A unified view of low complexity regions (LCRs) across species

Byron Lee[1,*], Nima Jaberi-Lashkari[1,*], and Eliezer Calo[1,2, †]

[1]Department of Biology and Massachusetts Institute of Technology, Cambridge MA, 02139
[2]David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge MA, 02139
[*]equal contribution
[†]Correspondence should be addressed to: Eliezer Calo (calo@mit.edu)

ABSTRACT

Low-complexity regions (LCRs) in proteins are important for higher-order assemblies of organisms, yet we lack a unified view of their sequences, features, relationships, and functions. Here, we use dotplots and dimensionality reduction to systematically define LCR type/copy relationships and create a map of LCR sequence space capable of integrating LCR features and functions. By defining LCR relationships across the proteome, we provide insight into how LCR type and copy number contribute to higher-order assemblies, such as the importance of K-rich LCR copy number for assembly of the nucleolar protein RPA43 *in vivo* and *in vitro*. With LCR maps, we reveal the underlying structure of LCR sequence space, and relate differential occupancy in this space to the conservation and emergence of higher-order assemblies, including the metazoan extracellular matrix and plant cell wall. Together, LCR relationships and maps uncovered the distribution and prevalence of E-rich LCRs in the nucleolus, and revealed previously undescribed regions of LCR sequence space with properties of higher order assemblies, including a teleost-specific T/H-rich sequence space. Thus, this unified view of LCRs enables discovery of how LCRs encode higher-order assemblies of organisms.

INTRODUCTION

Proteins which contain low complexity regions (LCRs) have been shown to direct the higher order assembly of membraneless bodies which enable the spatial compartmentalization of key biochemical processes in cells (Boeynaems et al., 2018; Gomes and Shorter, 2019). In light of the role of LCRs in the higher order assembly of membraneless compartments, interest has been renewed in how these functions are encoded by LCRs.

LCRs are contiguous regions in proteins of low sequence entropy, and several sequence features of these compositionally biased sequences can contribute to the incorporation of proteins into higher order assemblies. Experimental approaches, such as NMR and SAXS, have found examples where specific residues are required for the intermolecular interactions responsible for higher order assembly (Kim et al., 2019b; Martin et al., 2020). Computational identification of short linear motifs (SLiMs) have catalogued specific sub-sequences in LCRs which mediate certain interactions and post-translational modifications (Krystkowiak and Davey, 2017; Kumar et al., 2020), and biophysical predictions of LCRs have given insight into how certain physical properties may direct self-assembly of large compartments (Das and Pappu, 2013; Martin et al., 2020). Valency, defined by the number of binding sites in a molecule, facilitates the formation of higher order assemblies through interactions between multivalent scaffold proteins, which recruit low-valency clients (Banani et al., 2016). Valency can be encoded in any type of sequence (Li et al., 2012; Banani et al., 2016, 2017), yet it has only been studied in a few LCRs.

Numerous proteins have multiple LCRs, and the sequence relationships between these LCRs can impact protein function and higher order assembly. Recent work has shown that in proteins with multiple LCRs, the contributions of individual LCRs on protein function can depend on their identities (Hebert and Matera, 2000; Mitrea et al., 2016; Yang et al., 2020). Synthetic systems have shown that multiple copies of the same LCR can increase the valency of a protein (Schuster et al., 2018). However, the extent to which multiple copies of compositionally similar

LCRs contribute to valency in natural proteins has not been broadly studied. Furthermore, studies of proteins with compositionally distinct LCRs have shown that they can differentially contribute to the function of the protein, likely through their abilities to interact with different sequences (Hebert and Matera, 2000; Mitrea et al., 2016; Yang et al., 2020). Thus, the copy number and type of LCRs in proteins has large effects on their function for the few types of LCRs where they have been molecularly studied. However, we lack the global view required to understand how LCR relationships affect protein function.

More broadly, the importance of LCR features and relationships discussed above is not restricted to proteins of intracellular higher order assemblies. Structural assemblies such as the extracellular matrix (Forgacs et al., 2003; Rauscher and Pomès, 2017), spider silk (Xu and Lewis, 1990; Hinman and Lewis, 1992; Malay et al., 2020), and the siliceous skeleton of certain sponges (Shimizu et al., 2015) are comprised of proteins which share features with proteins involved in intracellular assemblies, such as multivalent scaffolding proteins abundant in LCRs. In fact, many of the proteins comprising these assemblies are composed of almost entirely LCRs which are known to mediate multivalent interactions (Rauscher et al., 2006; Malay et al., 2020). Thus, despite having vastly different functions and emergent physical properties, valency and hierarchical assembly seem to also play a role in these diverse extracellular assemblies. Given that LCRs are required for such diverse assemblies, how diverse are the sequences of natural LCRs, especially given their low sequence complexity? How do differences in the sequences, biophysical properties, copy number and type of LCRs correspond to differences in the higher order assemblies which they form? A unified view of LCRs which incorporates the sequences, features, relationships, and functions of LCRs may allow us to both gain detailed insights into how specific LCRs contribute to higher order assemblies, and gain a broader understanding of LCR sequences and their corresponding functions.

Here, we use systematic dotplot analysis to provide a comprehensive, unified view of LCRs. Our approach to identify LCRs has the unique capability of capturing the relationships between LCRs within proteins, allowing us to define LCR type and copy number across the proteome. On the basis of these features, we chose to study RPA43 and showed that the copy number of its LCRs is important for its higher order assembly. When paired with dimensionality reduction, our dotplot approach provides a complete view of LCR sequence space, highlighting the continuum of sequence in which natural LCRs exist. By integrating this view with additional features of LCRs such as biophysical predictions, we highlight the distribution, distinguishing features, and conserved prevalence of E-rich LCR sequences among nucleolar proteins. To understand the relationship between LCR sequence and higher order assemblies more broadly, we applied our approach to the proteomes of several species, where the conservation and emergence of higher order assemblies can be observed with respect to occupancy of LCR sequence space. Through this unified view, our understanding of LCRs can expand beyond isolated features or functions, enabling further study of how LCRs, and the higher order assemblies they make up, function in organisms.

RESULTS

Dotplots reveal the presence and organization of low complexity regions (LCRs) in proteins

To gain a view of LCRs and their relationships, we leveraged the dotplot matrix method of sequence comparison (Gibbs and Mcintyre, 1970; Pearson and Lipman, 1988). In self-comparison dotplots, every position in the protein is compared to every other position in the protein in a 2D-matrix. Any positions where the two corresponding amino acids are identical is assigned a value of one, while non-matching positions are assigned a value of zero. Self-comparison dotplots are symmetrical across the diagonal, which will always have a value of one as it represents the comparison of each position in the protein to itself. The low sequence complexity of LCRs leads to a strong signature of LCRs in dotplot matrices. Within a single LCR, the frequent recurrence of amino acids leads to many identical matches, which appear as a dense square region centered on the diagonal. Moreover, for proteins with multiple LCRs, compositionally similar LCRs will result in dense square regions off of the diagonal in the position corresponding to the intersection of both LCRs, but different LCRs will not intersect in this way. Therefore, dotplots are capable of identifying both the total number of LCRs in proteins and the relationships between similar and distinct LCRs for proteins with multiple LCRs.

For example, in the dotplot of G3BP1, a protein important for stress granule assembly (Yang et al., 2020), the dense squares along the diagonal clearly distinguish between the LCRs of G3BP1 and its other regions, which include its RNA-recognition motif (RRM) and dimerization domains (Figure 1A, dotted black outlines). Immediately apparent from the dotplot of G3BP1 is that its LCRs are not all the same type as dense squares do not occupy every intersection off the diagonal (Figure 1A, red outlines). The first two LCRs are acidic in composition, while the third is an RGG domain which plays a role in RNA-binding (Kim et al., 2019a). The presence of these compositionally distinct LCRs are critical for the ability of G3BP1 to form stress granules, as the acidic LCRs interact with and inhibit the RGG domain, preventing it from interacting with RNA, a necessary step of stress granule assembly (Guillén-Boixet et al., 2020; Yang et al., 2020). Thus, by highlighting the relationships between different LCRs, dotplots can provide key insights relevant to protein function.

While not all proteins have LCRs, some proteins almost entirely consist of LCRs and exhibit diverse LCR relationships and organization. For example, ACTB and SYTC lack LCRs, which is reflected by the lack of dense squares in their respective dotplots (Figure 1 - figure supplement 1A, B). Other examples, such as SMN, a component of nuclear gems (Liu and Dreyfuss, 1996), and the nucleolar protein KNOP1 (Grasberger and Bell, 2005; Larsson et al., 1999) have more complex architectures, with multiple copies of similar LCRs, which appear with roughly equal spacing (Figure 1 - figure supplement 1C, D). On the other hand, Nucleolin has multiple types of LCRs, one of which occurs in several copies. Interestingly, these LCR types are spatially segregated in the protein, with one set of LCRs in the N-terminal region, and a

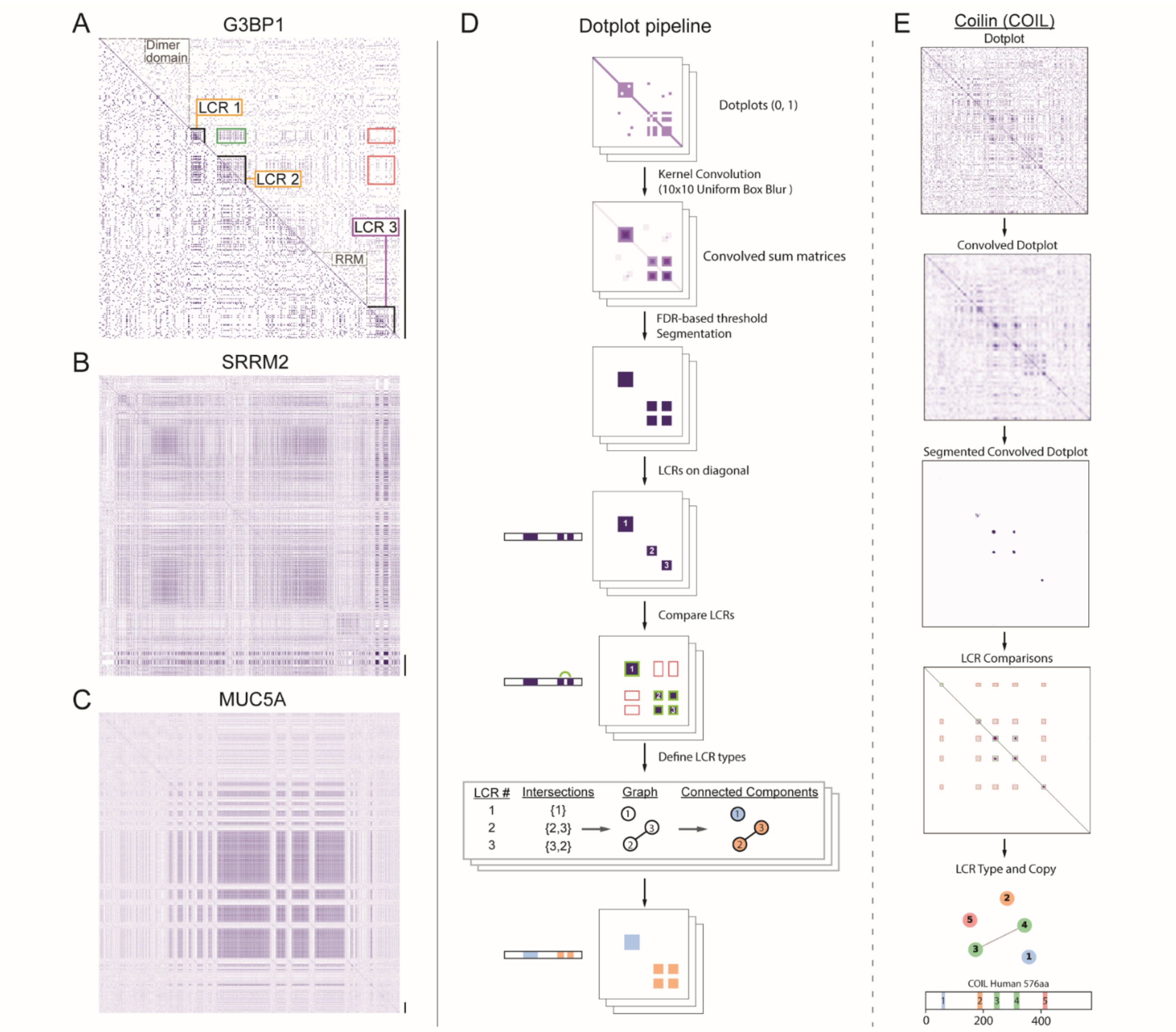


Figure 1: A systematic dotplot approach to reveal the relationships between low complexity regions (LCRs) in proteins

For all dotplots, the protein sequence lies from N-terminus to C-terminus from top to bottom, and left to right. Scale bars on the right of the dotplots represent 200 amino acids in protein length. (A-C) Single dotplots that have not been processed with the dotplot pipeline.

A) Dotplot of G3BP1. Top-right half of dotplot has been annotated with G3BP1s LCRs (solid black lines around diagonal), and functionally important non-LC sequences (dotted lines around diagonal). Off-diagonal comparisons are highlighted by green squares (for similar LCRs), or red squares (for dissimilar LCRs).

B) Dotplot of SRRM2.

C) Dotplot of MUC5A.

D) Schematic of dotplot pipeline, illustrating data generation and processing. Dotplots are generated, convolved using a uniform 10x10 kernel, and segmented based on a proteome-wide FDR-based threshold (same threshold applied to all proteins in the same proteome, see Methods for details). Using segmented dotplots, LCRs are identified as segments which lie along the diagonal. Pairwise off diagonal LCR comparisons are performed for each dotplot, and LCR relationships are represented as a graph. Connected components in this graph represent LCRs of the same type within each protein.

E) Sequential steps of the dotplot pipeline as performed for the human protein coilin (COIL). Shown from top to bottom are the raw dotplot, convolved dotplot, segmented convolved dotplot, LCR-comparison plot, graph representation of LCR relationships, and schematic showing LCR position and type as called by the dotplot pipeline. Numbers represent the LCR identifier within the protein from N-terminus to C-terminus. Different colors in schematic correspond to different LCR types.
See also Figure 1 - figure supplement 1,2,3,4.

different LCR in the C-terminus, highlighting the organizational complexity of LCRs that exists in some proteins (Figure 1 - figure supplement 1E).

As can be seen for SRRM2 and MUCIN5A, a large area in their dotplots consist of LCR signatures off the diagonal (Figure 1B, C), indicating that each of these proteins consist of long stretches of similar LCR sequences. For example, the dotplot of SRRM2 contains multiple regions of low complexity which cover an area corresponding to hundreds of amino acids (Figure 1A, B). SRRM2 and another LCR-containing protein SON (Figure 1 - figure supplement 1H) were recently found to act as essential scaffolds for formation of nuclear speckles (Sharma et al., 2010; Fei et al., 2017; Ilik et al., 2020), suggesting that proteins which each contain long stretches of similar LCR sequences could play important roles in certain higher order assemblies. In fact, many such proteins have been found to be essential for various higher order assemblies. These include UBP2L and PRC2C (Figure 1 - figure supplement 1F, G), which were only recently discovered to be essential for the formation of stress granules (Youn et al., 2018; Sanders et al., 2020). UBP2L was found in some conditions to be upstream of G3BP1 for stress granule formation (Cirillo et al., 2020).

Other proteins with long stretches of similar LCR sequences included mucins (MUC5A shown, Figure 1C), collagens and DSPP (Figure 1 - figure supplement 1I), proteins which are essential to the formation of extracellular assemblies with a diverse variety of physical properties. Mucins are key components of mucus, a liquid/gel-like assembly of glycoproteins (reviewed in (Lai et al., 2009)), while DSPP codes for a protein which scaffolds the mineralization of teeth (Stetler-Stevenson and Veis, 1986; Saito et al., 2000; Sreenath et al., 2003). Although proteins which each contain such long stretches of similar LCRs, such as SRRM2, UBP2L, MUCIN5A, and DSPP, are involved in such diverse biological processes, a commonality among them is their scaffolding roles. The fact that these proteins exhibit similar LCR relationships and roles in their respective assemblies suggests that the LCR relationships revealed by dotplots can inform how we understand protein functions.

The examples of dotplots make clear that functional information about LCR type and copy number can be extracted from dotplot matrices. LCR type has been shown in a handful of examples to be important for how these proteins, such as G3BP1 (Guillén-Boixet et al., 2020; Yang et al., 2020), function biologically. Additionally, copy number is likely an important contributor to valency for proteins which contain such a large proportion of LCRs, such as SRRM2 or MUC5A. However, there currently is not an approach to assess the global relationship between these features of LCRs and their functions. While several methods exist for identifying LCRs (Wootton and Federhen, 1993; Promponas et al., 2000; Albà et al., 2002; Harrison, 2017), these methods are unable to determine LCR types and their respective copy numbers. As a consequence, we have not been able to systematically understand how LCR sequence and organization influence their function. The ability of dotplots to both identify LCRs and provide information on LCR type and copy number presents an opportunity to develop a comprehensive and systematic tool to identify these features of proteins.

A systematic dotplot approach to identify and characterize LCRs proteome-wide

We developed a computational pipeline to extract both the positions and spatial relationships of LCRs using the 2D signature of LCRs in dotplots (Figure 1D, Methods). Because LCRs present themselves as dense squares in a 2D matrix, the identification of LCRs in dotplots was similar to the extraction of features from an image, allowing us to take advantage of image processing tools.

Specifically, we computationally extracted the LCRs of any protein by identifying high density regions in its dotplot through classic image processing methods, such as kernel convolution, thresholding, and segmentation (Figure 1D). To identify high density regions in dotplots, we performed kernel convolution on the dotplots with a uniform 10x10 kernel, which calculates a convolved pixel intensity value from 0 to 100 based on the number of dots in that window. Regions of high density will have higher convolved pixel intensities, while regions of low density will have lower convolved pixel intensities.

In order to define LCRs in the proteome, we employed a false discovery rate (FDR)-based approach to threshold the convolved pixel intensities. For a given proteome, we generated a background model by simulating an equally sized, length matched 'null proteome', whose sequences were generated from a uniform amino acid distribution (see methods for details). We compared the distribution of convolved pixel intensities across all proteins in the real proteome with those from the null proteome and identified the lowest convolved pixel intensity threshold which satisfied a stringent FDR of 0.002 (Figure 1D, Figure 1 - figure supplement 2A, B). This threshold was then applied to every protein in the human proteome to segment high-density regions in all dotplots. The segmented regions along the diagonal correspond to LCRs, while segmented regions off of the diagonal correspond to compositionally similar LCRs within the same protein (Figure 1D). We illustrate this process for Coilin, a scaffolding protein of Cajal bodies in the nucleus (Figure 1E), where dense regions in its dotplot are extracted by our systematic approach.

Across the human proteome, our approach identified 37,342 LCRs in 14,156 proteins (Figure 1 - figure supplement 2C), with nearly 60% of LCR-containing proteins in the human proteome containing more than one LCR (Figure 1 - figure supplement 2E). The Shannon entropy of these regions was significantly lower than that of randomly sampled sequences from the proteome, confirming that they are low complexity (Figure 1 - figure supplement 2D). Furthermore, we observe an inverse relationship between the convolved pixel intensity threshold used for segmentation and the resulting Shannon entropy of called LCRs (Figure 1 - figure supplement 2D). The tight relationship between these values shows that, in general, the density of points in dotplots is inversely related to the informational complexity of corresponding sequence.

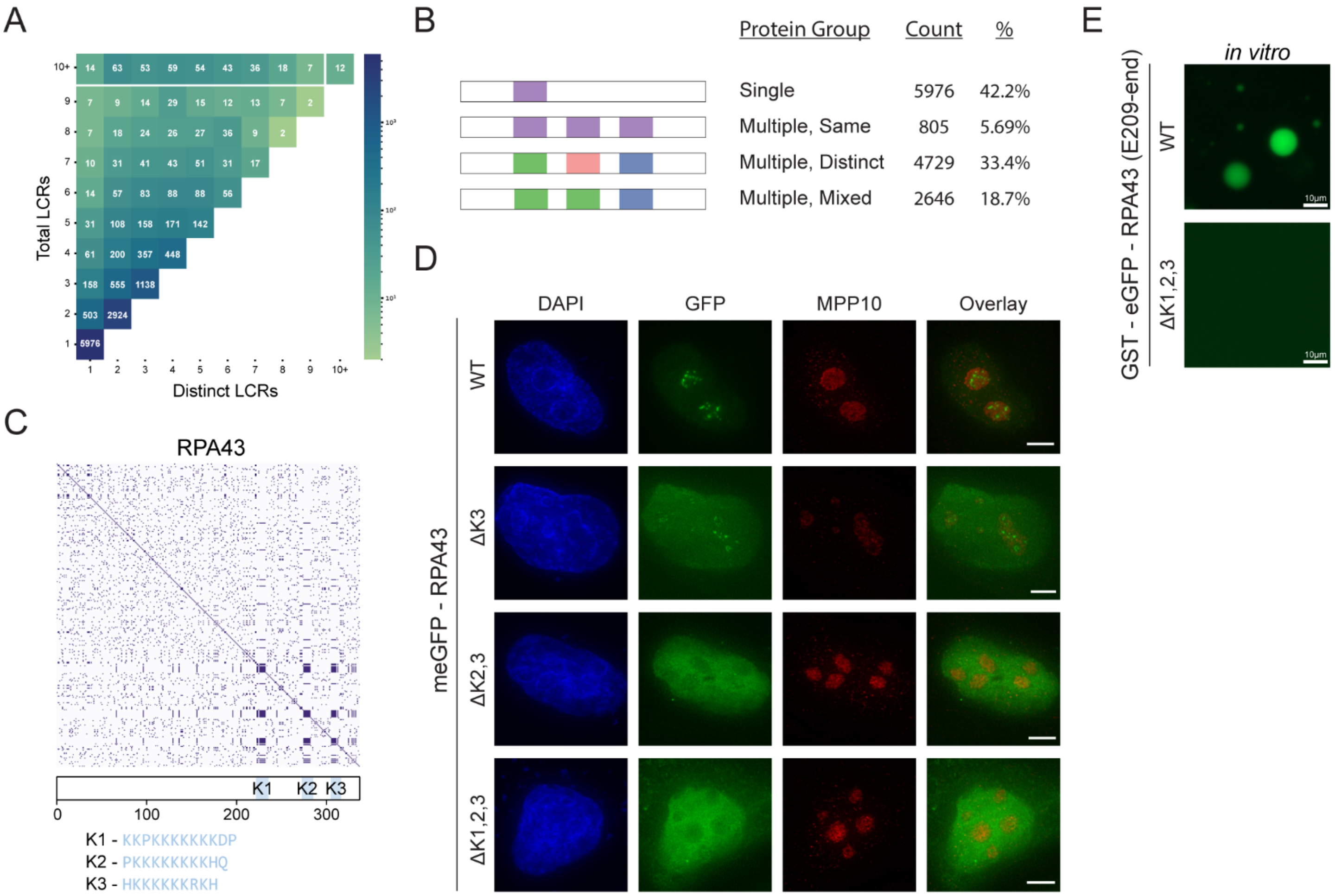
Finally, when compared to two commonly used LCR-callers, SEG (Wootton and Federhen, 1993) and fLPS (Harrison, 2017), our approach achieves a comparable or better performance in minimizing LCR entropy while maximizing total LCR sequence in the human proteome (Figure 1 - figure supplement 3A-D). Furthermore, we call LCRs in regions of proteins similar to those called by other methods, as can be seen in examples CO1A1 and ZN579 (Figure 1 - figure supplement 3E, F). While other approaches (Wootton and Federhen, 1993; Harrison, 2017) are more efficient at identifying the presence of LCRs, our approach allows for proteome-wide identification of LCRs without losing information about LCR type and copy number within proteins. Thus, by making 2D comparisons of LCRs within proteins across the proteome, our systematic dotplot approach provides more information on the relationship between LCRs within proteins, allowing us to ask deeper questions about the role of these features in protein function.

Comparison of LCRs defines type and copy number of LCRs across the proteome

The relationship between LCRs within LCR-containing proteins has not been studied on a proteome-wide scale, despite being important in the cases where it has been studied (Hebert and Matera, 2000; Mitrea et al., 2016; Yang et al., 2020). These relationships can now be studied systematically using our dotplot approach. To this end, we compared all LCRs within each protein for the human proteome. The relationship between two LCRs in a protein is determined by whether or not a segmented region of the dotplot exists off the diagonal in the region corresponding to the intersection of those two LCRs. If so, we designate these two compositionally similar LCRs as the same ‘type’ (green boxes in Figure 1D , Methods). The relationships between LCRs can be summarized as a graph where each LCR is a node, and off-diagonal intersections between pairs of LCRs are represented as edges (Figure 1D, Methods). Within each of these graphs, individual connected components are LCRs of the same type. The number of nodes is the total number of LCRs, while the number of connected components defines the number of distinct LCR types. This graph-based visualization is helpful for seeing LCR relationships within proteins, particularly for proteins with many LCRs (Figure 1 - figure supplement 4), and contributes to our understanding of potential valency provided by LCRs to natural proteins.

Our approach for calling LCR type and copy number is illustrated for numerous examples with a range of different types and copy numbers (Figure 1E, Figure 1 - figure supplement 4). For Coilin, we identify 4 distinct types of LCRs, with one of the types present in two copies (Figure 1E). Of these 4 types, two of them have been shown to play different yet important roles in Coilin localization to cajal bodies (Hebert and Matera, 2000). Thus, our systematic dotplot approach can capture LCR types of known functional importance.

With this, we compared the number of total and distinct LCRs for each protein (Figure 2A). We can see from this analysis that the range in combinations of total and distinct LCRs is diverse across proteins in the human proteome (Figure 2A, Figure 2 - figure supplement 1A), enabling different combinations of LCRs and functions. Based on the number of total and distinct LCRs in a given protein, proteins can be broadly categorized into four groups (Figure 2B), which each make a sizable fraction of the proteome and uniquely contribute to our understanding of how LCRs affect protein-function. We will refer to these groups as ‘single’, ‘multiple-same’, ‘multiple-distinct’, and ‘multiple-mixed’ to reflect the number of total and distinct LCRs that a protein possesses. The single LCR group, which lies in the bottom left corner (Figure 2A), corresponds to proteins with only a single LCR, in which we may assess the isolated function of an LCR. The multiple-same group lies along the vertical axis and corresponds to proteins with multiple LCRs, all of which are the same type (total LCRs >1, distinct LCRS=1). Since all of the LCRs for a given protein in this group are the same, this group is particularly useful for understanding the contribution of LCR copy number to the function of a protein. The multiple-distinct group lies along the diagonal, and corresponds to proteins with multiple LCRs, all of which are distinct from each other (total LCRs=distinct LCRs >1). This



Protein Group	Count	%
Single	5976	42.2%
Multiple, Same	805	5.69%
Multiple, Distinct	4729	33.4%
Multiple, Mixed	2646	18.7%

Figure 2: Proteome-wide definition of LCR type and copy number reveals copy number requirements for nucleolar integration of RPA43

A) Distribution of total and distinct LCRs for all LCR-containing proteins in the human proteome. The number in each square is the number of proteins in the human proteome with that number of total and distinct LCRs and is represented by the colorbar.

B) Illustration of different protein groups defined by their LCR combinations, and the number and percentage (%) of proteins that fall into each group. Group definitions are mutually exclusive.

C) Dotplot and schematic of RPA43. K-rich LCRs are highlighted in blue, and are labeled K1-K3. Sequences of K1-K3 are shown below the schematic.

D) Immunofluorescence of HeLa cells transfected with RPA43 constructs. HeLa cells were seeded on fibronectin-coated coverslips and transfected with the indicated GFP-RPA43 constructs, and collected ~48 h following transfection. DAPI, GFP, and MPP10 channels are shown. Scale bar is 5 μm.

E) Droplet formation assays using GFP-fused RPA43 C-terminus *in vitro*. Droplet assays were performed with 8.3 μM purified protein.

See also Figure 2 - figure supplement 1.

group allows for in-depth study of the relationships between different LCRs. Finally, the multiple-mixed group, which occupies the rest of the graph, corresponds to proteins with multiple LCRs of mixed types, where at least one type is present in at least two copies. This group likely corresponds to more complex proteins which may be affected by both the copy number and type of LCRs they contain. By characterizing the copy number and type of LCRs across the proteome, our approach allows for proteins to be selected on the basis of these features for further study.

LCR copy number impacts protein function

The group of proteins which have multiple LCRs of the same type presents an opportunity to specifically understand the role of LCR copy number in natural proteins. To highlight how these groups could inform us on LCR function, we sought to study the role of LCR copy number on higher order assembly by studying a protein in the ‘multiple same’ group.

We chose to study the RNA Polymerase I component RPA43, which localizes to the nucleolus (Dundr et al., 2002), a multi-component higher order assembly. RPA43 has three LCRs in its C-terminus which are all the same type (Figure 2C, Figure 1 - figure supplement 4A). To understand the common sequences in this LCR type, we manually checked the sequences determined by our systematic analysis. All three LCRs of RPA43 contained a 10-12 amino acid block of mostly K-residues (Figure 1 - figure supplement 4A, bottom row), which were the primary contributor to off-diagonal intersections between these LCRs and thus defined this LCR type. In order to test the importance of LCR copy number in RPA43 function, we chose to focus on the sequences in its three LCRs which make them the same type, the blocks of K-residues. We will refer to these K-rich blocks as K-rich LCRs of RPA43 (K1, K2, and K3 respectively).

While GFP-fused WT RPA43 localized correctly to the fibrillar center of the nucleolus, deletion of all three of its K-rich LCRs (ΔK1,2,3) led to its exclusion from the nucleolus, confirming that these LCRs are important for its higher order assembly (Figure 2D). It is important to note that this mutant retains the predicted disordered nature of RPA43s C-terminal region (Figure 2 - figure supplement 1B), showing that this phenotype cannot be explained by a predicted change in the solvent accessibility of this region. The fact that all of the LCRs of RPA43 are the same type, and that they together are required for nucleolar integration allows us to specifically study the role of LCR copy number in RPA43 higher order assembly.

We next generated RPA43 mutants lacking one or more of its LCRs. Surprisingly, RPA43 mutants with two copies of its LCRs correctly localized to the nucleolus, while those containing only one of its LCRs were excluded from the nucleolus (Figure 2D, Figure 2 - figure supplement 1C). This result held true regardless of what combination of LCRs were present (Figure 2 - figure supplement 1C), showing that these LCRs do not uniquely contribute to RPA43 localization. Rather, it is the copy number of these LCRs which is required for RPA43 integration into the nucleolus. Furthermore, the finding that RPA43 requires two copies of its LCRs suggests that a valency of at least two is required for it to integrate into the nucleolus via its K-rich LCRs.

Consistent with these results, while the recombinant GFP-fused RPA43 C-terminus phase separated into liquid droplets *in vitro*, the GFP-fused RPA43 C-terminus with its three K-rich LCRs specifically deleted did not (Figure 2E). Thus, the RPA43 C-terminus contains the sequences sufficient for higher order assembly, and the K-rich LCRs are necessary for this assembly. This result suggests that the K-rich LCRs are not merely linkers between other self-interacting elements, as deletion of such linkers tends to alter the physical properties of the assembly but not its presence (Martin et al., 2020). Rather, these results suggest that these K-rich LCRs either self-interact (perhaps mediated by solvent anions), or interact with a complementary element in the RPA43 C-terminus. Moreover, the observation that *in vivo* nucleolar localization and *in vitro* phase separation require the same sequences suggest that the interactions mediating RPA43's ability to form a higher order assembly are similar to those mediating its nucleolar integration. Together, these results not only highlight the ability of K-rich LCRs to mediate higher order assembly of RPA43, but highlight the importance of understanding LCR copy number more generally. Thus, our approach allows for targeted experiments to arrive at principles by which LCR copy number affects protein function.

A map of LCRs

In order to relate the sequences, features, relationships, and functions of LCRs across the proteome, we wanted to understand the full breadth of LCR sequences. By using a sequence map as a foundation to integrate these aspects of LCRs, we could begin to understand how differences in sequence correspond to differences in the features, relationships, and functions of LCRs. As such we took an unbiased approach to visualize the sequence space occupied by LCRs in the human proteome.

Using the LCRs identified by dotplots, we represented the amino acid composition of each LCR as a 20-dimensional vector where each dimension corresponds to the frequency of a different amino acid. Thus, each LCR will map to a point in 20-dimensional sequence space. To visualize LCR occupancy in this sequence space, we used Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) to generate a 2-dimensional map of all LCRs in the human proteome (Figure 3A, Figure 3 - figure supplement 1).

We can immediately see from this map that LCRs in the human proteome exhibit a rich diversity of sequence compositions, and do not fall into a handful of isolated groups. Generally, we see that this LCR space has many highly occupied regions (Figure 3A). Using Leiden clustering (Traag et al., 2019), we identified 19 clusters of LCRs in this space which mostly corresponded to high frequency of an amino acid, and serve as useful guides when referring to different regions of the map (Figure 3A). While most of these clusters correspond to LCRs with large contributions from a single amino acid, these clusters still have a substantial presence of many other amino acids. For example, the serine-rich cluster has regions within it that are also

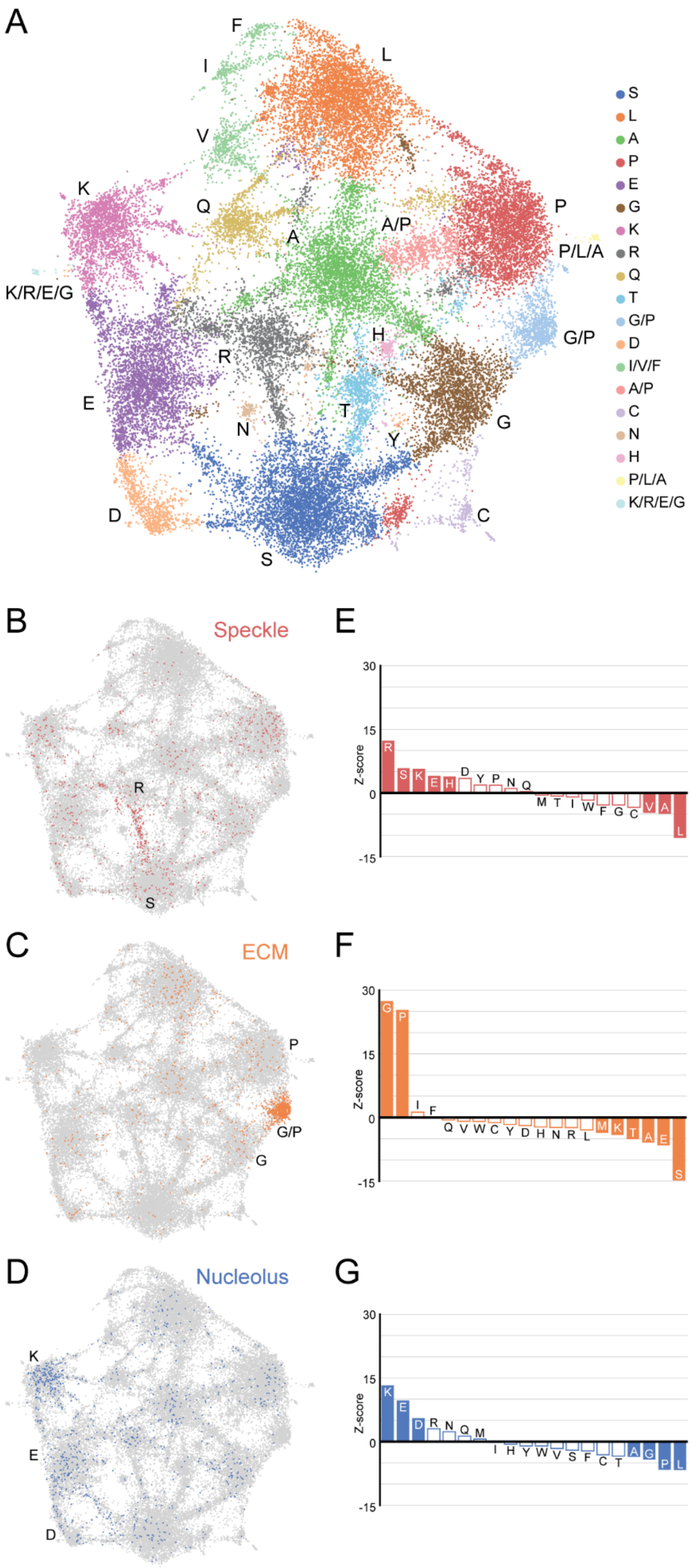


Figure 3: A map of LCRs captures known differences in higher order assemblies

A) UMAP of all LCRs in the human proteome. Each point is a single LCR and its position is based on its amino acid composition (see Methods for details). Clusters identified by the Leiden algorithm are highlighted with different colors. Labels indicate the most prevalent amino acid(s) among LCRs in corresponding Leiden clusters.

B) LCRs of annotated nuclear speckle proteins (obtained from Uniprot, see Methods) plotted on UMAP.

C) Same as B), but for extracellular matrix (ECM) proteins.

D) Same as B), but for nucleolar proteins.

E) Barplot of Wilcoxon rank sum tests for amino acid frequencies of LCRs of annotated nuclear speckle proteins compared to all other LCRs in the human proteome. Filled bars represent amino acids with Benjamini-Hochberg adjusted p-value < 0.001. Positive Z-scores correspond to amino acids enriched in LCRs of nuclear speckle proteins, while negative Z-scores correspond to amino acids depleted in LCRs of nuclear speckle proteins.

F) Same as E), but for extracellular matrix (ECM) proteins.

G) Same as E), but for nucleolar proteins.

See also Figure 3 - figure supplement 1,2,3,4.

enriched for other amino acids in addition to serine (Figure 3 - figure supplement 2A). These regions of the S-rich cluster are typically closer to the main cluster corresponding to the other amino acid, highlighting the richness in diversity of LCR compositions, even within one single cluster.

Strikingly, many clusters are ‘connected’ to other clusters through bridge-like connections, which are much more prominent between certain clusters (Figure 3A, Figure 3 - figure supplement 2). This indicates that some combinations of amino acids commonly co-occur within LCRs which occupy these bridges, while other combinations of amino acids do not co-occur as often. While cluster definitions are discrete, the amino acid compositions of the LCRs that lie along these bridges are continuous (Figure 3 - figure supplement 2B, C). In some cases, such as in the G/P-rich cluster between the main G- and P-rich clusters, these bridges are large enough to form their own clusters (Figure 3A, Figure 3 - figure supplement 2B). The observation that LCRs exhibit a gradual, continuous shift in LCR composition from one end of the bridge to the other raises the possibility that any properties sensitive to the composition of these LCRs may exhibit a similarly gradual and continuous variation, increasing the potential complexity of interactions formed by LCRs.

This map reveals the high degree of nuanced sequence variation that exists in natural LCRs and that certain amino acids coexist to varying degrees in LCRs. By capturing the variation in all LCRs, this global map provides an intuitive foundation for understanding how biological and physical properties of LCRs relate to their sequence.

Higher order assemblies map to specific regions in LCR sequence space

LCRs of certain compositions play important roles in specific higher order assemblies. To gain insight into what different regions of the map represent, we decided to see if higher order assemblies preferentially occupy certain regions in the map.

To do this, we mapped annotations of known higher order assemblies to the LCR map. Nuclear speckle proteins, which are commonly localized by LCRs known as RS-domains (Cáceres et al., 1997; Boucher et al., 2001), populated a bridge between the R and S clusters in LCR sequence space (Figure 3B), and were significantly enriched in both of these amino acids (Figure 3E). LCRs of extracellular matrix (ECM) proteins were heavily concentrated in a G/P-rich region (Figure 3C, F), reflecting the many, long collagen proteins in humans. LCRs of nucleolar proteins mapped to the K-rich cluster in LCR sequence space (Figure 3D, G), consistent with nucleolar localization signals possessing K-rich sequences (Scott et al., 2010). Other higher order assemblies also mapped to specific regions in the LCR sequence space, including the centrosome and nuclear pore complex (Figure 3 - figure supplement 3A, B). Wilcoxon rank-sum tests for each of the 20 amino acids confirmed that these spatial biases in the LCR map corresponded to actual differences in LCR composition, independent of Leiden cluster assignment (Figure 3E-G, Figure 3 - figure supplement 3E, F). Conversely, some higher order assemblies are known to not have many individual proteins which share a specific type of LCR, including stress granules for which RNA is a major contributor (Guillén-Boixet et al., 2020; Sanders et al., 2020), and PML bodies which depend on SUMOylation of non-LC sequences (Shen et al., 2006). As expected for these cases, there was neither a spatial bias in the LCR map, nor significantly enriched amino acids (Figure 3 - figure supplement 3C, D, G, H).

The ability of the map to highlight the biased LCR compositions of certain higher order assemblies demonstrates that we can capture how differences in sequence correspond to known differences in function. Thus, the LCR map allows us to interrogate the relationship between less understood regions of LCR space and protein function.

A unified view of LCRs reveals the prevalence of E-rich LCRs in nucleolar proteins

When examining the distribution of nucleolar protein LCRs across the LCR map, we found that in addition to the K-rich cluster, the E-rich cluster was significantly occupied (Figure

3D, G). While there is evidence for the importance of acidic stretches in phase separated bodies (Hebert and Matera, 2000; Mitrea et al., 2016; Guillén-Boixet et al., 2020; Yang et al., 2020), nucleolar LCRs were significantly more likely than LCRs of speckle proteins to have a high frequency of E residues, but not D residues (Figure 4A). This observation suggested that E-rich LCRs may play a nucleolus-specific role.

To see if the nucleolar E-rich LCRs had any features which could give insight into their role in the nucleolus, we first integrated our dataset with biophysical predictions relevant to higher order assemblies (Figure 3 - figure supplement 4). These included IUPred2 which predicts protein disorder, and ANCHOR which predicts the probability of a disordered sequence to become ordered upon binding to a globular protein partner (Mészáros et al., 2009, 2018). To maintain the context-dependent nature of these predictions, we first calculated the scores across full-length proteins and then extracted the values for LCRs based on position in the protein. These predictions could be plotted on the LCR map (Figure 3 - figure supplement 4), allowing us to gain insight into the relationship between LCR composition and these properties across all LCRs in the proteome.

By looking at the regions in the LCR map occupied by K- and E-rich LCRs, we noticed that while both K-rich and E-rich LCRs are predicted to be similarly disordered (Figure 3 - figure supplement 4A), the E-rich LCRs had higher ANCHOR scores overall (Figure 3 - figure supplement 4B). To determine if this trend held true for K-rich and E-rich LCRs of nucleolar proteins in particular, we analyzed LCRs in the top 25th percentile of K or E frequency in the nucleolus, which we refer to as K-enriched and E-enriched respectively (Figure 4B). The distributions of IUPred2 scores among these K- and E-enriched nucleolar LCRs are only subtly different (Figure 4C), consistent with what was gleaned from the LCR map. However, while K-enriched nucleolar LCRs exhibited a unimodal distribution of ANCHOR scores centered below 0.5, E-enriched LCRs exhibited a bimodal distribution (Figure 4D). One peak of this distribution was at approximately the same ANCHOR score as nucleolar K-rich LCRs (Figure 4D).

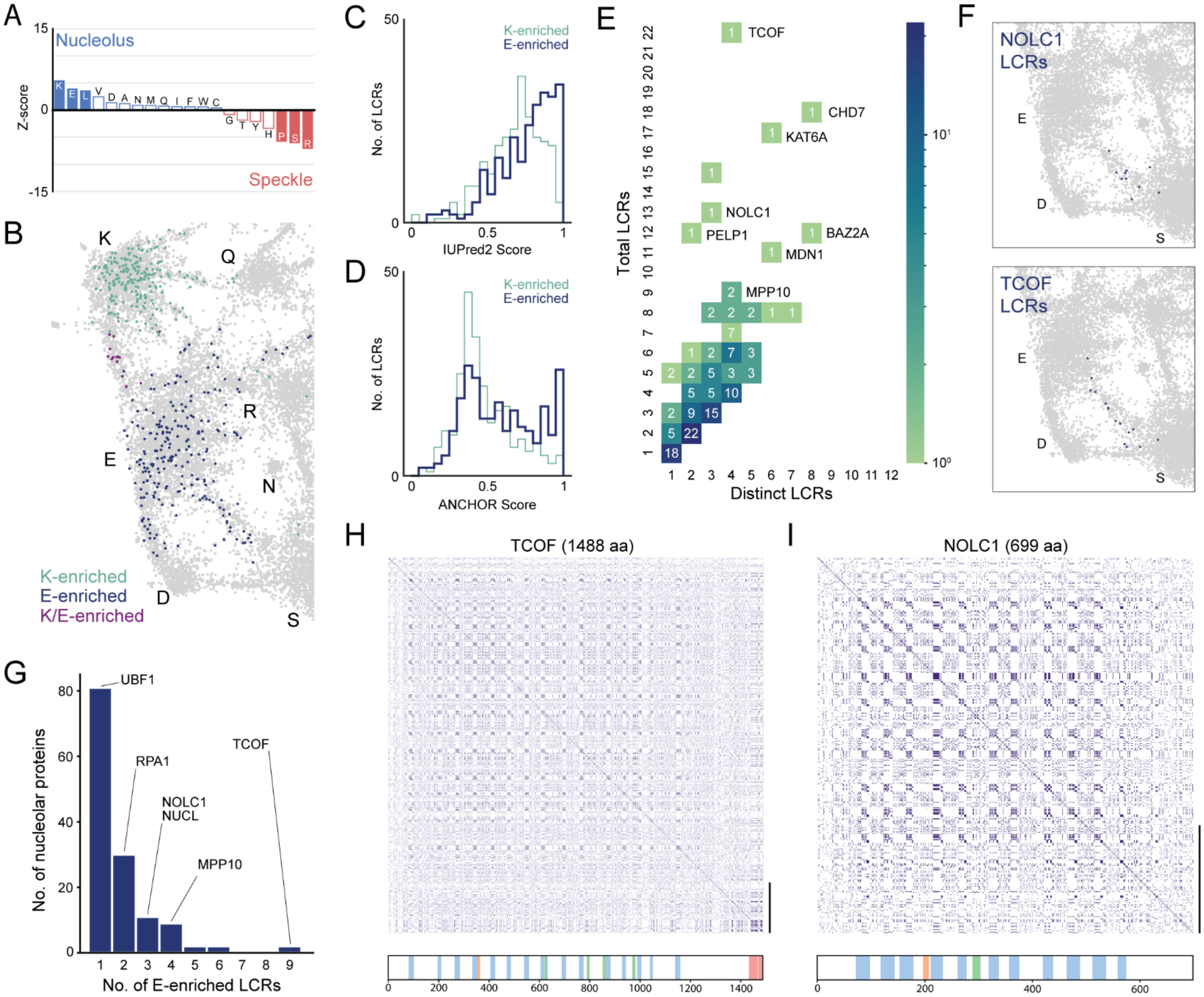


Figure 4: An integrated LCR map reveals the distribution and prevalence of E-rich LCRs among nucleolar proteins

A) Barplot of Wilcoxon rank sum tests for amino acid frequencies of LCRs of annotated nucleolar proteins compared to LCRs of annotated nuclear speckle proteins. Filled bars represent amino acids with Benjamini-Hochberg adjusted p-value < 0.001. Positive Z-scores correspond to amino acids enriched in LCRs of nucleolar proteins, while negative Z-scores correspond to amino acids enriched in LCRs of nuclear speckle proteins.

B) Nucleolar LCRs which are E-enriched (top 25% of nucleolar LCRs by E frequency), K-enriched (top 25% of nucleolar LCRs by K frequency), or K/E-enriched (both E- and K-enriched) plotted on close-up of K/E-rich regions of UMAP from Figure 3A.

C) Distribution of IUPred2 scores for K-enriched and E-enriched nucleolar LCRs.

D) Distribution of ANCHOR scores for K-enriched and E-enriched nucleolar LCRs.

E) Distribution of total and distinct LCRs for all nucleolar LCR-containing proteins in the human proteome with at least one E-enriched LCR. The number in each square is the number of proteins with that number of total and distinct LCRs and is represented by the colorbar. Several proteins with many LCRs are labeled directly to the right of their coordinates on the graph.

F) LCRs of NOLC1 and TCOF plotted on close-up of E-rich region of LCR UMAP of human proteome.

G) Distribution of the number of E-enriched LCRs for nucleolar proteins. Proteins with zero E-enriched LCRs are not included.

H) Dotplot of TCOF, and schematic showing positions of LCRs called from dotplot pipeline. Different colors in schematic correspond to different LCR types within TCOF. Scale bar on the right of dotplot represents 200 amino acids in protein length.

I) Dotplot of NOLC1, and schematic showing positions of LCRs called from dotplot pipeline. Different colors in schematic correspond to different LCR types within NOLC1. Scale bar on the right of dotplot represents 200 amino acids in protein length.

See also Figure 4 - figure supplement 1.

However, the second peak of this distribution had much higher anchor scores, approaching the maximal ANCHOR score of 1 (Figure 4D). This observation suggests that a subset of E-rich LCRs in the nucleolus may possess the ability to participate in modes of interaction different from K-rich LCRs, and raises the possibility that they fulfill non-overlapping roles in the structure of the nucleolus.

We sought to gain a better understanding of the contribution of E-rich LCRs to the nucleolus by looking at the type and copy number of these LCRs among the set of nucleolar proteins that possess E-enriched LCRs. Of the 319 LCR-containing nucleolar proteins, 137 had at least one LCR in the top 25th percentile of E frequency (Figure 4G). Moreover, the distribution of total vs distinct LCRs of nucleolar proteins containing E-enriched LCRs showed that many of these proteins were of the multiple-mixed type, with some even reaching 22 total LCRs across 4 distinct LCR types (Figure 4E). This allowed us to see the number of proteins which contain many of the same type of LCR. From this analysis, we can predict which proteins might act as clients and which might play scaffolding roles in the nucleolus with respect to interactions made by E-rich LCRs.

Proteins such as NOLC1 and TCOF have a high number of S and E-rich LCRs. In the total vs. distinct LCR comparison across the proteome, they were also high in total LCRs, while low in distinct LCRs (Figure 4E, G), suggesting that they might be scaffold-like proteins. NOLC1 and TCOF have a striking pattern of several evenly spaced E-rich LCRs, illustrated by their dotplots and UMAPs (Figure 4F, H, I, Figure 1 - figure supplement 4D, Figure 4 - figure supplement 1). Furthermore, these regions are required for NOLC1 and TCOF to interact with each other (Werner et al., 2015, 2018). Consistent with their scaffold-like property, NOLC1 and TCOF have been shown to recruit proteins important for key nucleolar functions, including key components of the Pol I transcription machinery (Werner et al., 2015, 2018). We find that most nucleolar E-enriched LCR-containing proteins contain only one or two of these LCRs (Figure 4G), and may act as clients in the nucleolus with respect to interactions mediated by E-rich

LCRs. These proteins include the catalytic component of RNA Pol I (RPA1) and the primary transcription factor for rDNA transcription (UBF1). The implication of E-rich LCRs in TCOF-NOLC1 interaction suggests that proteins with a high number of E-rich LCRs might enable nucleolar assembly through the recruitment of other E-rich containing proteins.

Thus, by integrating the sequences, features, and relationships of LCRs, such a unified view provides a framework for understanding of the role of LCRs in higher order assembly.

An expanded map of LCRs across species

The spatial relationship between regions of LCR space and higher order assemblies raises questions about if there is a conserved relationship between LCR sequence composition and the functions and physical properties of their parent proteins. In species where the existence of a given assembly such as the nucleolus is conserved, is occupancy of the sequence space also conserved? Similarly, does emergence of a certain higher order assembly across evolution such as the extracellular matrix correlate with the occupancy of a certain region of sequence space? Conversely, many species have distinct higher order assemblies with different functions and physical properties from those in humans, such as plants and fungal cell walls. Do these assemblies occupy a region of sequence space that is distinct from that occupied in humans, or do they use a sequence space that is occupied in humans?

To answer these questions, we wanted to capture the entire breadth of LCR sequence space across species, so that we could concurrently compare how they occupy this sequence space. We applied our dotplot and dimensionality reduction approach to the proteomes of *E. coli*, *S. cerevisiae*, *A. thaliana*, *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, and *H. sapiens*. This allowed us to simultaneously compare between prokaryotes and eukaryotes, among fungi, plants, and animals, and across metazoans.

We first generally compared the number of LCRs and LCR-containing proteins between species. At our stringent false discovery rate of 0.002, *E. coli* did not have any LCRs, while other species had many LCRs. So, we relaxed the FDR for *E. coli* to 0.05 to allow further analysis. While E. coli and S. cerevisiae had fewer LCRs, multicellular species had more, with both the number of LCRs and LCR-containing proteins increasing across metazoans (Figure 5 - figure supplement 1A, B). In particular, *D. melanogaster* had the most LCRs and highest proportion of LCR-containing proteins of the species analyzed Figure 5 - figure supplement 1A, B), consistent with previous studies (Huntley and Clark, 2007). The distribution of entropies for LCRs varied between species, but were all significantly lower than that of length-matched, randomly-sampled sequences from their respective proteomes (Figure 5 - figure supplement 1C).

We generated a map of the full breadth of LCR sequence space across these species (Figure 5A, Figure 5 - figure supplement 2), which we first used to examine the general distribution of LCRs of different species. The unicellular species, which did not have many LCRs overall, were absent in many of the spaces (Figure 5 - figure supplement 3). Most LCRs from *E. coli* were rich in hydrophobic residues, with few LCRs spread across polar amino acid clusters and almost none in charged amino acid clusters (Figure 5 - figure supplement 3). LCRs from *S. cerevisiae* were present but sparse in most clusters. In particular, they were relatively absent from the G, P, and G/P rich clusters, and abundant in the N-rich cluster. Among multicellular species, the larger clusters were mostly occupied, with some differences in the degree of occupancy in spaces such as the G/P, N, H and Q-rich clusters (Figure 5A, Figure 5 - figure supplement 3). In a few cases, entire clusters were specific to a species, such as the M-rich cluster predominantly occupied by *A. thaliana* and *D. rerio* (Figure 5 - figure supplement 2, 3), or the T/H-rich cluster specific to *D. rerio* (Figure 5 - figure supplement 2, 3, bottom left corner of map).

One set of regions with differences in LCRs between different species were the C-rich clusters. The C-rich regions separated into different clusters across the expanded map. While the main C-rich cluster contained LCRs from each multicellular species, various small C-rich

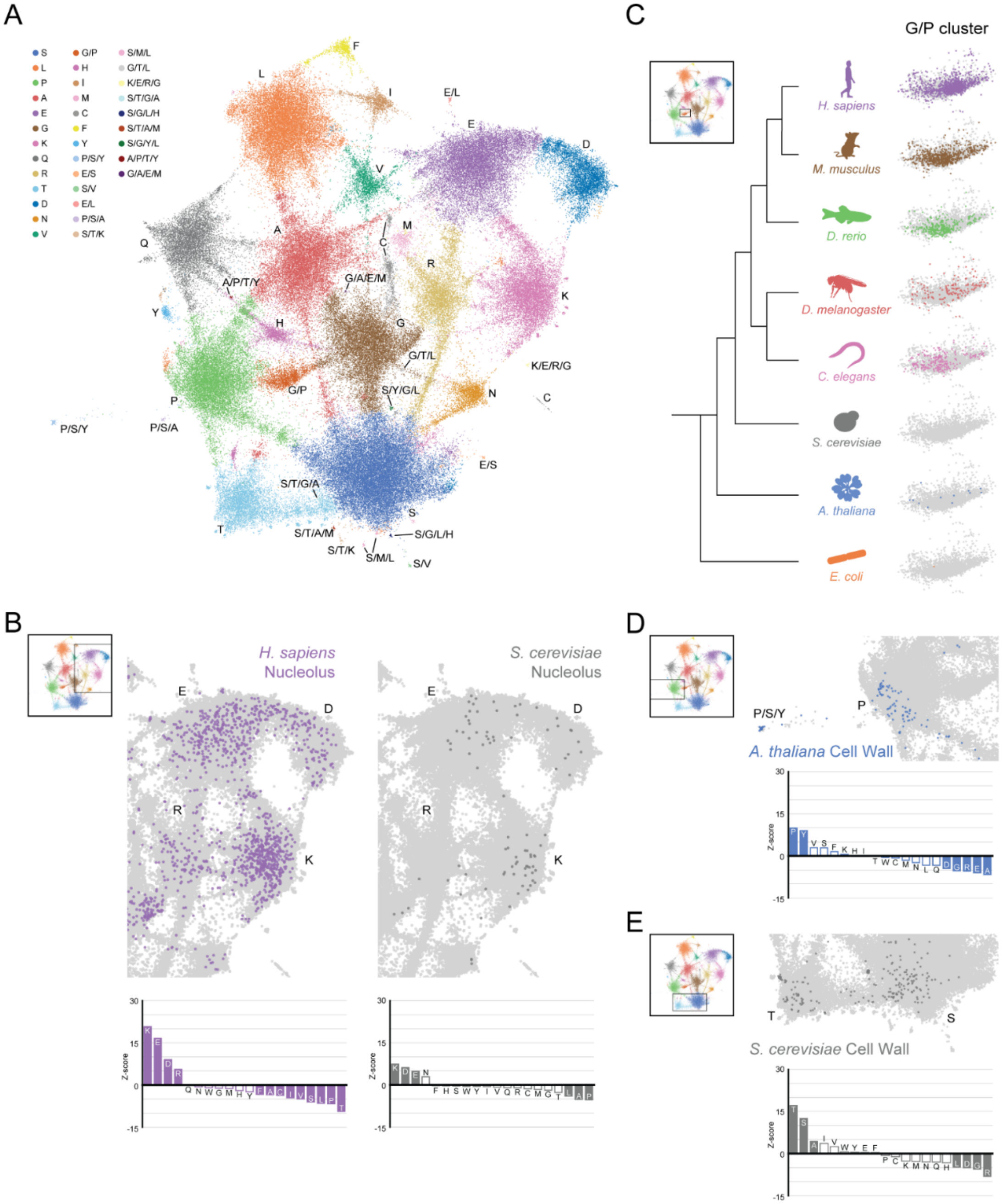


Figure 5: The conservation and emergence of higher order assemblies is captured in an expanded LCR map across species

A) UMAP of LCR compositions for all LCRs in the human (*H. Sapiens)*, mouse (*M. musculus*), zebrafish (*D. rerio*)*, fruit fly* (*D. melanogaster*)*,* worm (*C. elegans*), Baker's yeast (*S. cerevisiae*)*, A. thaliana, and E.coli* proteomes. Each point is a single LCR and its position is based on its amino acid composition (see Methods for details). Leiden clusters are highlighted with different colors. Labels indicate the most prevalent amino acid(s) among LCRs in corresponding leiden clusters.

B) Close-up view of UMAP in (A) with LCRs of human nucleolar (left) and yeast nucleolar (right) proteins indicated. (Bottom) Barplot of Wilcoxon rank sum tests for amino acid frequencies of LCRs of annotated human nucleolar proteins (left) and yeast nucleolar proteins (right) compared to all other LCRs in the UMAP (among all included species). Filled bars represent amino acids with Benjamini-Hochberg adjusted p-value < 0.001.

C) Close-up view of G/P-rich cluster from UMAP in (A) across species as indicated. LCRs within the G/P-rich cluster from each species are colored by their respective species. Species and their LCRs in the G/P-rich cluster are organized by their relative phylogenetic positions.

D) Close-up view of UMAP in (A) with LCRs of *A. thaliana* cell wall proteins indicated. Barplot of Wilcoxon rank sum tests for amino acid frequencies of LCRs of annotated *A. thaliana* cell wall proteins compared to all other LCRs in the UMAP (among all included species). Filled bars represent amino acids with Benjamini-Hochberg adjusted p-value < 0.001.

E) Same as (D) but with LCRs of *S. cerevisiae* cell wall proteins.

See also Figure 5 - figure supplement 1,2,3,4,5,6.

clusters exist, such as the one positioned left of the large Q-rich cluster (C/Q-rich; Figure 5 - figure supplement 2, 4A) and above the main C-rich cluster toward the V-rich cluster (C/V-rich; Figure 5 - figure supplement 2, 4B). These clusters were specific to *C. elegans* and *D. rerio*, respectively (Figure 5 - figure supplement 3, 4A, B), highlighting how specific LCR compositions, even when frequent amino acids are shared, can be specific to different metazoans.

Many of the spatial relationships between clusters were maintained in this expanded LCR sequence space. For example, certain bridges were observed in several species, while a few bridges were predominantly occupied in specific species, indicating that amino acid combinations in LCRs differed between species. For example, the H-rich cluster and its bridge connecting to the Q-rich cluster was expanded in *D. melanogaster* (Figure 5 - figure supplement 2, 3, 4C). LCRs in the Q/H-rich bridge have generally higher ANCHOR scores than those in the main Q-rich cluster (Figure 5 - figure supplement 5B), suggesting that H residues are combined with varying degrees of Q residues to allow for more nuanced differences in LCR properties.

The nuanced differences in LCR sequences and physical properties (Figure 5 - figure supplement 5) which we could detect in our expanded LCR map enabled us to look more deeply into how this LCR sequence space relates to the functions of those LCRs between species.

Conserved and diverged higher order assemblies are captured in LCR sequence space

We wanted to see if higher order assemblies which were conserved or diverged between species corresponded to similarities and differences in the occupancy of LCR space. For example, we mapped nucleolar annotations from *S. cerevisiae* and *H. sapiens* to compare the occupancy of nucleolar LCRs in these species. The space occupied by nucleolar LCRs from yeast and human were both common to the K-rich cluster as well as the E/D-rich clusters (Figure 5B, Figure 5 - figure supplement 6A, B), suggesting that the compositions of LCRs participating in the nucleolus are conserved across a large evolutionary distance, including the

E-rich sequences we discussed above (Figure 4). Similarly, when comparing between speckle annotations for *A. thaliana* and *H. sapiens*, we found that the R/S-rich bridge between the R-rich and S-rich clusters was occupied for each (Figure 5 - figure supplement 6C-F). In areas in which higher order assemblies are conserved between species, occupancy of LCR sequence space is generally conserved.

Furthermore, changes in the LCR sequence space corresponded to differences in higher order assemblies, such as the extracellular matrix which occupied the G/P cluster in humans. While *E. coli*, *S. cerevisiae*, and *A. thaliana* had nearly no LCRs in the G/P cluster, this cluster was much more occupied in metazoans (Figure 5C), corresponding with the emergence of collagens, a hallmark of the metazoan lineage (Hynes, 2012). This difference in G/P occupancy could not be explained by differences in the total number of LCRs in these species, since *A. thaliana* had more LCRs than *C. elegans* but much lower occupancy in the G/P cluster. When looking across metazoans, although this cluster was occupied in *C. elegans* and *D. melanogaster*, it was more heavily occupied in vertebrates. Many LCRs existed in this cluster in *D. rerio*, and even more in *M. musculus*, and *H. sapiens*, which spanned most of the space in this G/P cluster. Again, the difference in G/P cluster occupancy could not be explained by the total number of LCRs in each species, as *D. melanogaster* had more LCRs than all of the vertebrates, but lower G/P occupancy (Figure 5C, Figure 5 - figure supplement 3). The gradual differences in occupancy of the G/P cluster between the metazoan species (Figure 5C) correlated with the expansion of the extracellular matrix across metazoans (reviewed in (Hynes, 2012)), highlighting that the LCR map traces the progression of a higher order assembly across evolution.

Across longer evolutionary distances, different species-specific higher order assemblies mapped to unique regions of LCR sequence space. LCRs of cell wall proteins of *A. thaliana*, for example, primarily mapped to the P-rich cluster and a nearby P/S/Y-rich cluster (Figure 5D, Figure 5 - figure supplement 6G), reflecting the set of hydroxyproline-rich cell wall proteins which include extensins, arabinogalactan proteins (AGPs) and proline-rich proteins (PRPs). Extensins, which have SPPPP motifs, are known to be important scaffolds for the assembly of the cell wall, in which they are thought to form self-assembling networks to organize pectin (Cannon et al., 2008; Sede et al., 2018). In *S. cerevisiae*, LCRs of cell wall proteins mapped to the S-rich and T-rich clusters (Figure 5E, Figure 5 - figure supplement 6H), which included flocculation proteins. These S- and T-rich LCRs are often sites for O-mannosylation in mannoproteins, which is crucial for the integrity of the cell wall (Gentzsch and Tanner, 1996; González et al., 2012; Neubert et al., 2016).

Our approach allowed us to answer several general questions about the relationships between the sequence space occupied by LCRs and their functions. Firstly, we show that when a given assembly is conserved, occupancy of the corresponding LCR sequence space is also conserved. Secondly, the emergence of a higher order assembly can correspond to the population of a previously unoccupied sequence space. Finally, higher order assemblies with different physical properties occupy different regions of sequence space, even when they fulfill similar roles in their respective species. While these principles may not always hold for every sequence space or assembly, they may guide how we interpret the spaces and assemblies which have yet to be explored.

A teleost-specific T/H cluster contains scaffold-like proteins

Given the relationship between some regions of LCR sequence space and higher order assemblies, we looked for previously undescribed regions of sequence space which are differentially occupied across species.

We found various species-specific regions which lacked detailed annotations, one of which was the T/H-rich cluster specific to *D. rerio* (Figure 6A, Figure 5 - figure supplement 3). Many of the LCRs in this cluster included direct TH repeats (Figure 6A), making this a cluster of particular interest because these amino acid residues may have properties of mixed-charge

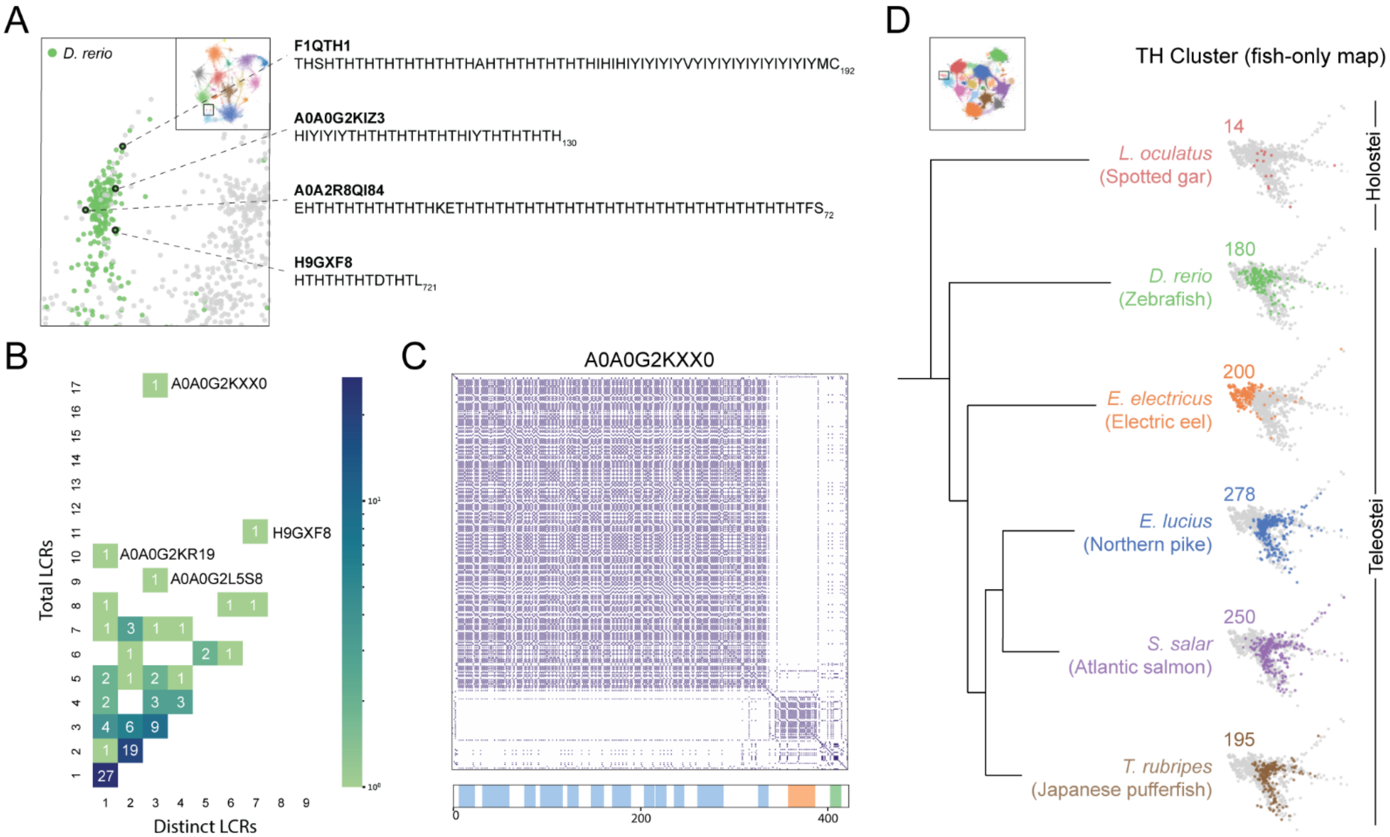


Figure 6: A conserved, teleost-specific T/H rich cluster exhibits signatures of higher order assemblies

A) Close up of T/H-rich region in UMAP shown in Figure 5A. LCRs of *D. rerio* are indicated in green, LCRs of all other species in UMAP are indicated in grey. Specific LCRs are circled and the dotted lines point to their parent protein and sequences (right). For all LCRs shown, the subscript at the end of the sequence corresponds to the ending position of the LCR in the sequence of its parent protein.

B) Distribution of total and distinct LCRs for all *D. rerio* proteins with at least one LCR in the T/H-rich region. The number in each square is the number of proteins with that number of total and distinct LCRs and is represented by the colorbar. Several proteins with many LCRs are labeled directly to the right of their coordinates on the graph.

C) Dotplot of A0A0G2KXX0, the *D. rerio* protein in the T/H-rich region with the largest number of total LCRs. Schematic showing positions of LCRs called from dotplot pipeline are shown below. Different colors in schematic correspond to different LCR types within A0A0G2KXX0.

D) T/H-rich cluster in UMAP generated from LCRs in proteomes of zebrafish (*D. rerio*), Spotted gar (*L. oculatus*), Electric eel (*E. electricus*), Northern pike (*E. lucius*), Atlantic salmon (*S. salar*), and Japanese pufferfish (*T. rubripes*). LCRs within the T/H-rich cluster from each species are colored by their respective species. The number above each UMAP cluster is the number of LCRs from each species inside that cluster. Species and their LCRs in the T/H-rich cluster are organized by their relative phylogenetic positions and members of Teleostei and Holostei are indicated.

See also Figure 6 - figure supplement 1.

domains under certain conditions. Given the phosphorylation state of threonine and protonation of histidine at certain pH levels, T/H-rich LCRs may behave like other LCRs composed of mixed charges, such as K/E, R/D, R/E, and R/S, many of which are known to form higher order assemblies (Greig et al., 2020). Therefore, we decided to further investigate this T/H-rich cluster, which contained 97 proteins with T/H-rich LCRs. To see if there could be signatures of higher order assemblies in this cluster, we analyzed the total vs. distinct LCRs to look for proteins which may be more client or scaffold-like in terms of their LCR relationships. This analysis showed that proteins with T/H-rich LCRs have a wide distribution of total and distinct LCRs (Figure 6B), in which many proteins exist with a low copy of LCRs and few proteins exist with a high copy of LCRs. Of particular interest were proteins with a high number of T/H LCRs. Such proteins could be similar to proteins like SRRM2, which together with another nuclear speckle protein SON, scaffold the nuclear speckle (Sharma et al., 2010; Fei et al., 2017; Ilik et al., 2020). The presence of scaffold-like proteins was confirmed by subsequently checking the dotplots of these proteins and their LCR sequences. For example, in the plot of total vs. distinct LCRs for the T/H-rich cluster, protein A0A0G2KXX0 had 17 total LCRs and only 3 distinct LCR types. Of these, 15 were T/H-rich LCRs, with only 1 LCR in each of the other distinct types (Figure 6C), suggesting that T/H LCRs, and all of the properties which come with a T/H composition, exist in high valency, scaffold-like proteins.

Given that we identified the T/H-rich LCR in zebrafish, we wanted to see if this composition of LCRs was generally conserved in fishes. To test this, we used our dotplot and UMAP approach to identify and cluster LCRs from a range of fishes from the clade Actinopterygii (Hughes et al., 2018), to which zebrafish belongs. The six species we analyzed were zebrafish, electric eel, northern pike, Atlantic salmon, Japanese pufferfish, and spotted gar (*D. rerio*, *E. electricus*, *E. lucius*, *S. salar*, *T. rubripes*, and *L. oculatus*, respectively). Of these fishes, all but the spotted gar substantially occupied the T/H-rich cluster. The fishes which heavily occupied the T/H-rich cluster each contained ~200 T/H LCRs, while the spotted gar was only lightly occupied with 14 LCRs (Figure 6D, Figure 6 - figure supplement 1A). Even when accounting for the total number of LCRs of each species, the spotted gar had the lowest percentage of LCRs in the T/H cluster (Figure 6 - figure supplement 1A, B). This difference in occupancy of the T/H cluster correlated exactly with evolutionary relationships between these fishes. Those containing T/H-rich LCRs belonged to Teleostei, while the spotted gar, which did not, belonged to Holostei, a group which diverged from Teleostei in Actinopterygii. Moreover, the seven species we analyzed outside of Actinopterygii did not have T/H-rich LCRs either (Figure 6A, Figure 5 - figure supplement 3), which strongly suggests that T/H-rich LCRs are indeed teleost specific. The finding that T/H-rich LCRs are conserved in teleosts, one of the largest clades of fishes, suggests that these conditionally mixed-charge LCRs may play a conserved, important role in these species.

Our approach was able to unearth conserved LCR compositions, with scaffold-like distributions within their parent proteins. These results not only demonstrate the existence of unexplored LCRs with signatures of higher order assemblies, but also highlight the ability of our approach to systematically explore the vast diversity of proteins across species.

DISCUSSION

Here, we have established a systematic approach to study LCRs, providing a unified view of how the sequences, features, relationships and functions of LCRs relate to each other. This unified view enabled us to gain insight into the role of LCRs in multivalent interactions, higher order assemblies, and organismal structures. Moreover, this framework for understanding LCRs raises fundamental questions about how LCRs encode their functions.

How can low complexity sequences capture the diversity of LCR function?

While the functions of proteins are encoded in their sequence, it has been difficult to assign functions to LCRs. Any mapping between LCR function and sequence space presents a question of how the many disparate functions of LCRs can exist in a space which only employs a few amino acids at a time.

In our LCR map, we find that natural LCRs distribute across a continuum of sequence space. Such nuanced differences in amino acid composition might enable similarly nuanced differences in the functions they encode. One known example of such nuanced LCR function is in the acidic LCR of G3BP1, which interacts with and inhibits its RNA-binding RGG LCR (Guillén-Boixet et al., 2020; Yang et al., 2020). This inhibitory activity of the acidic LCR is independent of the acidic LCR's primary sequence, and is abolished by substitution of negatively charged glutamic acid for neutral glutamine residues (Yang et al., 2020). These results suggest that gradual changes in the ratio of glutamine to glutamic acid may alter the inhibitory activity of such an LCR. Given that we observe a bridge connecting the E and Q clusters, such a range in activity may exist across proteins in the human proteome. We observe various other bridges, highlighting that meaningful functional differences may exist in the nuanced compositional differences of naturally occurring LCRs.

Differences in composition also imply differences in sequence. It follows that functional consequences downstream of sequence, such as post-translational modifications, can be affected by differences in composition. We have shown that several bridge-like connections exist between the clusters for serine and other amino acids in the LCR map. One well understood kinase, CK2, binds and phosphorylates serines in acidic contexts (Rusin et al., 2017), changing the physical properties of this sequence by making it more negatively charged. Interestingly, bridge-like connections exist between both S and D, and S and E in the LCR map, raising the possibility that their physical properties can be regulated to different extents by CK2. Notably, assembly of NOLC1 and TCOF, which have many LCRs in the bridge between S and E, is known to be regulated by CK2 phosphorylation (Werner et al., 2015, 2018). Similarly, a bridge-like connection exists between E and K, two oppositely charged residues. Lysine acetylation eliminates the positive charge in lysine, and has been shown *in vivo* to predominantly occur in K/E/D-rich contexts (Lundby et al., 2012). Interestingly, K/E/D-rich sequences are prevalent in nucleolar LCRs, and evidence points to lysine acetylation preventing nucleolar integration (Fantini et al., 2010; Lirussi et al., 2012). Together, these observations suggest that different nucleolar LCRs along the K/E bridge may be differentially regulated by lysine acetylation, potentially affecting the structure of the nucleolus. Thus, differences in post-translational modifications may provide an additional layer by which LCRs can encode biological functions.

While these examples highlight potential consequences of nuanced differences in LCRs in certain regions of LCR space, the functional consequences of nuanced differences in other regions of LCR space can now be systematically studied with our approach.

Implications of bridges between certain amino acids in LCR space

Looking more generally at the LCR maps, the presence or absence of certain bridges connecting clusters may correspond to informative relationships between pairs of amino acids. We found that various bridges exist in the map, including the bridges between L and each of I, F, and V, the K - E - D axis, and the G/P and R/S bridges.

Some of these bridges represent mixtures of similar residue properties, such as hydrophobic or negatively-charged amino acids. These findings are consistent with the hypothesis that some sets of amino acids with similar physical properties may be redundant, and thus varying combinations of them are not selected against. Interestingly, while R and K are both positively-charged, basic residues, the region between these clusters was poorly populated, suggesting that these residues may not always be interchangeable in LCRs. This is consistent with known differences between R and K, such as the ability of R to participate in stacking interactions. In fact, recent evidence showed that the physical properties of R and K substantially differ, while the difference between D and E is much more subtle (Fossat et al., 2021; Greig et al., 2020; Wang et al., 2018). Thus, while co-occurrence of similar amino acids may not be entirely surprising, a lack of co-occurrence between seemingly similar amino acids may point towards interesting differences between them.

Likewise, while dissimilar amino acids may not often co-occur in LCRs, the presence of bridges with dissimilar amino acids may represent combinations which have emergent functions. It has not escaped our notice that R/S and G/P are combinations of dissimilar amino acids which all correspond to functional, conserved higher order assemblies--the speckle and extracellular matrix. In these cases, it is known how the combinations of amino acids may enable emergent properties, such as mixed charge domains (Greig et al., 2020) or tight-packing polyproline helices (Ramachandran and Kartha, 1955; Rich and Crick, 1955; Cowan and McGAVIN, 1955). However, certain combinations exist in which the properties are not well understood, such as H/Q, N/S or T/H, among others. Thus, we hypothesize that the existence of bridges between dissimilar amino acids may correspond to LCRs with specific emergent properties. These types of LCRs represent open, unexplored regions of LCR space for which the relationship between sequence and function has yet to be determined.

A unified LCR map relates disparate higher order assemblies across species

The ability of the LCR map to capture certain higher order assemblies raises questions of what the LCRs in other parts of the map may tell us about their functions. While we do not interpret that all LCRs must be involved in higher order assembly, the observation that LCRs of certain higher order assemblies populated different regions of same sequence space allows us to consider if there are similarities among these assemblies which give us insight into LCR function. For example, the nucleolus is a liquid assembly of protein, RNA, and DNA essential for ribosome biogenesis, while the extracellular matrix is a solid/gel-like assembly of glycoproteins scaffolded by long collagen fibers. Despite the physical differences in each of these assemblies, the LCRs are essential for their higher order assembly. The K-rich LCRs of nucleolar proteins such as RPA43 are required for their higher order assembly and integration into the nucleolus, while the G-P-P motif-containing LCRs in various collagens form key assemblies in the ECM (Timpl et al., 1981; Mould and Hulmes, 1987; Hansen and Bruckner, 2003) and the G/V/P-rich LCRs of elastin assemble to provide ECM elasticity (Urry et al., 1974; Rauscher et al., 2006). Although these examples are vastly different in physical properties, a common theme is that the LCRs enable the integration and assembly of various biomolecules in biological structures.

If this is the case, we may gain insight into the structures and organizations of species by comparing the differences in the sequence space occupied by their LCRs. One fruitful comparison was between the human extracellular matrix and plant cell wall, which greatly affect cellular organization in their respective species. Each of these have taken a role in the extracellular space, yet they have different chemical compositions, structures, and proteins. The LCR spaces occupied by these proteins are unique for each extracellular assembly, corresponding to differences in the specific interactions and processes required for their formation. While human ECM and plant cell wall proteins both occupy spaces which have a substantial presence of prolines, the specific differences in the regions they occupy give insight into their unique properties. For example, LCRs of ECM proteins occupy the G/P-rich cluster and the presence of glycines in ECM collagen proteins is crucial for tight packing of helices to form the collagen triple helix (Beck et al., 2000), which is the basis for higher order assembly of most of the tissues in the human body. On the other hand, while plant cell wall proteins also use polyproline II helices, these P-rich LCRs occupy a different region in the map from LCRs in the ECM. Moreover, plant cell wall proteins contain different P-rich LCR compositions which delineate between extensins and other proline-rich cell wall proteins. Such differences, in whether or not the contiguous prolines are interrupted, have been proposed to explain the origins of plant cell wall proteins with different properties (Kieliszewski and Lamport, 1994; Lamport et al., 2011), supporting the idea that functional divergence of LCRs can occur through relatively local differences in sequence space. Our map of LCR sequence space not only shows differences in the LCRs of ECM and cell wall proteins, but also shows that two different groups of LCRs exist among cell wall proteins. Thus, this view of LCR sequence space captures key sequence determinants of higher order assemblies, even when the differences between them are subtle.

As the functions of other regions of LCR sequence space are uncovered or mapped, such as the teleost-specific T/H-rich cluster we identified, species with different higher order assemblies and cellular organizations may be found to occupy similar or different spaces. By viewing these higher order assemblies from the perspective of LCRs, we suggest that the principles which explain extracellular and organismal structure may be similar to the principles which explain membraneless subcellular compartmentalization. For now, we can only speculate that they may not be isolated processes, but different regions across a unified LCR space.

ACKNOWLEDGMENTS

We thank all members of the Calo lab, as well as Eeshit D. Vaishnav, Connor Kenny, Christopher B. Burge, Amy E. Keating, and David P. Bartel for helpful discussions and feedback on the manuscript. We would also like to thank the Swanson Biotechnology Center Microscopy and Barbara K. Ostrom (1978) Bioinformatics core facilities in the Koch Institute at MIT. Funding: National Institute of General Medical Sciences R35GM142634 to EC, The Ludwig Center at MIT to NJ, NIH Pre-Doctoral Training Grant T32GM007287 to BL and The Phew Charitable Trust to EC.

AUTHOR CONTRIBUTIONS

NJ and BL - conceptualized the study, developed the dotplot pipeline, performed all analyses and experiments, wrote and edited the manuscript. EC - supervised the study, acquired funding, edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests

MATERIALS AND METHODS

Experimental Methods

Plasmids

Note: All RPA43 constructs (both Mammalian expression and bacterial expression) contain a GSAAGGSG peptide linker between GFP and RPA43.

Mammalian expression constructs

Plasmid	Source	Identifier
pcDNA3.1(+) meGFP - RPA43	This paper	RP104 (RPA43 WT)
pcDNA3.1(+) meGFP - RPA43 (ΔK223-P234, P274-Q284, H306-H315)	This paper	RP105 (RPA43 ΔK1,2,3)
pcDNA3.1(+) meGFP - RPA43 (ΔH306-H315)	This paper	RP108 (RPA43 ΔK3)
pcDNA3.1(+) meGFP - RPA43 (ΔK223-P234, P274-Q284)	This paper	RP109 (RPA43 ΔK1,2)
pcDNA3.1(+) meGFP - RPA43 (ΔK223-P234, H306-H315)	This paper	RP110 (RPA43 ΔK1,3)
pcDNA3.1(+) meGFP - RPA43 (ΔP274-Q284, H306-H315)	This paper	RP111 (RPA43 ΔK2,3)
pcDNA3.1(+) meGFP - RPA43 (ΔK223-P234)	This paper	RP112 (RPA43 ΔK1)
pcDNA3.1(+) meGFP - RPA43 (ΔP274-Q284)	This paper	RP113 (RPA43 ΔK2)

Bacterial expression and purification constructs

Plasmid	Source	Identifier
pGEX6p1 GST-SBP-eGFP - RPA43 (E209-end)	This paper	RP106 (RPA43 C-term WT)
pGEX6p1 GST-SBP-eGFP - RPA43 (E209-end) (ΔK223-P234, P274-Q284, H306-H315)	This paper	RP107 (RPA43 C-term ΔK1,2,3)

Cell lines

HeLa cells were obtained from ATCC. Cells tested negative for mycoplasma.

Cell Culture

HeLa cells were cultured in 5% CO2 on cell culture-treated 10 cm plates (Genesee Scientific, 25-202) in Dulbecco's Modified Eagle Medium (DMEM, Genesee Scientific, 25-500) supplemented with 10% Fetal bovine serum (FBS, Gemini Bio-products, 100-106) and 1% Penicillin/Streptomycin (Gibco, 10378-016). Cells were split 1:10 every 3 days by using trypsin (Gibco, 25200072).

Protein purification

All protein purification constructs used were cloned into a version of the pGEX-6P-1 plasmid modified to include eGFP followed by a GSAAGGSG peptide linker. All RPA43 C-terminal (amino acid positions 209-338) fragments were fused to the C-terminus of this linker. After sequence verification, plasmids encoding the final constructs were transformed into 20 µL of Rosetta (DE3) competent cells (EMD Millipore, 70954) and grown overnight at 37°C in 5mL LB containing 100 µg/mL Ampicillin (Fisher Scientific, BP1760) and 34 µg/mL Chloramphenicol (Fisher Scientific, BP904-100). Overnight cultures were added to 250 mL of Superbroth containing Ampicillin and Chloramphenicol (same concentrations as above) and grown at 37°C to an OD_{600} ~ 0.6-0.8. Cultures were cooled to 4°C, expression of proteins was induced by the addition of IPTG to a final concentration of 0.5mM, and cultures were grown on a shaker overnight at 15°C. Cells were pelleted by centrifugation for 35 minutes at 9790 x g at 4°C, and pellets were frozen at -80°C.

Pellets were thawed and lysed on ice in 15 mL lysis buffer containing freshly added lysozyme and benzonase prepared according to manufacturer instructions (Qiagen Qproteome Bacterial Protein Prep Kit, Cat. No. 37900), 1mM PMSF (ThermoFisher Scientific, 36978), and 1.5 cOmplete mini EDTA-free protease inhibitor cocktail tablets (Millipore Sigma, 11836170001) per 250 mL culture. Lysates were incubated on ice for 20 minutes with occasional inversion, and sonicated for 5 cycles (30 secs on, 30 secs off, high intensity) on a Bioruptor 300 at 4-6°C. Cellular debris and unlysed cells were pelleted by centrifugation for 30 minutes at 12,000 x g at 4°C.

Cleared lysates were syringe filtered (Pall Life Sciences, Product ID 4187) and added to 0.625 mL of glutathione-sepharose beads (GE Healthcare, GE17-0756), which were pre-equilibrated in equilibration buffer (1X PBS, 250mM NaCl, 0.1% Tween-20) by performing four 10 mL washes for 5 minutes each with end-over-end rotation at 4°C. After addition of filtered lysates, beads were incubated for 2 h at 4°C on an end-over-end rotator. Beads were centrifuged at 500 x g for two minutes and unbound lysate was removed. Beads were washed three times for 10 minutes with 10 mL cold wash buffer (150mM NaCl, 10mM $MgCl_2$, 10mM Na_2HPO_4, 2mM ATP) at 4°C with end-over-end rotation. Three to five 0.5mL elutions were performed at 4°C on a nutator, with freshly prepared elution buffer (100mM TRIS pH 8, 20mM reduced glutathione, 5mM EDTA pH 8, 2mM ATP), each for 10 minutes. Elutions were collected, concentrated, and subsequently buffer exchanged into protein storage buffer (25 mM Tris pH 7.5, 150 mM KCl, 0.5 mM EDTA, 0.5 mM DTT freshly added, 10% glycerol) using Amicon Ultra-0.5 centrifugal filter units with a 10kDa cutoff (Millipore Sigma, UFC5010). Protein concentrations were determined, after which proteins were diluted to 100 μM in protein storage buffer, aliquoted, and stored at -80°C.

Droplet formation assays

Droplet formation assays were performed in droplet formation buffer (50 mM Tris pH 7.0, 150 mM NaCl), in the presence of a final concentration of 10% PEG-8000 (New England Biolabs, B1004), in a total volume of 12 μL. Droplet formation was initiated by the addition of 1 μL of purified protein (in protein storage buffer) to 11 μL of pre-mixed Droplet formation buffer and PEG-8000 on ice (8.6 μL of Droplet formation buffer + 2.4 μL 50% PEG-8000). The final protein concentration in the reaction was 8.3 μM. After the addition of purified protein, the reaction was mixed by pipetting, 10 µL was loaded onto a microscope slide (Fisher Scientific, 12-544-2), and droplets were immediately imaged using a fluorescent microscope (Evos FL) at 40X magnification. Representative images were chosen for Figure 2.

Droplet formation assays were repeated over the course of about 6 months, with each replicate corresponding to the same experiment carried out on different days, using the same preparation of purified protein.

Immunofluorescence

Glass coverslips (Fisherbrand, 12-545-80) were placed in 24-well plates (Genesee Scientific, 25-107) and coated in 3 µg/mL of fibronectin (EMD Millipore, FC010) for 30 minutes at room temperature. HeLa cells were seeded in each well at 50,000 cells per well. 24 hours after seeding, the cells were transfected with GFP-tagged protein plasmids using Lipofectamine 2000 (Invitrogen, 11668027). Each well was transfected using 100 ng of plasmid and 1 µL of Lipofectamine 2000 in a total of 50 µL of OptiMEM (Gibco, 31985070) according to the Lipofectamine 2000 instructions. Cells on glass coverslips were collected for immunofluorescence 48 hours after transfection. Cells were collected by washing with 1x PBS (Genesee Scientific, 25-508) and fixation in 4% paraformaldehyde (PFA) for 15 minutes at room temperature, followed by another 3 washes with 1x PBS. Cells were permeabilized and blocked by incubation in blocking buffer (1% BSA (w/v), 0.1% Triton X-100 (v/v), 1x PBS) for 1 hour at room temperature. Coverslips were then incubated overnight at 4°C in a 1:100 dilution of primary antibody (anti-MPP10, Novus Biologicals, NBP1-84341) in blocking buffer. After 3 washes with blocking buffer, coverslips were incubated for 2 hours in a 1:1000 dilution of secondary antibody (anti-rabbit, Invitrogen, 32260). Coverslips were washed 3 times with blocking buffer, then once with 1x PBS, and mounted on glass slides using ProLong Diamond antifade mountant with DAPI (Invitrogen, P36962). Slides were sealed using clear nail polish, allowed to dry, and stored at 4°C. Slides were imaged on a DeltaVision TIRF microscope using

100X oil immersion objective lens. The same set of exposure conditions (one exposure per channel) was used across all slides. Raw images were deconvoluted, from which a max projection image was generated. Deconvolution and max projection were performed using Deltavision SoftWoRx software. Displayed images were scaled such that the distribution of signal was representative. Image analysis was performed using Fiji (https://imagej.net/software/fiji/). For each transfected construct, representative cells were chosen. Cells that were excluded were cells that were not appreciably transfected, and cells that highly overexpressed the transfected constructs.

The immunofluorescence experiment was performed more than three times over the course of about 1 year, with each replicate corresponding to the same experiment carried out on different days.

External data

Proteome Datasets

Proteomes were downloaded from UniProt for all species analyzed (see table below). Every proteome was greater than 90% complete based on Benchmarking Universal Single-Copy Ortholog (BUSCO) assessment score for proteome completeness. One protein sequence was downloaded per gene in FASTA format. Thus, all protein names used in the manuscript are UniProt protein names (i.e. "NUCL" in "NUCL_HUMAN").

Species	**Proteome ID**	**Date accessed**
Homo sapiens	UP000005640	March 15, 2021
Mus musculus	UP000000589	March 15, 2021
Danio rerio	UP000000437	March 15, 2021
Drosophila Melanogaster	UP000000803	March 15, 2021
Caenorhabditis elegans	UP000001940	March 15, 2021

Saccharomyces cerevisiae	UP000002311	March 15, 2021
Arabidopsis thaliana	UP000006548	March 15, 2021
Escherichia coli	UP000000625	March 15, 2021
Electrophorus electricus	UP000314983	July 14, 2021
Esox lucius	UP000265140	July 14, 2021
Lepisosteus oculatus	UP000018468	August 5, 2021
Salmo salar	UP000087266	July 14, 2021
Takifugu rubripes	UP000005226	July 13, 2021

Higher order assembly annotations

Annotations for higher order assemblies were downloaded from Uniprot, based on their subcellular location annotations. Only entries which were Swiss-Prot reviewed (i.e. entry belongs to the Swiss-Prot section of UniProtKB) were included in the annotations. Annotations were accessed in FASTA format. Annotations for stress granule were taken from a published experiment (Jain et al., 2016). Stress granule protein sequences from the “Tier1” list of stress granule proteins were downloaded from UniProt in FASTA format.

Species	**Annotation**	**Date accessed**
Homo sapiens	Nucleus speckle (SL0186)	September 30, 2020
	Extracellular matrix (SL0111)	October 27, 2020
	Nucleolus (SL0188)	October 7, 2020
	Nuclear pore complex (SL0185)	May 6, 2021
	Centrosome (SL0048)	April 12, 2021
	PML body (SL0465)	October 8, 2020
	Stress granule (Jain et al., 2016)	May 18, 2021
Arabidopsis thaliana	Nucleus speckle (SL0186)	August 5, 2021
	Cell wall (SL0041)	June 17, 2021

Saccharomyces cerevisiae	Nucleolus (SL0188)	March 16, 2021
	Cell wall (SL0041)	June 17, 2021

Core approach

See Figure 1D for overview and flowchart. All code was written in Python 3. Run on Google Colaboratory or the Luria server at MIT. Python modules used were NumPy (1.20.1), BioPython (1.78), Pandas (1.2.3), Mahotas (1.4.11), SciPy (1.6.2), Scanpy (1.7.2), AnnData (0.7.5), NetworkX (2.3), Matplotlib (3.4.1), Seaborn (0.11.1). Code, dotplot module outputs, and other relevant files can be found on zenodo (https://doi.org/10.5281/zenodo.5555373).

Dotplot generation (Module 1)

Self-comparison dotplots of every protein sequence of every proteome were generated using a custom implementation to make dotplots in which every identically matching amino acid equals 1 and every non-matching position equals 0. For each dotplot, protein sequences from the proteome FASTA file were integer-encoded such that each of the 20 amino acids corresponds to a unique integer from 1 to 20, inclusive. For the null proteome, length-matched sequences were randomly generated from uniformly distributed integers from 1 to 20. A total of two arrays of this sequence x N, row-wise and column-wise, were generated, such that each array was a matrix of size N x N, where N is the protein sequence length. The two matrices were subtracted such that any identical amino acid matches equaled 0 and non-matches were non-zero. The final dotplot matrix was generated by replacing any 0 values with 1 and replacing any non-zero values with 0. Dotplot matrices were saved to .npz files using the file saving and compression implementation from NumPy. For images of dotplots, matrices were plotted directly.

LCR calling (Module 2, part 1)

LCRs were called by identifying high density regions in protein dotplots through classic image processing methods, such as kernel convolution, thresholding, and segmentation (Figure 1D).

To identify high density regions in dotplots, we performed kernel convolution on the dotplots with a uniform 10x10 kernel, which calculates a convolved pixel intensity value from 0 to 100 based on the number of dots in that window. This kernel relates to the minimum length of an LCR.

We used the convolved dotplots to determine this “high density” cutoff to define LCRs. Specifically, we used a false discovery rate (FDR)-based approach to threshold the convolved pixel intensities in a way that reliably identifies high density regions and treats the same sequence similarly regardless of the proteome it comes from. For a given proteome, we generated a background model by simulating an equally sized, length-matched ‘null proteome’, whose sequences were generated from a uniform amino acid distribution. Using a uniform amino acid distribution for the null proteome minimizes proteome-specific effects on whether a sequence is considered to contribute to a region of high density in a dotplot. Moreover, matching the lengths of the proteomes accounts for differences in the length distributions of proteins in different proteomes. We compared the distribution of convolved pixel intensities from the across all convolved dotplots of proteins in the real proteome with those from the null proteome and identified the lowest convolved pixel intensity which satisfied a stringent FDR of 0.002 (Figure 1D, Figure 5 - figure supplement 1). FDR was defined by the number of pixels from the null set which pass the threshold divided by the total number of pixels which pass the threshold (from the real and null sets combined). This threshold was then applied to every protein in the proteome to generate segmented dotplots, in which high-density regions (referred to as segmented regions) had values of 1 while other regions had values of 0. The positions from -4 and +5 of the boundaries of the segmented regions were included as the start and stop of the LCR to account for the convolution kernel size. The exception to this was LCRs which existed within that distance from the start or stop of a protein, in which the protein start or stop was designated the start or stop accordingly. Only segmented regions which intersected with the diagonal were called as LCRs.

LCR type and copy number determination (Module 2, part 2)

To computationally determine the types of LCRs and the copy number for each type, we determined the presence of segmented regions at the intersection between called LCRs in the segmented dotplot (Figure 1D, E). For each protein, we represented the LCRs as a network in which the LCRs were nodes and intersections between LCRs were edges (Figure 1D, E, Figure 1 - figure supplement 4). The total number of nodes equals the total number of LCRs in the protein. The number of connected components of this network equals the number of distinct LCR types in the protein. Therefore, the number of nodes within a given connected component equals the number of LCRs of that type. NetworkX (version 2.3) was used to calculate these values, and plot the network representation of LCR relationships within proteins.

Entropy calculation, random length-matched sequence sampling

Shannon entropy was calculated for each LCR sequence and a length-matched sequence which was randomly sampled from the respective proteome. Random length-matched sequence sampling was done by indexing the position of all proteins in the proteome from 1 to the length of the proteome (i.e. the sum of lengths of all proteins), and randomly selecting a position between 1 and the length of the proteome minus the length of the sequence of interest. The randomly sampled sequence was the sequence of the matched length, starting at the selected position. Shannon entropy for both the LCR and randomly sampled sequence was calculated using Scipy’s implementation.

Other LCR calling methods (SEG/fLPS)

LCRs were called with other methods, SEG (Wootton and Federhen, 1993) and fLPS (Harrison, 2017) for comparison.

SEG was run on the human proteome using 'default', 'intermediate', and 'strict' settings, as defined by the PLAtform of TOols for LOw COmplexity (PlaToLoCo) (Jarnot et al., 2020). Settings used from PlaToLoCo (http://platoloco.aei.polsl.pl/#!/help, accessed May 20, 2021) are restated here for completeness. 'Default': *W* = 12, *K1* = 2.2, *K2* = 2.5; 'Intermediate': *W* = 15, *K1* = 1.9, *K2* = 2.5 (Huntley and Golding, 2002); 'Strict': *W* = 15, *K1* = 1.5, *K2* = 1.8 (Radó-Trilla and Albà, 2012). From the output, we extracted the LCR coordinates for use in downstream entropy calculations. SEG was downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/seg/seg/ on May 20, 2021.

fLPS was run on the human proteome using 'default' and 'strict' settings, as defined by the PLAtform of TOols for LOw COmplexity (PlaToLoCo) (Jarnot et al., 2020), with a uniform background amino acid composition. Settings used from PlaToLoCo (http://platoloco.aei.polsl.pl/#!/help, accessed May 20, 2021) are restated here for completeness. 'default: m = 15, M = 500, t = 0.001, c = equal; 'strict': m = 5, M = 25, t = 0.00001, c = equal. From the output of fLPS, 'whole' rows were dropped in order to remove LCR calls covering the full length of a protein, which obscured LCR calls of subsequences of proteins. We then extracted the LCR coordinates for use in downstream entropy calculations. fLPS was downloaded from: https://github.com/pmharrison/flps/blob/master/fLPS.tar.gz on May 20, 2021.

Generation of LCR maps (UMAP dimensionality reduction, Leiden clustering)

LCR maps contained a 2 dimensional representation of different LCR amino acid compositions. For each LCR in the proteome, the amino acid composition was calculated as the frequency of each amino acid in the LCR, and was represented as a vector in 20-dimensional space. The 20-dimensional vectors of all LCRs were saved in AnnData format as an array in which rows were LCRs and columns were the amino acid frequencies. LCR maps were generated by dimensionality reduction from 20 to 2 dimensions using Scanpy's implementation of UMAP (random_state=73, n_components=2; n_neighbors=200 for Figure 5A and n_neighbours=default for Figures 3A and 6D; (McInnes et al., 2020; Wolf et al., 2018)). Amino acid distributions on the LCR map were generated by coloring each point on a color scale corresponding to the frequency of the amino acid represented. Leiden clustering was performed using Scanpy (random_state=73), and the most represented amino acids in each cluster was determined by looking at the amino acid distributions in each cluster by eye.

Annotation of LCR maps (higher order assemblies, biophysical predictions)

Annotation of LCRs belonging to higher order assemblies (see table above) was done by adding annotations to the AnnData object and coloring the LCRs using Scanpy's plotting implementation. Wilcoxon rank sum (MannWhitneyU) tests for amino acid enrichment in LCRs of higher order assemblies were performed using Scanpy. For the Wilcoxon rank sum tests comparing one annotation against all other LCRs, default settings were used. For the Wilcoxon rank sum tests comparing between two annotation sets of LCRs, one annotation set was set as the reference.

Biophysical predictions were calculated and mapped for all LCRs. For IUPred2A and ANCHOR2 predictions, which are context dependent, the scores at each position were calculated for full-length proteins in the proteome using a modified version of the official python script ((Mészáros et al., 2018); https://iupred2a.elte.hu/download_new, accessed May 31, 2021) to allow for batch predictions. LCR positions identified by our dotplot approach were used to extract the corresponding ANCHOR and IUPred2A scores for each position in each LCR. The mean ANCHOR and IUPred2A scores for each LCR were calculated and used to color the UMAP plot. The IUPred2A and ANCHOR2 scoring was run with the default 'long' setting and '-a' to include ANCHOR predictions. Kappa scores (Das and Pappu, 2013; Holehouse et al., 2017) for mixed-charge distribution were calculated for each LCR using the localCIDER package (version 0.1.19).

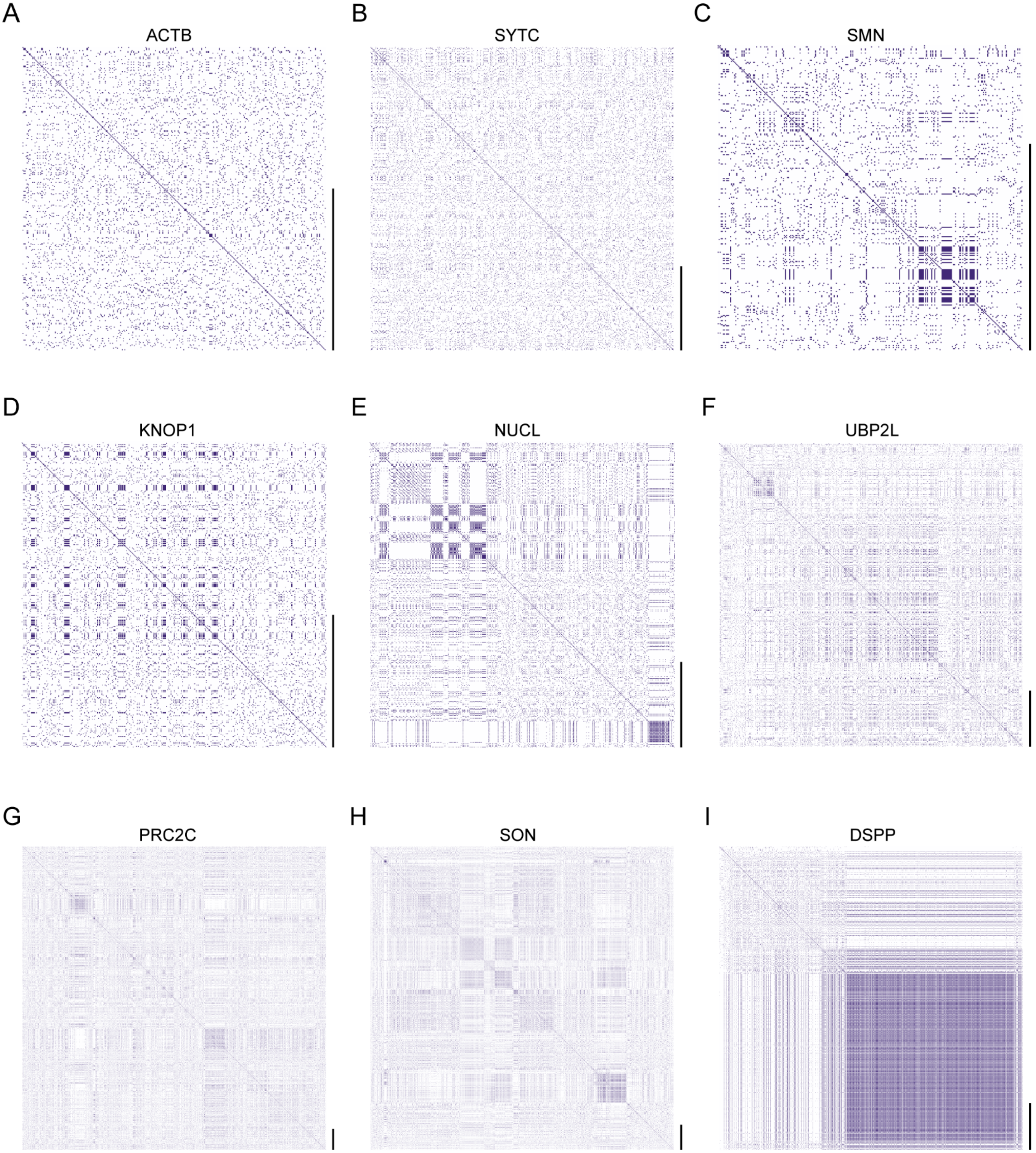


Figure 1 - figure supplement 1: Dotplots of various human proteins
Raw dotplot matrices for A) ACTB, B) SYTC, C) SMN, D), KNOP1, E) NUCL, F) UBP2L, G) PRC2C, H) SON, and I) DSPP. For all dotplots, the protein sequence lies from N-terminus to C-terminus from top to bottom, and left to right. Scale bars on the right of the dotplots represent 200 amino acids in protein length.

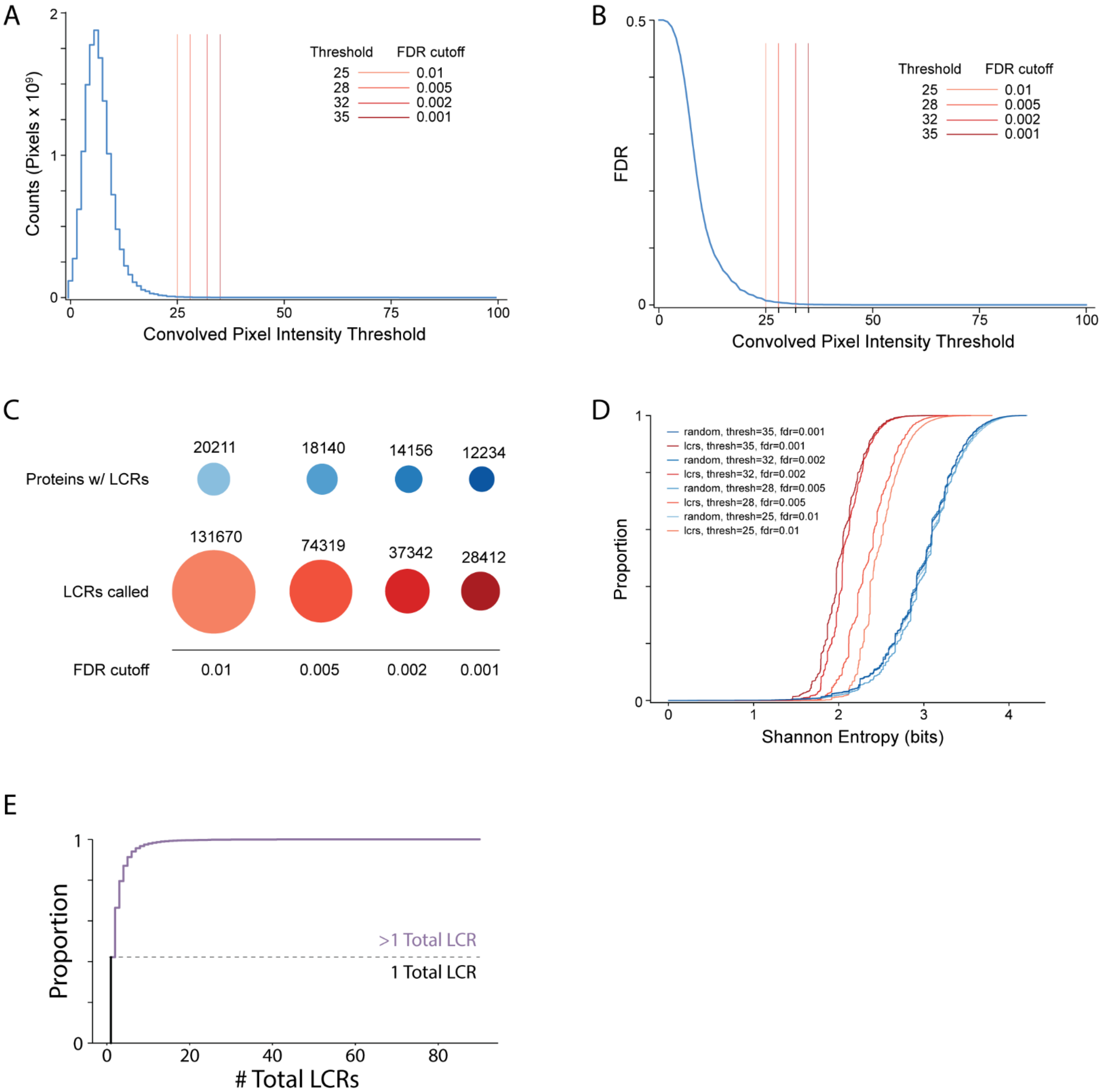


Figure 1 - figure supplement 2: Summary statistics from systematic dotplot analysis of human proteome

A) Histogram of convolved pixel intensities across dotplots of all proteins in the human proteome. Vertical lines indicate certain FDRs and their corresponding convolved pixel intensity thresholds. Four specific thresholds and their corresponding FDRs are labelled. FDR was defined by the number of pixels from the null set which pass the threshold divided by the total number of pixels which pass the threshold (from the real and null sets combined) (see Methods for details).

B) Plot of FDR vs. convolved pixel intensity threshold for dotplots of all proteins in the human proteome. Four specific thresholds and their corresponding FDRs are labelled.

C) The number of LCR-containing proteins and number of LCRs called from systematic dotplot analysis on the human proteome at different FDR cutoffs.

D) The cumulative distribution of Shannon entropies of LCRs identified using the dotplot pipeline with specific FDR cutoffs (red), and paired Shannon entropies of randomly sampled, length matched sequences from the proteome (blue).

E) Cumulative distribution plot of number of total LCRs of all LCR-containing proteins in the human proteome. Dotted line separates the proportion of proteins with only 1 LCR from those with >1 LCR.

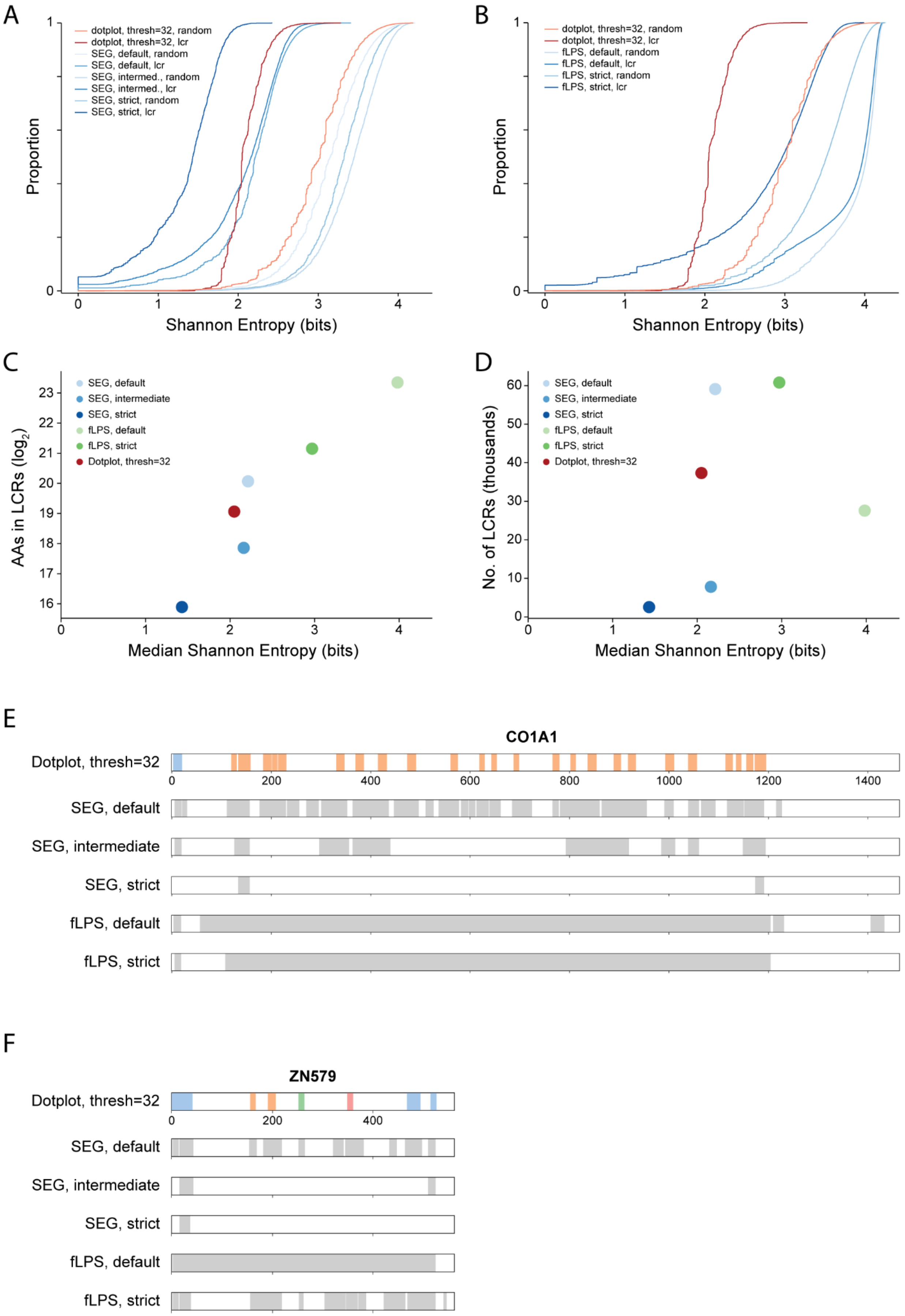

A
Proportion
Shannon Entropy (bits)
0
1
0
1
2
3
4
dotplot, thresh=32, random
dotplot, thresh=32, lcr
SEG, default, random
SEG, default, lcr
SEG, intermed., random
SEG, intermed., lcr
SEG, strict, random
SEG, strict, lcr
B
Proportion
Shannon Entropy (bits)
dotplot, thresh=32, random
dotplot, thresh=32, lcr
fLPS, default, random
fLPS, default, lcr
fLPS, strict, random
fLPS, strict, lcr
C
AAs in LCRs ($\log_2$)
Median Shannon Entropy (bits)
16
17
18
19
20
21
22
23
SEG, default
SEG, intermediate
SEG, strict
fLPS, default
fLPS, strict
Dotplot, thresh=32
D
No. of LCRs (thousands)
Median Shannon Entropy (bits)
0
10
20
30
40
50
60
SEG, default
SEG, intermediate
SEG, strict
fLPS, default
fLPS, strict
Dotplot, thresh=32
E
CO1A1
Dotplot, thresh=32
0
200
400
600
800
1000
1200
1400
SEG, default
SEG, intermediate
SEG, strict
fLPS, default
fLPS, strict
F
ZN579
Dotplot, thresh=32
0
200
400
SEG, default
SEG, intermediate
SEG, strict
fLPS, default
fLPS, strict

Figure 1 - figure supplement 3: Comparison of systematic dotplot analysis to existing LCR calling software, SEG and fLPS

A) Cumulative distributions of Shannon entropy of LCRs called using dotplots (Threshold=32, FDR=0.002) or SEG (default, intermediate, or strict, see Methods for details), and paired Shannon entropies of randomly sampled, length matched sequences from the proteome. Red lines represent dotplot approach, blue lines represent SEG. Dark and light shades correspond to called LCRs and randomly sampled sequences respectively.

B) Same as A) but using fLPS (default or strict, see Methods for details).

C) Total number of amino acids in LCRs (Log2) vs median Shannon entropy of called LCRs by dotplot approach, SEG, and fLPS.

D) Number of called LCRs (thousands) vs median Shannon entropy of called LCRs called by dotplot approach, SEG, and fLPS.

E) Schematic of LCR coordinates called by dotplot approach, SEG, and fLPS for CO1A1. Different colors in schematic correspond to different LCR types for the dotplot approach.

F) Same as E), but for ZN579.

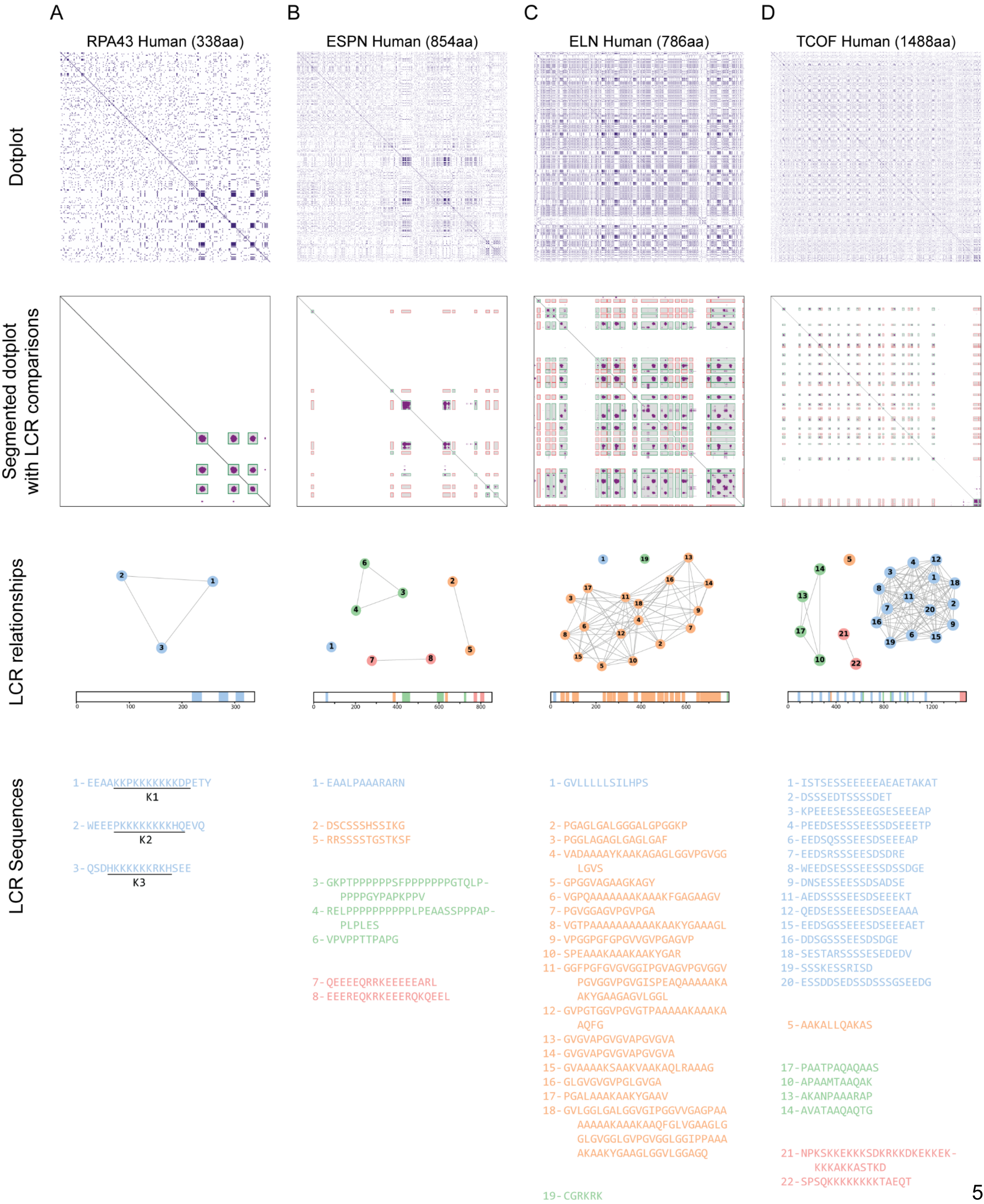

A
B
C
D
RPA43 Human (338aa)
ESPN Human (854aa)
ELN Human (786aa)
TCOF Human (1488aa)
Dotplot
Segmented dotplot with LCR comparisons
LCR relationships
LCR Sequences
0
100
200
300
0
200
400
600
800
0
200
400
600
0
400
800
1200
1-EEAAKKPKKKKKKKDPETY
K1
2-WEEEPKKKKKKKKHQEVQ
K2
3-QSDHKKKKKKRKHSEE
K3
1-EAALPAAARARN
2-DSCSSSHSSIKG
5-RRSSSSTGSTKSF
3-GKPTPPPPPPSFPPPPPPPGTQLP-
PPPPGYPAPKPPV
4-RELPPPPPPPPPPLPEAASSPPPAP-
PLPLES
6-VPVPPTTPAPG
7-QEEEEQRRKEEEEEARL
8-EEEREQKRKEEERQKQEEL
1-GVLLLLLSILHPS
2-PGAGLGALGGGALGPGGKP
3-PGGLAGAGLGAGLGAF
4-VADAAAAYKAAKAGAGLGGVPGVGG
LGVS
5-GPGGVAGAAGKAGY
6-VGPQAAAAAAAKAAAKFGAGAAGV
7-PGVGGAGVPGVPGA
8-VGTPAAAAAAAAAAKAAKYGAAAGL
9-VPGGPGFGPGVVGVPGAGVP
10-SPEAAAKAAAKAAKYGAR
11-GGFPGFGVGVGGIPGVAGVPGVGGV
PGVGGVPGVGISPEAQAAAAAKA
AKYGAAGAGVLGGL
12-GVPGTGGVPGVGTPAAAAAKAAAKA
AQFG
13-GVGVAPGVGVAPGVGVA
14-GVGVAPGVGVAPGVGVA
15-GVAAAAKSAAKVAAKAQLRAAAG
16-GLGVGVGVPGLGVGA
17-PGALAAAKAAKYGAAV
18-GVLGGLGALGGVGIPGGVVGAGPAA
AAAAAKAAAKAAQFGLVGAAGLG
GLGVGGLGVPGVGGLGGIPPAAA
AKAAKYGAAGLGGVLGGAGQ
19-CGRKRK
1-ISTSESSEEEEEAEAETAKAT
2-DSSSEDTSSSSDET
3-KPEEESESSEEGSESEEEAP
4-PEEDSESSSEESSDSEEETP
6-EEDSQSSSEESDSEEEAP
7-EEDSRSSSEESDSDRE
8-WEEDSESSSEESSDSSDGE
9-DNSESSEESSDSADSE
11-AEDSSSSEESDSEEEKT
12-QEDSESSEEESDSEEAAA
15-EEDSGSSEEESDSEEEAET
16-DDSGSSSEESDSDGE
18-SESTARSSSSESEDEDV
19-SSSKESSRISD
20-ESSDDSEDSSDSSSGSEEDG
5-AAKALLQAKAS
17-PAATPAQAQAAS
10-APAAMTAAQAK
13-AKANPAAARAP
14-AVATAAQAQTG
21-NPKSKKEKKKSDKRKKDKEKKKEK-
KKKAKKASTKD
22-SPSQKKKKKKKKTAEQT

Figure 1 - figure supplement 4: Sequential steps of dotplot pipeline performed for several example proteins
Raw dotplots (top row), segmented dotplot with LCR comparisons (second row), LCR relationship summaries (third row), and LCR sequences (bottom row) for A) RPA43, B) ESPN, C) ELN, and D) TCOF.

For all dotplots, the protein sequence lies from N-terminus to C-terminus from top to bottom, and left to right. For segmented dotplots with LCR comparisons (second row), green squares represent matching LCRs, and red squares represent non-matching LCRs. All LCRs along the diagonal are green since they match with themselves. LCR relationship summaries (third row) contain a graph-based representation of LCR relationships and a schematic of LCRs within the protein. For the graph-based representation, LCRs are represented by nodes, and LCRs which match off of the diagonal are connected by edges. LCRs part of the same connected component are designated as the same type, and colored the same. Numbers represent the LCR identifier within the protein from N-terminus to C-terminus. Schematic under network representation shows coordinates of called LCRs and their types, with colors corresponding to the connected components in network representation for each protein. For LCR sequences (bottom row), the LCR number and sequence of each LCR is shown. These numbers are the same as those in the graph representation. Raw dotplots for RPA43 and TCOF are also included in main Figures (2C and 4H respectively), but are also shown here for completeness of illustrating the processing steps.

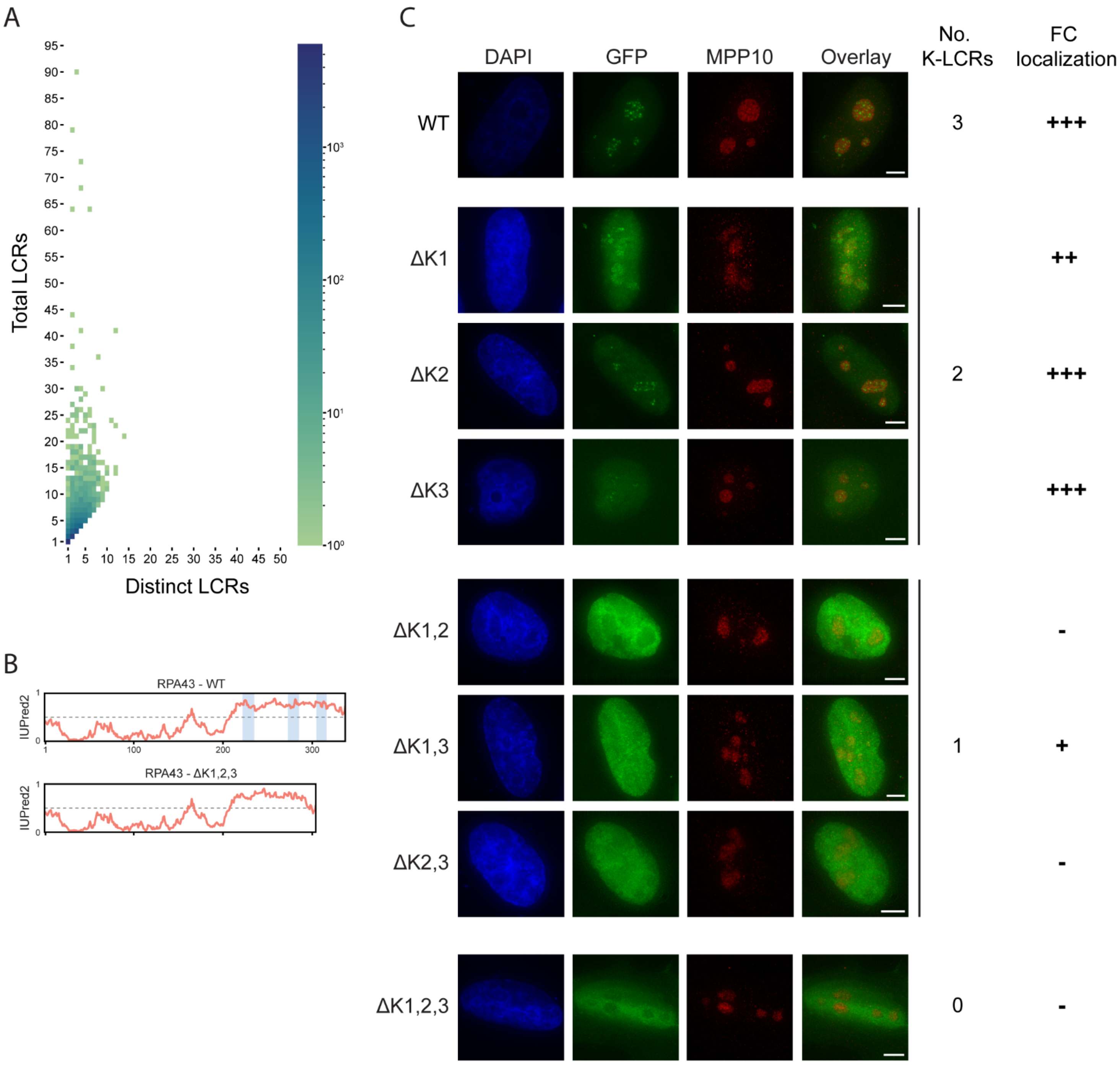


Figure 2 - figure supplement 1: Supplementary information for LCR type and copy number

A) Distribution of total and distinct LCRs for all LCR-containing proteins in the human proteome from Figure 2A, without binning proteins with 10+ total LCRs and/or 10+ distinct LCRs. The number in each square is the number of proteins in the human proteome with that number of total and distinct LCRs and is represented by the colorbar.

B) Disorder tendency (predicted by IUPred2A) of WT or ΔK1,2,3 RPA43. Coordinates of the three K-rich LCRs of RPA43 are indicated in blue.

C) Immunofluorescence of RPA43 constructs in HeLa cells. HeLa cells were seeded on fibronectin-coated coverslips and transfected with the indicated GFP-RPA43 constructs, and collected ~48 h following transfection. DAPI, GFP, and MPP10 channels are shown. Scale bar is 5 µm. The number of K-rich LCRs present and fibrillar center (FC) localization scoring is shown to the right of each construct ('+++' to '+' = strong FC localization to uniform nuclear localization, '-' = nucleolar exclusion).

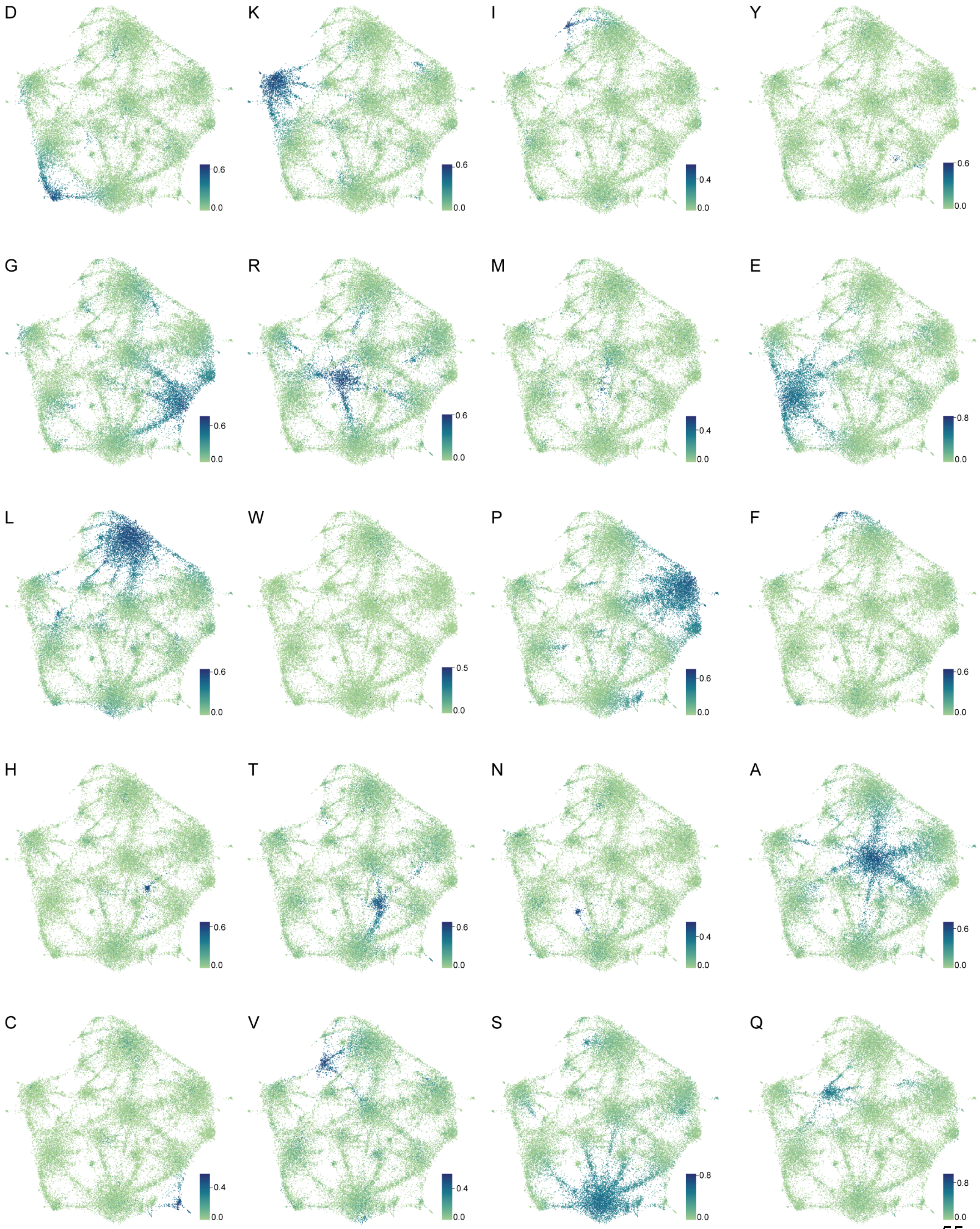

D
0.6
0.0
K
0.6
0.0
I
0.4
0.0
Y
0.6
0.0
G
0.6
0.0
R
0.6
0.0
M
0.4
0.0
E
0.8
0.0
L
0.6
0.0
W
0.5
0.0
P
0.6
0.0
F
0.6
0.0
H
0.6
0.0
T
0.6
0.0
N
0.4
0.0
A
0.6
0.0
C
0.4
0.0
V
0.4
0.0
S
0.8
0.0
Q
0.8
0.0

Figure 3 - figure supplement 1: Amino acid frequency distributions on human proteome UMAP from Figure 3A. Color of each dot corresponds to the frequency of the given amino acid in every LCR, as defined by each respective colorbar.

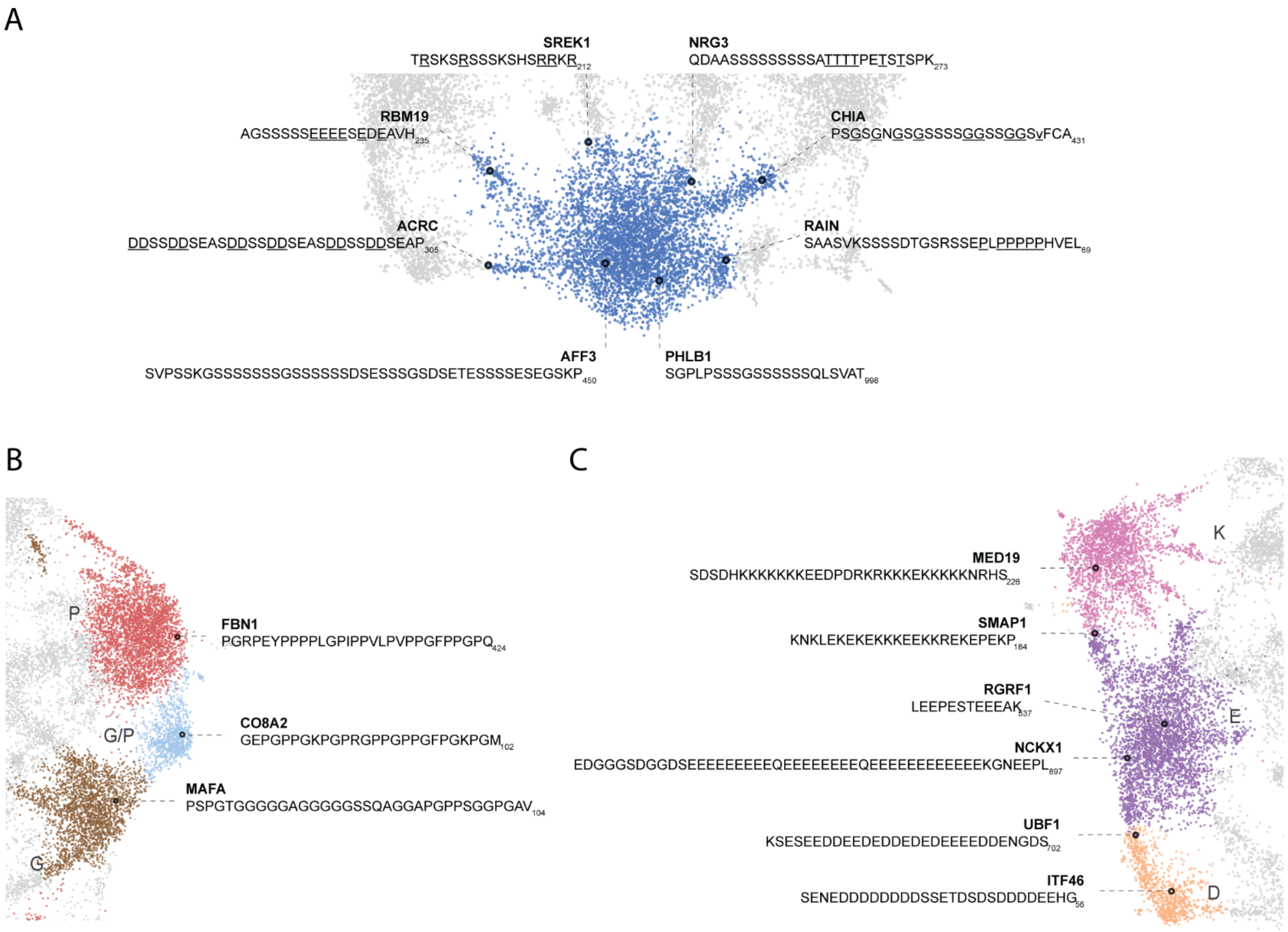


Figure 3 - figure supplement 2: Nuanced sequence differences among LCRs correspond to their positions in the UMAP
Close up view of specific clusters in human proteome UMAP (shown in Figure 3A), with several LCR sequences and their parent proteins annotated. For all LCRs shown the subscript at the end of the sequence corresponds to the ending position of the LCR in the sequence of its parent protein.

A) Close-up view of S-rich Leiden cluster (bottom of UMAP in Figure 3A). For LCRs along bridges connecting to leiden clusters of other amino acids, the residues of that other amino acid are underlined. For example, the LCR from ACRC lies in the bridge between the S and D clusters, so the D residues are underlined to highlight their frequency.

B) Close-up view of P-rich, G/P-rich, and G-rich Leiden clusters (right side of UMAP in Figure 3A).

C) Close-up view of K-rich, E-rich, and D-rich Leiden clusters (left side of UMAP in Figure 3A).

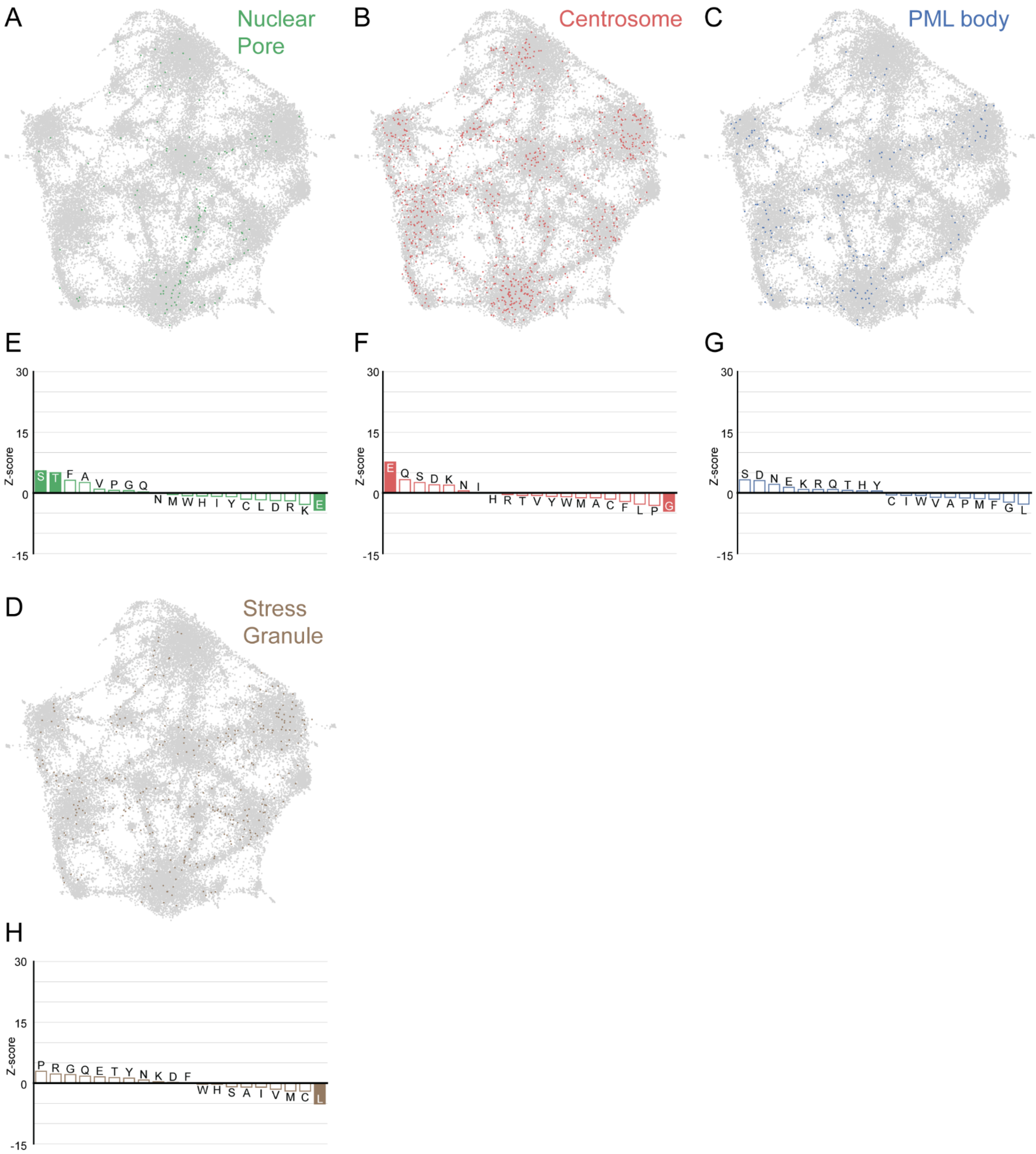


Figure 3 - figure supplement 3: LCRs of known higher order assemblies annotated on onto human proteome UMAP from Figure 3A

A) LCRs of annotated nuclear pore proteins (obtained from Uniprot, see Methods) plotted on UMAP.

B - D) Same as A), but for Centrosome, PML body, and Stress Granule (Jain et al., 2016) LCRs.

E) Barplot of Wilcoxon rank sum tests for amino acid frequencies of LCRs of nuclear pore proteins compared to all other LCRs in the human proteome. Filled bars represent amino acids with Benjamini-Hochberg adjusted p-value < 0.001. Positive Z-scores correspond to amino acids significantly enriched in LCRs of nuclear pore proteins, while negative Z-scores correspond to amino acids significantly depleted in LCRs of nuclear pore proteins.

F - H) Same as E), but for Centrosome, PML body, and Stress Granule LCRs, respectively.

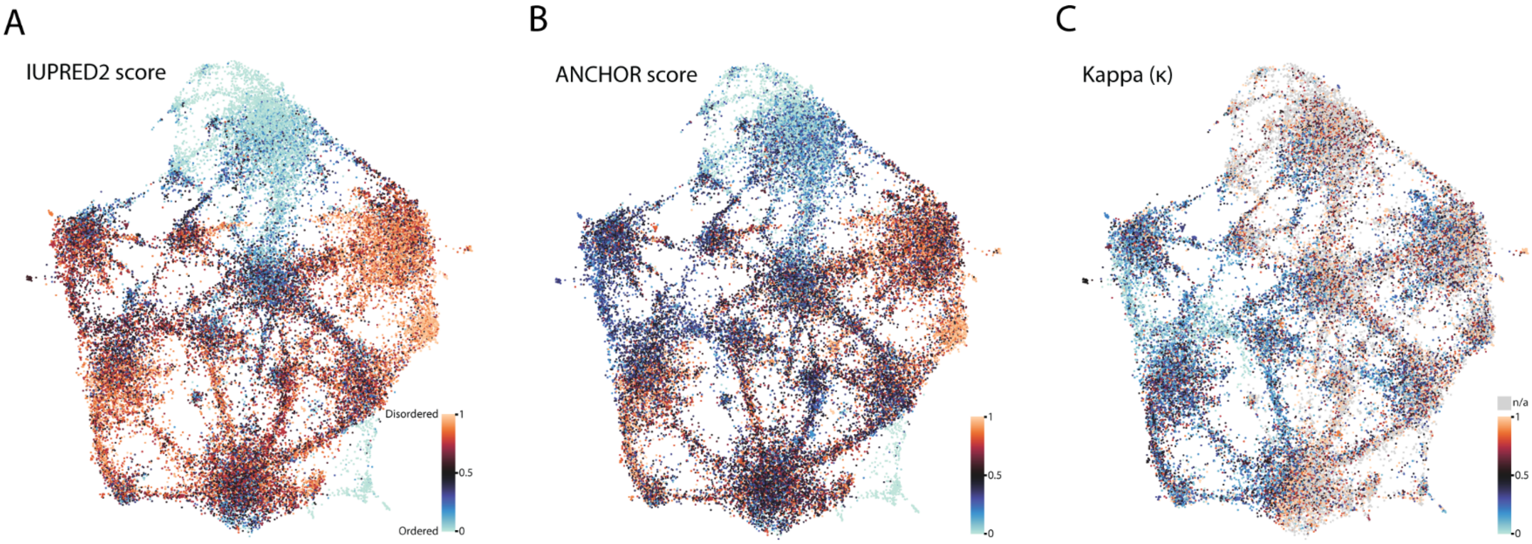


Figure 3 - figure supplement 4: Biophysical predictions of LCRs mapped onto human proteome UMAP from Figure 3A.

A) Predicted disorder (IUPred2A) for all LCRs in human proteome.

B) ANCHOR scores for all LCRs in human proteome.

C) Kappa scores (Das and Pappu, 2013) for all LCRs in human proteome.

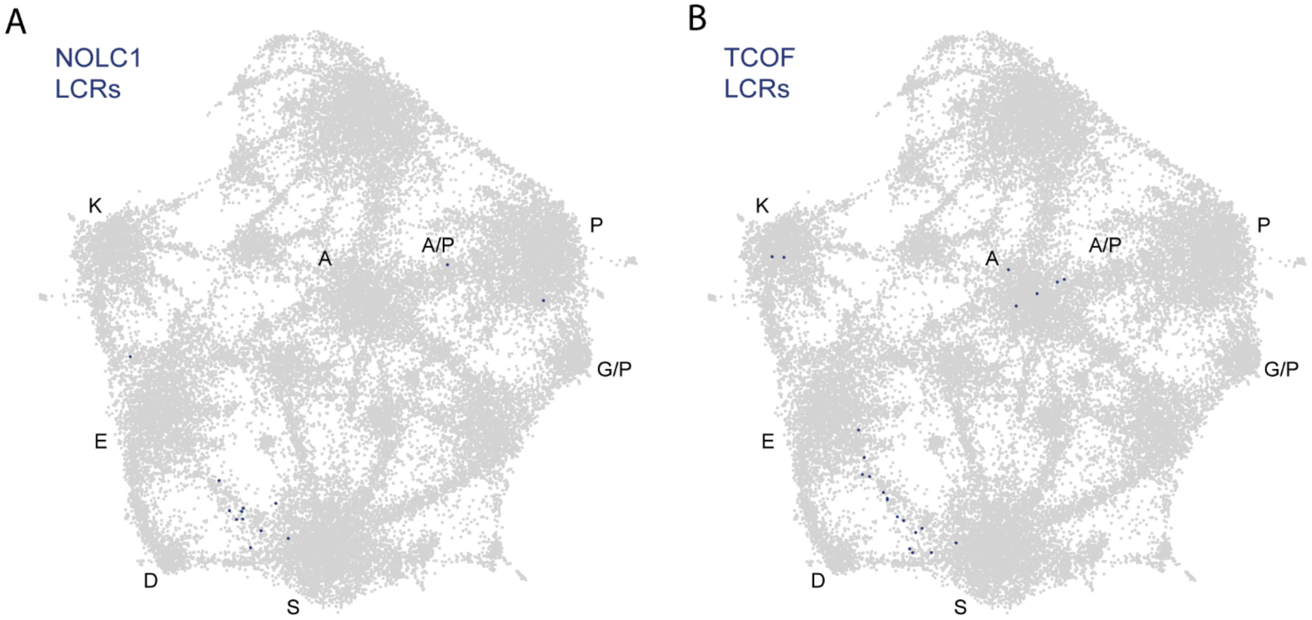


Figure 4 - figure supplement 1: LCRs of NOLC1 and TCOF mapped onto human proteome UMAP from Figure 3A.

A) NOLC1 LCRs displayed on UMAP from Figure 3A. This is a full view of the close-up in Figure 4F, and is included for completeness.

B) TCOF LCRs displayed on UMAP from Figure 3A. This is a full view of the close-up in Figure 4F, and is included for completeness.

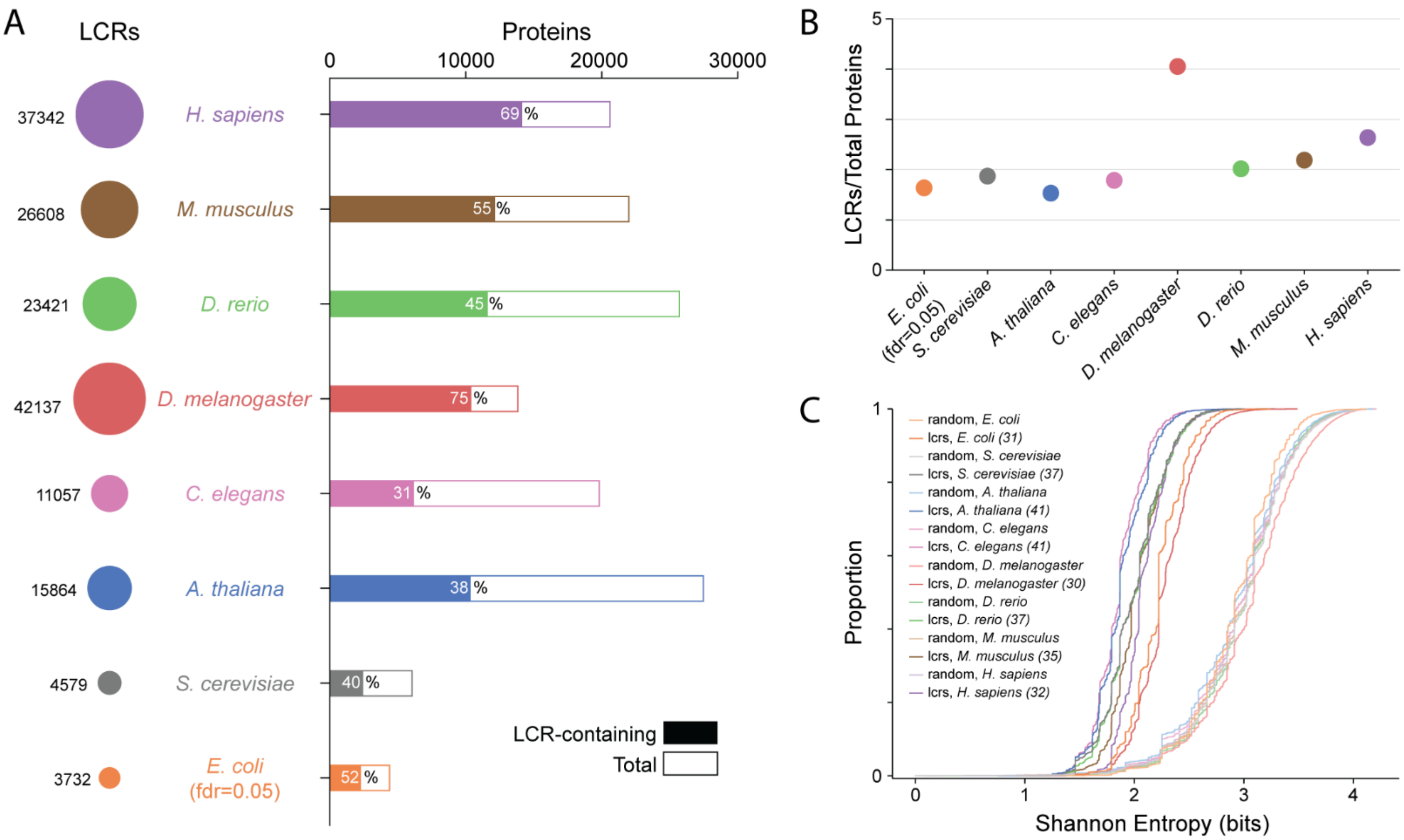


Figure 5 - figure supplement 1: Summary statistics from systematic dotplot analysis across species

A) Summary information for systematic dotplot analysis on proteomes of human (*H. Sapiens*), mouse (*M. musculus*), zebrafish (*D. rerio*), *fruit fly* (*D. melanogaster*), *worm* (*C. elegans*), Baker's yeast (*S. cerevisiae*), *A. thaliana, and E.coli*. Circles on the left column correspond to the number of LCRs in each proteome. Bar plot corresponds to the total number of proteins (open bar) and LCR-containing proteins (shaded bar) in each proteome. Percentage of LCR-contain proteins out of total proteins in the respective proteome is inset in each bar.

B) The average number of LCRs per protein for each proteome in A).

C) Cumulative Shannon entropy distributions of LCRs called using dotplot approach for all proteomes in A) and paired Shannon entropies of randomly sampled, length matched sequences from the same proteomes. Dark and light shades correspond to called LCRs and randomly selected sequences respectively. An FDR of 0.05 was used for *E. Coli*, and an FDR of 0.002 was used for all other species. The corresponding convolved pixel intensity thresholds for each proteome are indicated in parentheses.

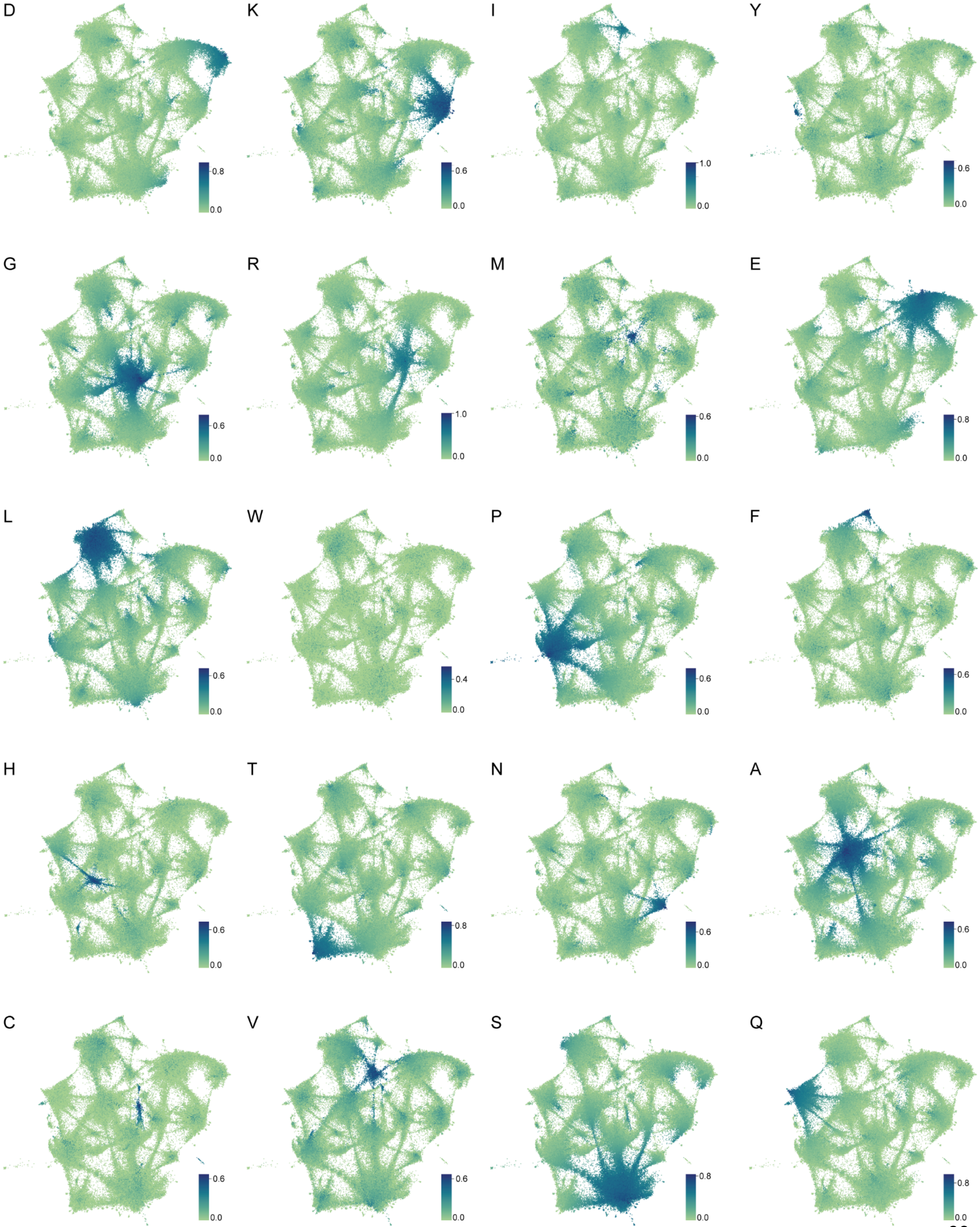

D
0.8
0.0
K
0.6
0.0
I
1.0
0.0
Y
0.6
0.0
G
0.6
0.0
R
1.0
0.0
M
0.6
0.0
E
0.8
0.0
L
0.6
0.0
W
0.4
0.0
P
0.6
0.0
F
0.6
0.0
H
0.6
0.0
T
0.8
0.0
N
0.6
0.0
A
0.6
0.0
C
0.6
0.0
V
0.6
0.0
S
0.8
0.0
Q
0.8
0.0

Figure 5 - figure supplement 2: Amino acid frequency distributions mapped onto expanded UMAP from Figure 5A
Frequency of each amino acid in LCRs from the proteomes of human (*H. Sapiens)*, mouse (*M. musculus*), zebrafish (*D. rerio*)*, fruit fly* (*D. melanogaster*)*,* worm (*C. elegans*)*,* Baker's yeast (*S. cerevisiae*)*, A. thaliana, and E.coli* displayed on the UMAP from Figure 5A. Color of each dot corresponds to the frequency of the given amino acid in every LCR, as defined by each respective colorbar.

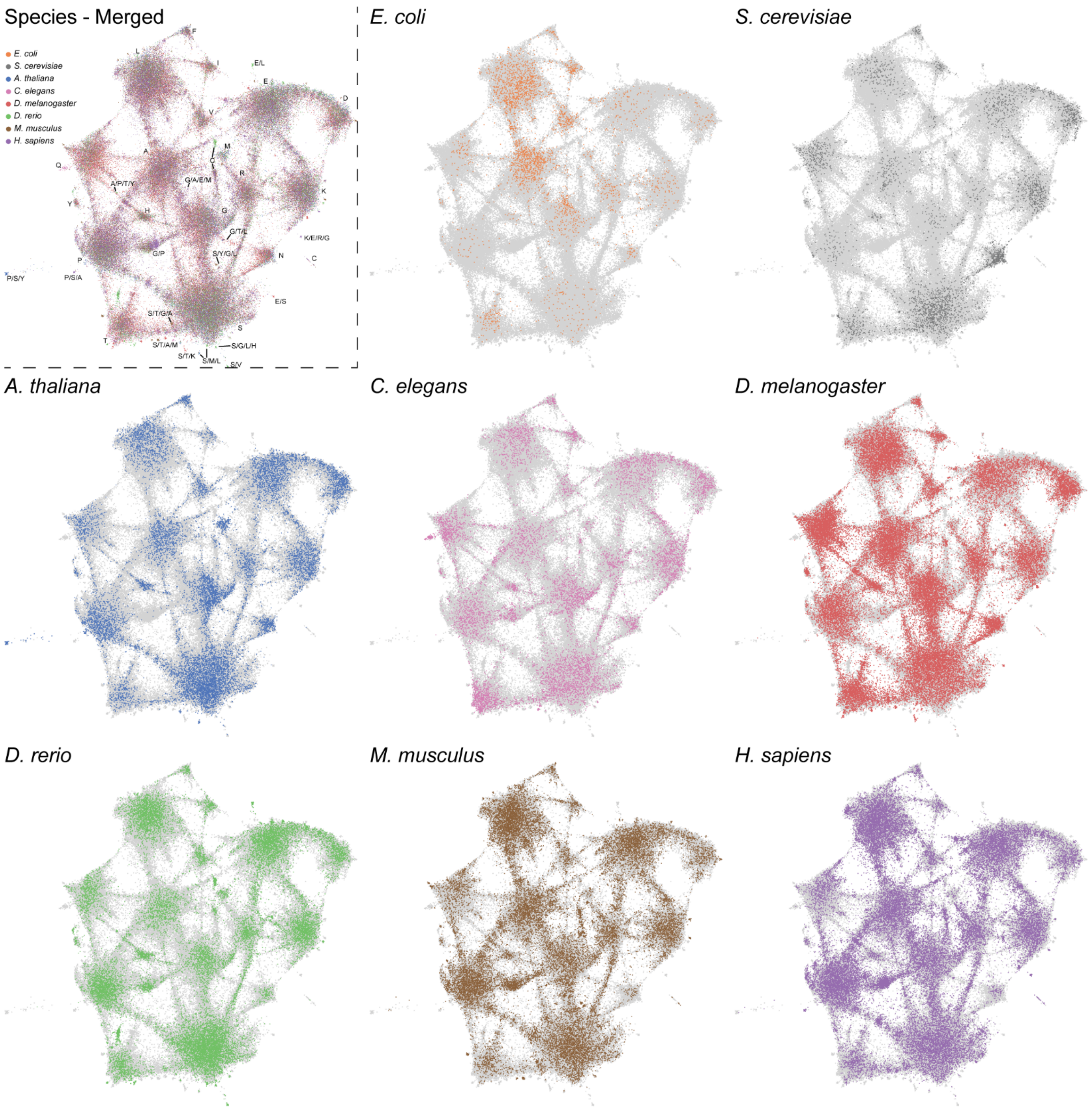


Figure 5 - figure supplement 3: LCRs of individual species mapped onto expanded UMAP from Figure 5A
UMAPs of LCRs in proteomes of human (*H. Sapiens)*, mouse (*M. musculus*), zebrafish (*D. rerio*), *fruit fly* (*D. melanogaster*), worm (*C. elegans*), Baker's yeast (*S. cerevisiae*), *A. thaliana, and E.coli* (same as that in Figure 5A). Top left panel contains UMAP with all LCRs colored by species. Labels indicate the most prevalent amino acid(s) among LCRs in corresponding leiden clusters. Other panels contain UMAP with LCRs of each species colored separately as indicated. In panels where LCRs of only one species are colored, light grey regions in the UMAP represent LCRs from other species.

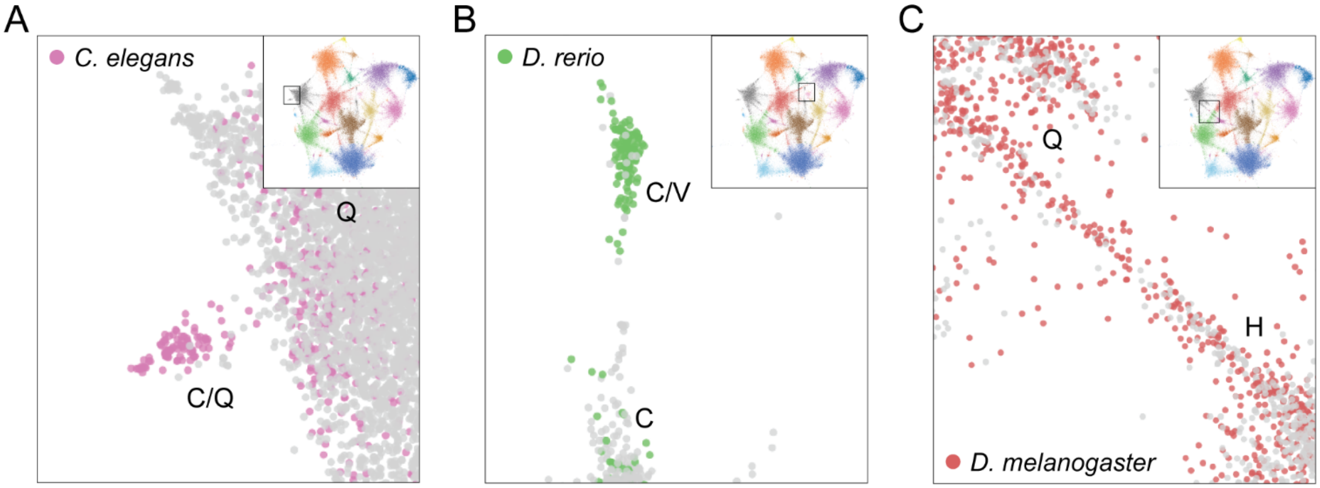


Figure 5 - figure supplement 4: Examples of species-specific clusters in the expanded UMAP from Figure 5A

A) Close-up view of C/Q-rich cluster (upper-left side of UMAP in Figure 5A). Pink circles indicate LCRs from *C. elegans*. Grey circles indicate LCRs from other species.

B) Close-up view of C/V-rich cluster (upper-middle region of UMAP in Figure 5A). Green circles indicate LCRs from *D. rerio.* Grey circles indicate LCRs from other species.

C) Close-up view of H/Q-rich bridge (middle-left side of UMAP in Figure 5A). Red circles indicate LCRs from *D. melanogaster*. Grey circles indicate LCRs from other species.

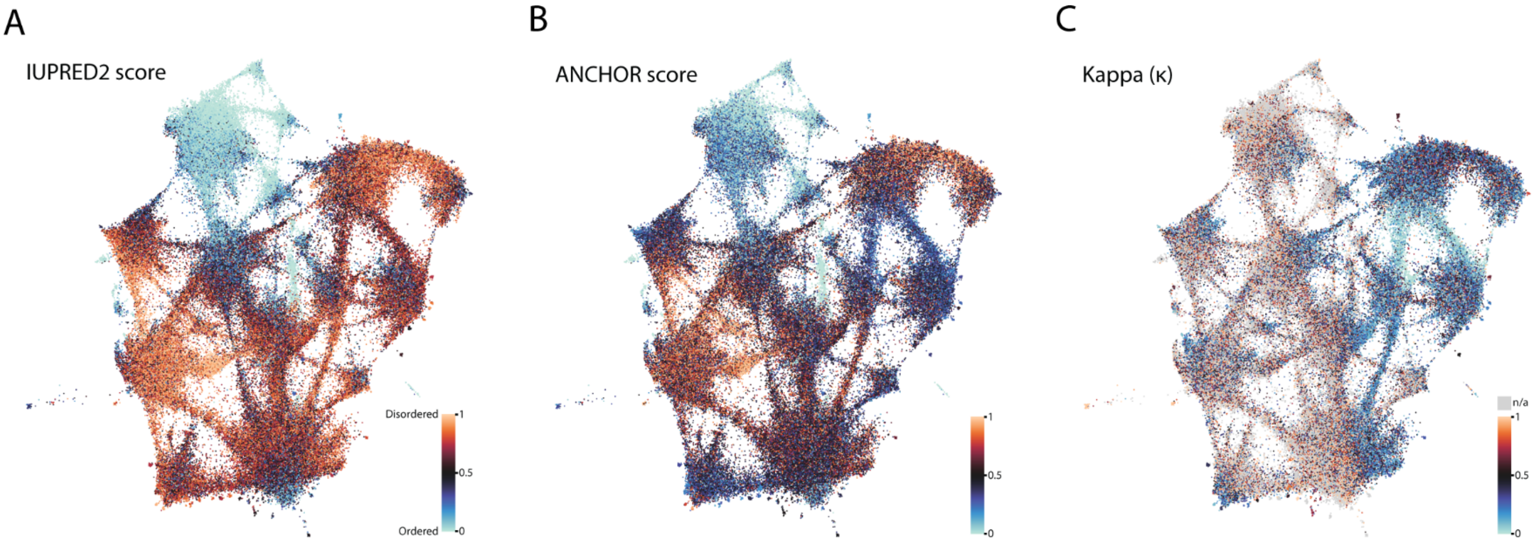


Figure 5 - figure supplement 5: Biophysical predictions of LCRs mapped onto the expanded UMAP from Figure 5A. Mapping biophysical predictions of LCRs onto UMAP of LCRs from proteomes of human (*H. Sapiens)*, mouse (*M. musculus*), zebrafish (*D. rerio*)*, fruit fly* (*D. melanogaster*)*,* worm (*C. elegans*)*,* Baker's yeast (*S. cerevisiae*)*, A. thaliana, and E.coli* (same UMAP as that shown in Figure 5A).

A) Predicted disorder (IUPred2A) for all LCRs on UMAP.

B) ANCHOR scores for all LCRs on UMAP.

C) Kappa scores (Das and Pappu, 2013) for all LCRs on UMAP.

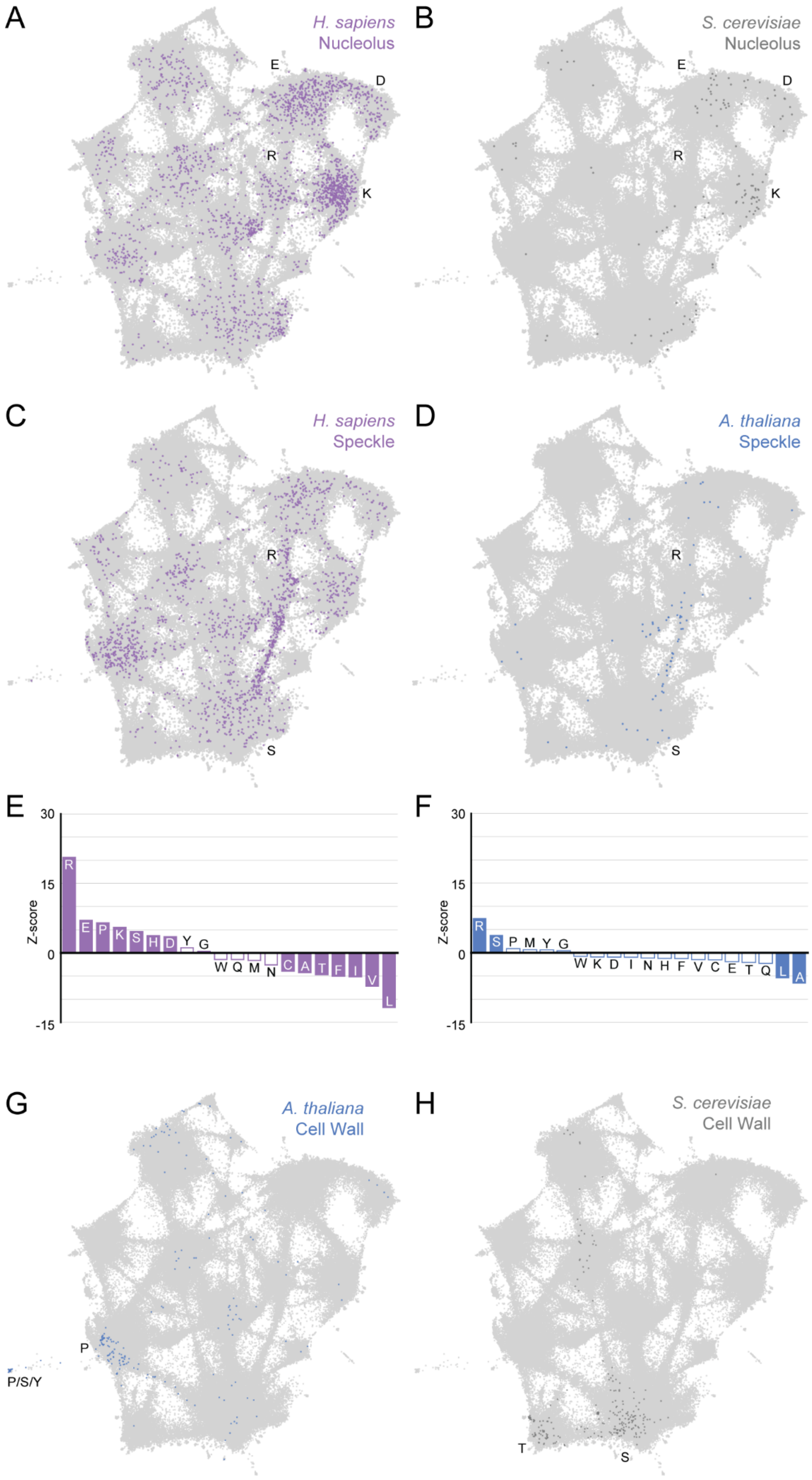

A
H. sapiens
Nucleolus
E
D
R
K
B
S. cerevisiae
Nucleolus
E
D
R
K
C
H. sapiens
Speckle
R
S
D
A. thaliana
Speckle
R
S
E
30
15
0
-15
Z-score
R E P K S H D Y G W Q M N C A T F I V L
F
30
15
0
-15
Z-score
R S P M Y G W K D I N H F V C E T Q L A
G
A. thaliana
Cell Wall
P
P/S/Y
H
S. cerevisiae
Cell Wall
T
S

Figure 5 - figure supplement 6: Higher order assemblies in different species annotated on the expanded UMAP from Figure 5A

Mapping higher order assembly annotations of LCRs onto UMAP of LCRs from proteomes of human (*H. Sapiens)*, mouse (*M. musculus*), zebrafish (*D. rerio*)*, fruit fly* (*D. melanogaster*)*,* worm (*C. elegans*)*,* Baker's yeast (*S. cerevisiae*)*, A. thaliana, and E.coli* (same UMAP as that shown in Figure 5A). A, B, G, and H are full views of insets shown in Figure 5, and included for completeness.

A) Full view of expanded UMAP with LCRs of annotated *H. sapiens* nucleolar proteins indicated.

B) Same as (A), but for annotated *S. cerevisiae* nucleolar proteins.

C) Same as (A), but for annotated *H. sapiens* nuclear speckle proteins.

D) Same as (A), but for annotated *A. thaliana* nuclear speckle proteins.

E) Barplot of Wilcoxon rank sum tests for amino acid frequencies of LCRs of annotated *H. sapiens* nuclear speckle proteins compared to all other LCRs. Filled bars represent amino acids with Benjamini-Hochberg adjusted p-value < 0.001. Positive Z-scores correspond to amino acids enriched in LCRs of *H. sapiens* nuclear speckle proteins, while negative Z-scores correspond to amino acids depleted in LCRs of *H. sapiens* nuclear speckle proteins.

F) Same as (E), but for annotated *A. thaliana* nuclear speckle proteins.

G) Same as (A), but for annotated *A. thaliana* cell wall proteins.

H) Same as (A), but for annotated *S. cerevisiae* cell wall proteins.

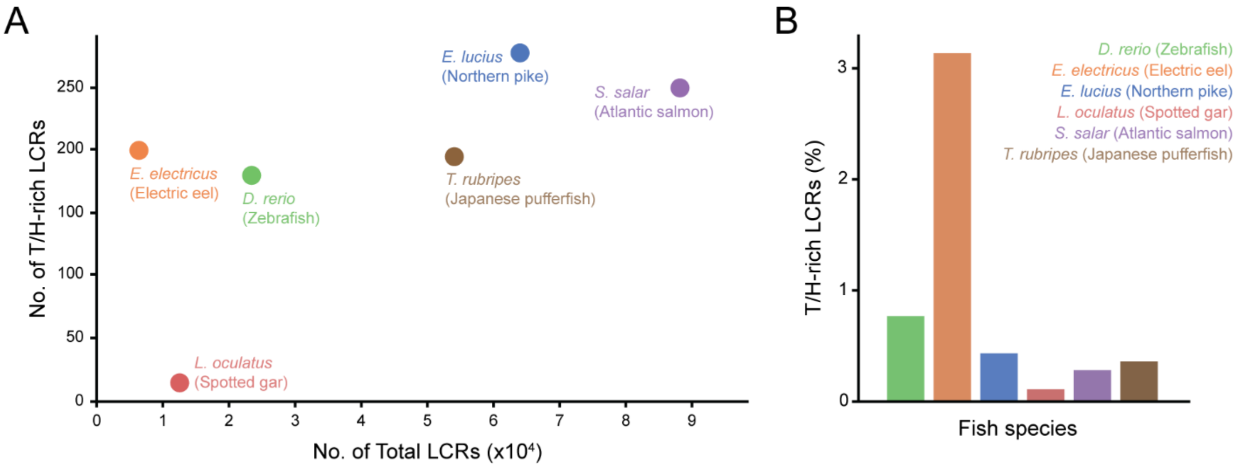


Figure 6 - figure supplement 1: Number and proportion of T/H-rich LCRs across fish species

A) Number of T/H-rich LCRs vs. total number of LCRs in proteomes of zebrafish (*D. rerio*), Spotted gar (*L. oculatus*), Electric eel (*E. electricus*), Northern pike (*E. lucius*), Atlantic salmon (*S. salar*), and Japanese pufferfish (*T. rubripes*).

B) Barplot of the T/H-rich LCRs in the proteomes of the fish species in (A), shown as the percentage of the total number of LCRs.

REFERENCES

Albà, M.M., Laskowski, R.A., and Hancock, J.M. (2002). Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinformatics *18*, 672–678.

Banani, S.F., Rice, A.M., Peeples, W.B., Lin, Y., Jain, S., Parker, R., and Rosen, M.K. (2016). Compositional Control of Phase-Separated Cellular Bodies. Cell *166*, 651–663.

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nat. Rev. Mol. Cell Biol. *18*, 285–298.

Beck, K., Chan, V.C., Shenoy, N., Kirkpatrick, A., Ramshaw, J.A.M., and Brodsky, B. (2000). Destabilization of osteogenesis imperfecta collagen-like model peptides correlates with the identity of the residue replacing glycine. Proc. Natl. Acad. Sci. *97*, 4273–4278.

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 Cell Biol. *28*, 420–435.

Boucher, L., Ouzounis, C.A., Enright, A.J., and Blencowe, B.J. (2001). A genome-wide survey of RS domain proteins. RNA *7*, 1693–1701.

Cáceres, J.F., Misteli, T., Screaton, G.R., Spector, D.L., and Krainer, A.R. (1997). Role of the Modular Domains of SR Proteins in Subnuclear Localization and Alternative Splicing Specificity. J. Cell Biol. *138*, 225–238.

Cannon, M.C., Terneus, K., Hall, Q., Tan, L., Wang, Y., Wegenhart, B.L., Chen, L., Lamport, D.T.A., Chen, Y., and Kieliszewski, M.J. (2008). Self-assembly of the plant cell wall requires an extensin scaffold. Proc. Natl. Acad. Sci. *105*, 2226–2231.

Cirillo, L., Cieren, A., Barbieri, S., Khong, A., Schwager, F., Parker, R., and Gotta, M. (2020). UBAP2L Forms Distinct Cores that Act in Nucleating Stress Granules Upstream of G3BP1. Curr. Biol. *30*, 698-707.e6.

Cowan, P.M., and McGAVIN, S. (1955). Structure of Poly-L-Proline. Nature *176*, 501–503.

Das, R.K., and Pappu, R.V. (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc. Natl. Acad. Sci. *110*, 13392–13397.

Dundr, M., Hoffmann-Rohrer, U., Hu, Q., Grummt, I., Rothblum, L.I., Phair, R.D., and Misteli, T. (2002). A Kinetic Framework for a Mammalian RNA Polymerase in Vivo. Science *298*, 1623–1626.

Fantini, D., Vascotto, C., Marasco, D., D’Ambrosio, C., Romanello, M., Vitagliano, L., Pedone, C., Poletto, M., Cesaratto, L., Quadrifoglio, F., et al. (2010). Critical lysine residues within the overlooked N-terminal domain of human APE1 regulate its biological functions. Nucleic Acids Res. *38*, 8239–8256.

Fei, J., Jadaliha, M., Harmon, T.S., Li, I.T.S., Hua, B., Hao, Q., Holehouse, A.S., Reyer, M., Sun, Q., Freier, S.M., et al. (2017). Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. J. Cell Sci. *130*, 4180–4192.

Forgacs, G., Newman, S.A., Hinner, B., Maier, C.W., and Sackmann, E. (2003). Assembly of Collagen Matrices as a Phase Transition Revealed by Structural and Rheologic Studies. Biophys. J. *84*, 1272–1280.

Fossat, M.J., Zeng, X., and Pappu, R.V. (2021). Uncovering Differences in Hydration Free Energies and Structures for Model Compound Mimics of Charged Side Chains of Amino Acids. J. Phys. Chem. B *125*, 4148–4161.

Gentzsch, M., and Tanner, W. (1996). The PMT gene family: protein O-glycosylation in Saccharomyces cerevisiae is vital. EMBO J. *15*, 5752–5759.

Gibbs, A.J., and Mcintyre, G.A. (1970). The Diagram, a Method for Comparing Sequences. Eur. J. Biochem. *16*, 1–11.

Gomes, E., and Shorter, J. (2019). The molecular language of membraneless organelles. J. Biol. Chem. *294*, 7115–7127.

González, M., Brito, N., and González, C. (2012). High abundance of Serine/Threonine-rich regions predicted to be hyper-O-glycosylated in the secretory proteins coded by eight fungal genomes. BMC Microbiol. *12*, 213.

Grasberger, H., and Bell, G.I. (2005). Subcellular recruitment by TSG118 and TSPYL implicates a role for zinc finger protein 106 in a novel developmental pathway. Int. J. Biochem. Cell Biol. *37*, 1421–1437.

Greig, J.A., Nguyen, T.A., Lee, M., Holehouse, A.S., Posey, A.E., Pappu, R.V., and Jedd, G. (2020). Arginine-Enriched Mixed-Charge Domains Provide Cohesion for Nuclear Speckle Condensation. Mol. Cell *77*, 1237-1250.e4.

Guillén-Boixet, J., Kopach, A., Holehouse, A.S., Wittmann, S., Jahnel, M., Schlüßler, R., Kim, K., Trussina, I.R.E.A., Wang, J., Mateju, D., et al. (2020). RNA-Induced Conformational Switching and Clustering of G3BP Drive Stress Granule Assembly by Condensation. Cell *181*, 346-361.e17.

Hansen, U., and Bruckner, P. (2003). Macromolecular Specificity of Collagen Fibrillogenesis: FIBRILS OF COLLAGENS I AND XI CONTAIN A HETEROTYPIC ALLOYED CORE AND A COLLAGEN I SHEATH*. J. Biol. Chem. *278*, 37352–37359.

Harrison, P.M. (2017). fLPS: Fast discovery of compositional biases for the protein universe. BMC Bioinformatics *18*, 476.

Hebert, M.D., and Matera, A.G. (2000). Self-association of Coilin Reveals a Common Theme in Nuclear Body Localization. Mol. Biol. Cell *11*, 4159–4171.

Hinman, M.B., and Lewis, R.V. (1992). Isolation of a clone encoding a second dragline silk fibroin. Nephila clavipes dragline silk is a two-protein fiber. J. Biol. Chem. *267*, 19320–19324.

Holehouse, A.S., Das, R.K., Ahad, J.N., Richardson, M.O.G., and Pappu, R.V. (2017). CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. Biophys. J. *112*, 16–21.

Hughes, L.C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C.C., Thompson, A.W., Arcila, D., Betancur-R, R., Li, C., Becker, L., et al. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. *115*, 6249–6254.

Huntley, M.A., and Clark, A.G. (2007). Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. Mol. Biol. Evol. *24*, 2598–2609.

Huntley, M.A., and Golding, G.B. (2002). Simple sequences are rare in the Protein Data Bank. Proteins Struct. Funct. Bioinforma. *48*, 134–140.

Hynes, R.O. (2012). The evolution of metazoan extracellular matrix. J. Cell Biol. *196*, 671–679.

Ilik, İ.A., Malszycki, M., Lübke, A.K., Schade, C., Meierhofer, D., and Aktaş, T. (2020). SON and SRRM2 are essential for nuclear speckle formation. ELife *9*, e60579.

Jain, S., Wheeler, J.R., Walters, R.W., Agrawal, A., Barsic, A., and Parker, R. (2016). ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. Cell *164*, 487–498.

Jarnot, P., Ziemska-Legiecka, J., Dobson, L., Merski, M., Mier, P., Andrade-Navarro, M.A., Hancock, J.M., Dosztányi, Z., Paladin, L., Necci, M., et al. (2020). PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. Nucleic Acids Res. *48*, W77–W84.

Kieliszewski, M.J., and Lamport, D.T.A. (1994). Extensin: repetitive motifs, functional sites, post-translational codes, and phylogeny. Plant J. *5*, 157–172.

Kim, S.S.-Y., Sze, L., and Lam, K.-P. (2019a). The stress granule protein G3BP1 binds viral dsRNA and RIG-I to enhance interferon-β response. J. Biol. Chem. *294*, 6430–6438.

Kim, T.H., Tsang, B., Vernon, R.M., Sonenberg, N., Kay, L.E., and Forman-Kay, J.D. (2019b). Phospho-dependent phase separation of FMRP and CAPRIN1 recapitulates regulation of translation and deadenylation. Science *365*, 825–829.

Krystkowiak, I., and Davey, N.E. (2017). SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. Nucleic Acids Res. *45*, W464–W469.

Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J.A., Bukirova, D., Čalyševa, J., et al. (2020). ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Res. *48*, D296–D306.

Lai, S.K., Wang, Y.-Y., Wirtz, D., and Hanes, J. (2009). Micro- and macrorheology of mucus. Adv. Drug Deliv. Rev. *61*, 86–100.

Lamport, D.T.A., Kieliszewski, M.J., Chen, Y., and Cannon, M.C. (2011). Role of the Extensin Superfamily in Primary Cell Wall Architecture. Plant Physiol. *156*, 11–19.

Larsson, M., Brundell, E., Jörgensen, P.-M., Ståhl, S., and Höög, C. (1999). Characterization of a novel nucleolar protein that transiently associates with the condensed chromosomes in mitotic cells. Eur. J. Cell Biol. *78*, 382–390.

Li, P., Banjade, S., Cheng, H.-C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J.V., King, D.S., Banani, S.F., et al. (2012). Phase transitions in the assembly of multivalent signalling proteins. Nature *483*, 336–340.

Lirussi, L., Antoniali, G., Vascotto, C., D'Ambrosio, C., Poletto, M., Romanello, M., Marasco, D., Leone, M., Quadrifoglio, F., Bhakat, K.K., et al. (2012). Nucleolar accumulation of APE1 depends on charged lysine residues that undergo acetylation upon genotoxic stress and modulate its BER activity in cells. Mol. Biol. Cell *23*, 4079–4096.

Liu, Q., and Dreyfuss, G. (1996). A novel nuclear structure containing the survival of motor neurons protein. EMBO J. *15*, 3555–3565.

Lundby, A., Lage, K., Weinert, B.T., Bekker-Jensen, D.B., Secher, A., Skovgaard, T., Kelstrup, C.D., Dmytriyev, A., Choudhary, C., Lundby, C., et al. (2012). Proteomic Analysis of Lysine Acetylation Sites in Rat Tissues Reveals Organ Specificity and Subcellular Patterns. Cell Rep. *2*, 419–431.

Malay, A.D., Suzuki, T., Katashima, T., Kono, N., Arakawa, K., and Numata, K. (2020). Spider silk self-assembly via modular liquid-liquid phase separation and nanofibrillation. Sci. Adv. *6*, eabb6030.

Martin, E.W., Holehouse, A.S., Peran, I., Farag, M., Incicco, J.J., Bremer, A., Grace, C.R., Soranno, A., Pappu, R.V., and Mittag, T. (2020). Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science *367*, 694–699.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat.

Mészáros, B., Simon, I., and Dosztányi, Z. (2009). Prediction of Protein Binding Regions in Disordered Proteins. PLOS Comput. Biol. *5*, e1000376.

Mészáros, B., Erdős, G., and Dosztányi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. *46*, W329–W337.

Mitrea, D.M., Cika, J.A., Guy, C.S., Ban, D., Banerjee, P.R., Stanley, C.B., Nourse, A., Deniz, A.A., and Kriwacki, R.W. (2016). Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying R-rich linear motifs and rRNA. ELife *5*, e13571.

Mould, A.P., and Hulmes, D.J. (1987). Surface-induced aggregation of type I procollagen. J. Mol. Biol. *195*, 543–553.

Neubert, P., Halim, A., Zauser, M., Essig, A., Joshi, H.J., Zatorska, E., Larsen, I.S.B., Loibl, M., Castells-Ballester, J., Aebi, M., et al. (2016). Mapping the O-Mannose Glycoproteome in Saccharomyces cerevisiae*. Mol. Cell. Proteomics *15*, 1323–1337.

Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. *85*, 2444–2448.

Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C., and Ouzounis, C.A. (2000). CAST: an iterative algorithm for the complexity analysis of sequence tracts. Bioinformatics *16*, 915–922.

Radó-Trilla, N., and Albà, Mm. (2012). Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evol. Biol. *12*, 155.

Ramachandran, G.N., and Kartha, G. (1955). Structure of Collagen. Nature *176*, 593–595.

Rauscher, S., and Pomès, R. (2017). The liquid structure of elastin. ELife *6*, e26526.

Rauscher, S., Baud, S., Miao, M., Keeley, F.W., and Pomès, R. (2006). Proline and Glycine Control Protein Self-Organization into Elastomeric or Amyloid Fibrils. Structure *14*, 1667–1676.

Rich, A., and Crick, F.H.C. (1955). The Structure of Collagen. Nature *176*, 915–916.

Rusin, S.F., Adamo, M.E., and Kettenbach, A.N. (2017). Identification of Candidate Casein Kinase 2 Substrates in Mitosis by Quantitative Phosphoproteomics. Front. Cell Dev. Biol. *5*, 97.

Saito, T., Yamauchi, M., Abiko, Y., Matsuda, K., and Crenshaw, M.A. (2000). In vitro apatite induction by phosphophoryn immobilized on modified collagen fibrils. J. Bone Miner. Res. Off. J. Am. Soc. Bone Miner. Res. *15*, 1615–1619.

Sanders, D.W., Kedersha, N., Lee, D.S.W., Strom, A.R., Drake, V., Riback, J.A., Bracha, D., Eeftens, J.M., Iwanicki, A., Wang, A., et al. (2020). Competing Protein-RNA Interaction Networks Control Multiphase Intracellular Organization. Cell *181*, 306-324.e28.

Schuster, B.S., Reed, E.H., Parthasarathy, R., Jahnke, C.N., Caldwell, R.M., Bermudez, J.G., Ramage, H., Good, M.C., and Hammer, D.A. (2018). Controllable protein phase separation and modular recruitment to form responsive membraneless organelles. Nat. Commun. *9*, 2985.

Scott, M.S., Boisvert, F.-M., McDowall, M.D., Lamond, A.I., and Barton, G.J. (2010). Characterization and prediction of protein nucleolar localization sequences. Nucleic Acids Res. *38*, 7388–7399.

Sede, A.R., Borassi, C., Wengier, D.L., Mecchia, M.A., Estevez, J.M., and Muschietti, J.P. (2018). Arabidopsis pollen extensins LRX are required for cell wall integrity during pollen tube growth. FEBS Lett. *592*, 233–243.

Sharma, A., Takata, H., Shibahara, K., Bubulya, A., and Bubulya, P.A. (2010). Son Is Essential for Nuclear Speckle Organization and Cell Cycle Progression. Mol. Biol. Cell *21*, 650–663.

Shen, T.H., Lin, H.-K., Scaglioni, P.P., Yung, T.M., and Pandolfi, P.P. (2006). The Mechanisms of PML-Nuclear Body Formation. Mol. Cell *24*, 331–339.

Shimizu, K., Amano, T., Bari, M.R., Weaver, J.C., Arima, J., and Mori, N. (2015). Glassin, a histidine-rich protein from the siliceous skeletal system of the marine sponge Euplectella, directs silica polycondensation. Proc. Natl. Acad. Sci. U. S. A. *112*, 11449–11454.

Sreenath, T., Thyagarajan, T., Hall, B., Longenecker, G., D'Souza, R., Hong, S., Wright, J.T., MacDougall, M., Sauk, J., and Kulkarni, A.B. (2003). Dentin sialophosphoprotein knockout mouse teeth display widened predentin zone and develop defective dentin mineralization similar to human dentinogenesis imperfecta type III. J. Biol. Chem. *278*, 24874–24880.

Stetler-Stevenson, W.G., and Veis, A. (1986). Type I collagen shows a specific binding affinity for bovine dentin phosphophoryn. Calcif. Tissue Int. *38*, 135–141.

Timpl, R., Wiedemann, H., van Delden, V., Furthmayr, H., and Kühn, K. (1981). A network model for the organization of type IV collagen molecules in basement membranes. Eur. J. Biochem. *120*, 203–211.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

Urry, D.W., Long, M.M., Cox, B.A., Ohnishi, T., Mitchell, L.W., and Jacobs, M. (1974). The synthetic polypentapeptide of elastin coacervates and forms filamentous aggregates.

Biochim. Biophys. Acta BBA - Protein Struct. *371*, 597–602.

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.e16.

Werner, A., Iwasaki, S., McGourty, C.A., Medina-Ruiz, S., Teerikorpi, N., Fedrigo, I., Ingolia, N.T., and Rape, M. (2015). Cell-fate determination by ubiquitin-dependent regulation of translation. Nature *525*, 523–527.

Werner, A., Baur, R., Teerikorpi, N., Kaya, D.U., and Rape, M. (2018). Multisite dependency of an E3 ligase controls monoubiquitylation-dependent cell fate decisions. ELife *7*, e35407.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Wootton, J.C., and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. *17*, 149–163.

Xu, M., and Lewis, R.V. (1990). Structure of a protein superfiber: spider dragline silk. Proc. Natl. Acad. Sci. *87*, 7120–7124.

Yang, P., Mathieu, C., Kolaitis, R.-M., Zhang, P., Messing, J., Yurtsever, U., Yang, Z., Wu, J., Li, Y., Pan, Q., et al. (2020). G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. Cell *181*, 325-345.e28.

Youn, J.-Y., Dunham, W.H., Hong, S.J., Knight, J.D.R., Bashkurov, M., Chen, G.I., Bagci, H., Rathod, B., MacLeod, G., Eng, S.W.M., et al. (2018). High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. Mol. Cell *69*, 517-532.e11.