.5
0.52
0.54
0.56
0.58
0.6
0.62
0.64
0.66
0.68
0.7
Scaled Kyte-Doolittle Hydropathy
KD-zeroKRDE-win51-noTM
KD-hu1627-pdb-win51

Figure 3.

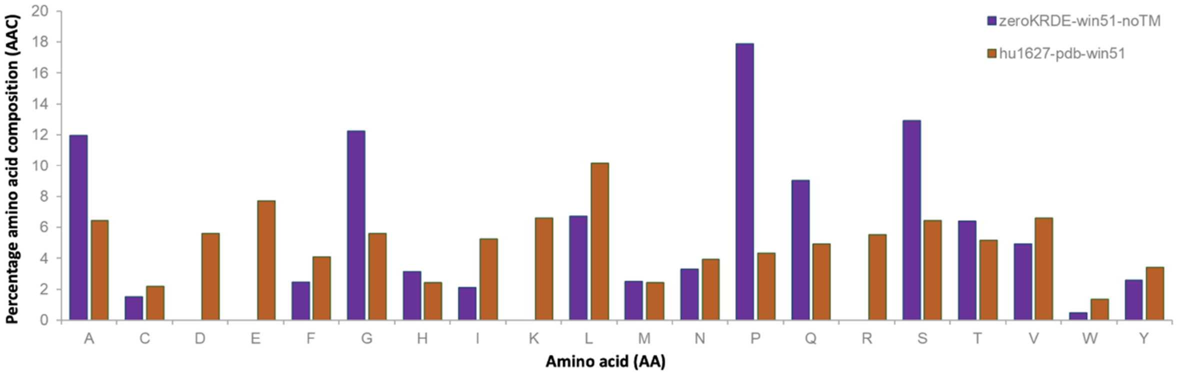

Percentage amino acid composition (AAC)
0
2
4
6
8
10
12
14
16
18
20
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
Amino acid (AA)
zeroKRDE-win51-noTM
hu1627-pdb-win51

Figure 4.

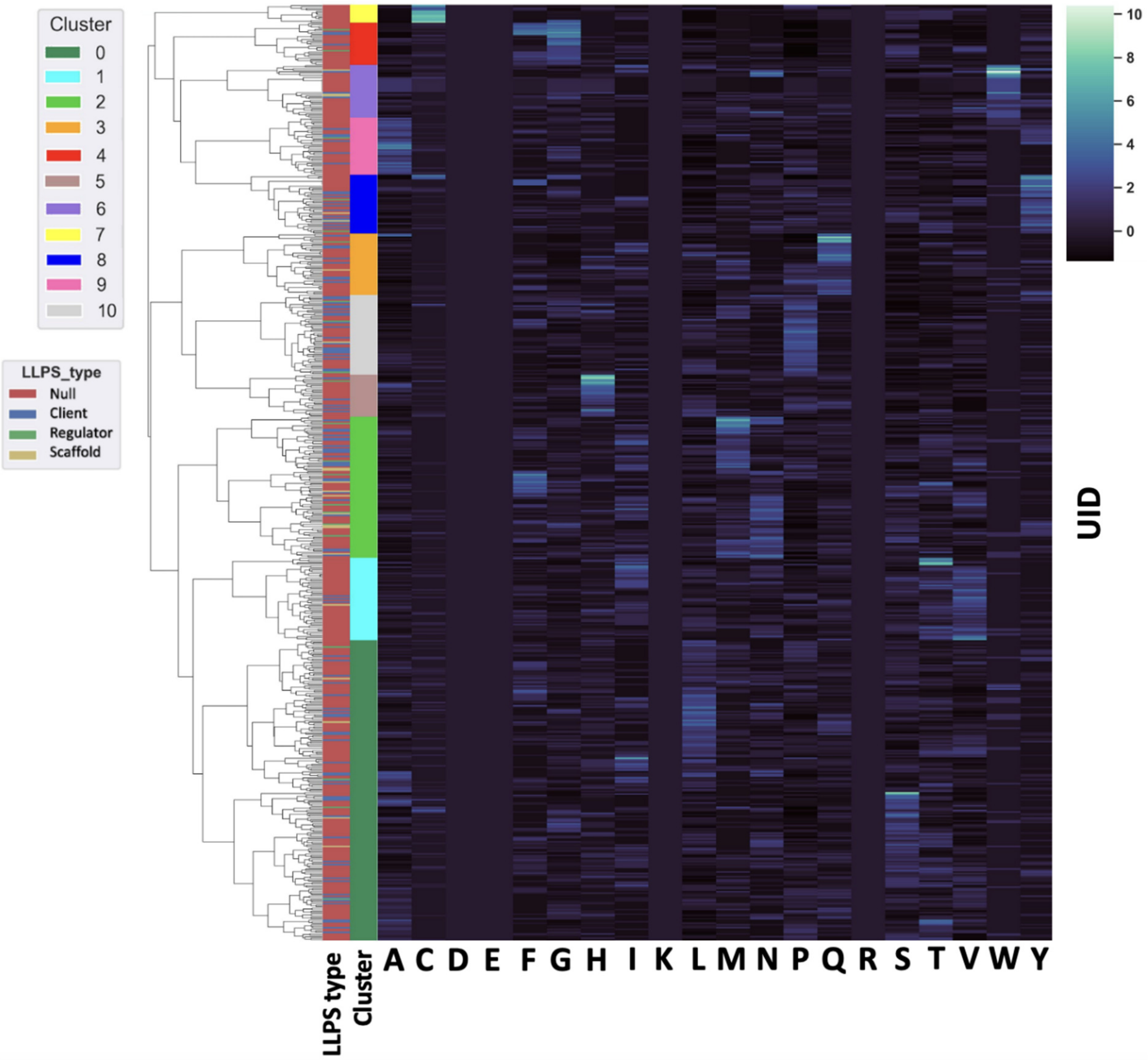

Cluster
0
1
2
3
4
5
6
7
8
9
10
LLPS_type
Null
Client
Regulator
Scaffold
10
8
6
4
2
0
UID
LLPS type
Cluster
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y

Figure S1.

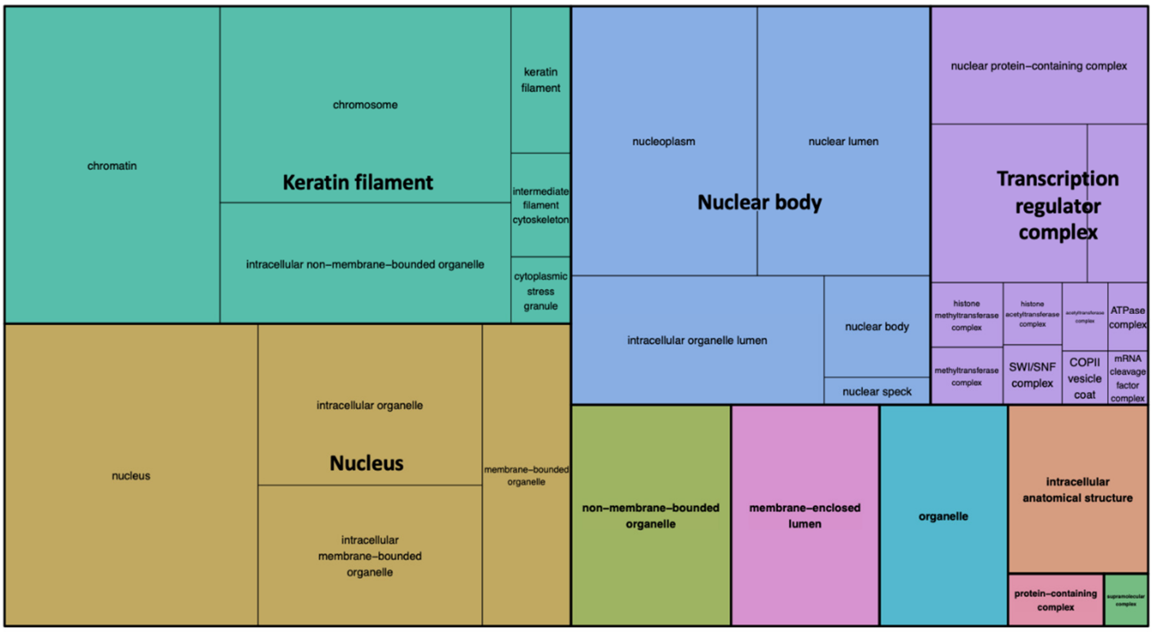

chromatin
chromosome
Keratin filament
intracellular non–membrane–bounded organelle
keratin filament
intermediate filament cytoskeleton
cytoplasmic stress granule
nucleoplasm
nuclear lumen
Nuclear body
intracellular organelle lumen
nuclear body
nuclear speck
nuclear protein–containing complex
Transcription regulator complex
histone methyltransferase complex
histone acetyltransferase complex
acetyltransferase complex
ATPase complex
methyltransferase complex
SWI/SNF complex
COPII vesicle coat
mRNA cleavage factor complex
nucleus
intracellular organelle
Nucleus
intracellular membrane–bounded organelle
membrane–bounded organelle
non–membrane–bounded organelle
membrane–enclosed lumen
organelle
intracellular anatomical structure
protein–containing complex
supramolecular complex

Figure S2.

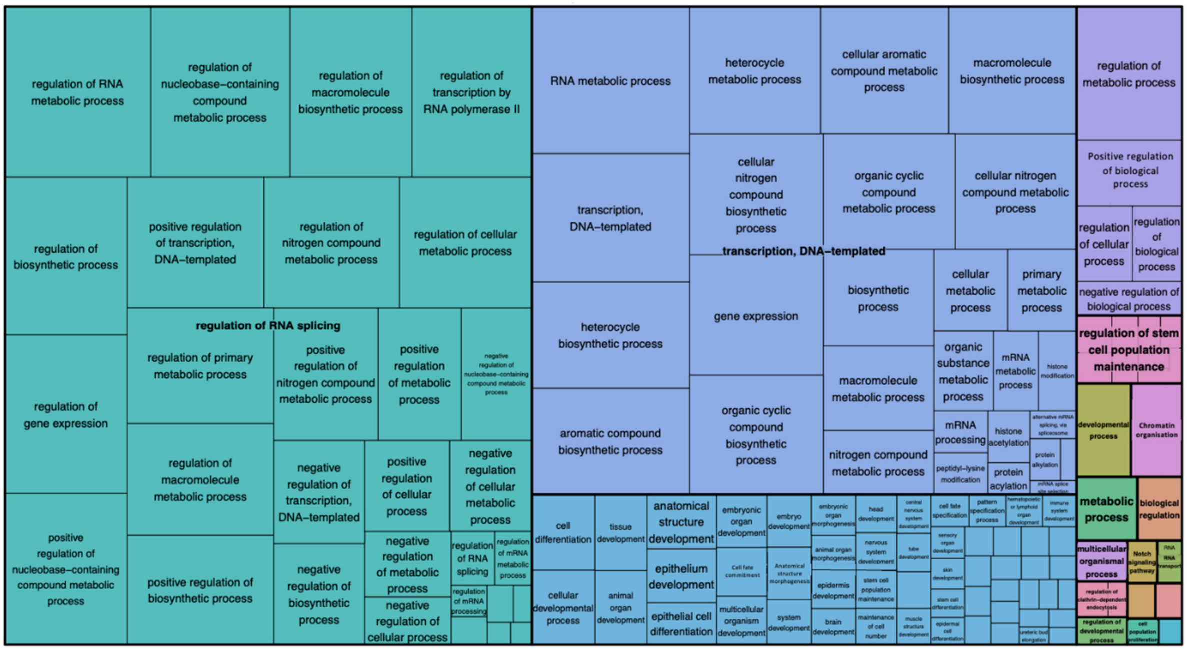
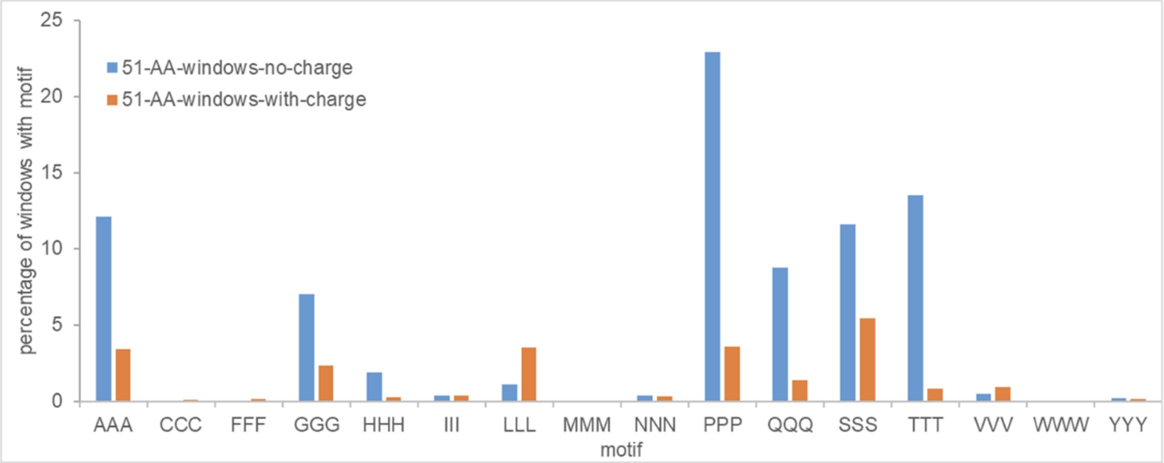

regulation of RNA metabolic process
regulation of nucleobase–containing compound metabolic process
regulation of macromolecule biosynthetic process
regulation of transcription by RNA polymerase II
regulation of biosynthetic process
positive regulation of transcription, DNA–templated
regulation of nitrogen compound metabolic process
regulation of cellular metabolic process
regulation of RNA splicing
regulation of gene expression
regulation of primary metabolic process
positive regulation of nitrogen compound metabolic process
positive regulation of metabolic process
negative regulation of nucleobase–containing compound metabolic process
regulation of macromolecule metabolic process
negative regulation of transcription, DNA–templated
positive regulation of cellular process
negative regulation of cellular metabolic process
positive regulation of nucleobase–containing compound metabolic process
positive regulation of biosynthetic process
negative regulation of biosynthetic process
negative regulation of metabolic process
negative regulation of cellular process
regulation of RNA splicing
regulation of mRNA metabolic process
regulation of mRNA processing
RNA metabolic process
heterocycle metabolic process
cellular aromatic compound metabolic process
macromolecule biosynthetic process
transcription, DNA–templated
cellular nitrogen compound biosynthetic process
organic cyclic compound metabolic process
cellular nitrogen compound metabolic process
transcription, DNA–templated
gene expression
biosynthetic process
cellular metabolic process
primary metabolic process
heterocycle biosynthetic process
organic substance metabolic process
mRNA metabolic process
histone modification
macromolecule metabolic process
aromatic compound biosynthetic process
organic cyclic compound biosynthetic process
mRNA processing
histone acetylation
alternative mRNA splicing, via spliceosome
protein alkylation
nitrogen compound metabolic process
peptidyl–lysine modification
protein acylation
mRNA splice site selection
cell differentiation
tissue development
anatomical structure development
embryonic organ development
embryo development
embryonic organ morphogenesis
head development
central nervous system development
cell fate specification
pattern specification process
hematopoietic or lymphoid organ development
immune system development
sensory organ development
epithelium development
Cell fate commitment
Anatomical structure morphogenesis
animal organ morphogenesis
nervous system development
tube development
skin development
cellular developmental process
animal organ development
epidermis development
stem cell population maintenance
stem cell differentiation
epithelial cell differentiation
multicellular organism development
system development
brain development
maintenance of cell number
muscle structure development
epidermal cell differentiation
ureteric bud elongation
regulation of metabolic process
Positive regulation of biological process
regulation of cellular process
regulation of biological process
negative regulation of biological process
regulation of stem cell population maintenance
developmental process
Chromatin organisation
metabolic process
biological regulation
multicellular organismal process
Notch signaling pathway
RNA
RNA transport
regulation of clathrin–dependent endocytosis
regulation of developmental process
cell population proliferation

Figure S3.


25
20
15
10
5
0
percentage of windows with motif
51-AA-windows-no-charge
51-AA-windows-with-charge
AAA
CCC
FFF
GGG
HHH
III
LLL
MMM
NNN
PPP
QQQ
SSS
TTT
VVV
WWW
YYY
motif