A unified view of 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

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

57 More broadly, the importance of LCR features and relationships discussed above is not
58 restricted to proteins of intracellular higher order assemblies. Structural assemblies such as the
59 extracellular matrix (Forgacs et al., 2003; Rauscher and Pomès, 2017), spider silk (Xu and
60 Lewis, 1990; Hinman and Lewis, 1992; Malay et al., 2020), and the siliceous skeleton of certain
61 sponges (Shimizu et al., 2015) are comprised of proteins which share features with proteins
62 involved in intracellular assemblies, such as multivalent scaffolding proteins abundant in LCRs.
63 In fact, many of the proteins comprising these assemblies are composed of almost entirely
64 LCRs which are known to mediate multivalent interactions (Rauscher et al., 2006; Malay et al.,
65 2020). Thus, despite having vastly different functions and emergent physical properties, valency
66 and hierarchical assembly seem to also play a role in these diverse extracellular assemblies.
67 Given that LCRs are required for such diverse assemblies, how diverse are the sequences of
68 natural LCRs, especially given their low sequence complexity? How do differences in the
69 sequences, biophysical properties, copy number and type of LCRs correspond to differences in
70 the higher order assemblies which they form? A unified view of LCRs which incorporates the
71 sequences, features, relationships, and functions of LCRs may allow us to both gain detailed
72 insights into how specific LCRs contribute to higher order assemblies, and gain a broader
73 understanding of LCR sequences and their corresponding functions.

74 Here, we use systematic dotplot analysis to provide a comprehensive, unified view of
75 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

102 a dense square region centered on the diagonal. Moreover, for proteins with multiple LCRs,
103 compositionally similar LCRs will result in dense square regions off of the diagonal in the
104 position corresponding to the intersection of both LCRs, but different LCRs will not intersect in
105 this way. Therefore, dotplots are capable of identifying both the total number of LCRs in proteins
106 and the relationships between similar and distinct LCRs for proteins with multiple LCRs.

107 For example, in the dotplot of G3BP1, a protein important for stress granule assembly
108 (Yang et al., 2020), the dense squares along the diagonal clearly distinguish between the LCRs
109 of G3BP1 and its other regions, which include its RNA-recognition motif (RRM) and dimerization
110 domains (Figure 1A, dotted black outlines). Immediately apparent from the dotplot of G3BP1 is
111 that its LCRs are not all the same type as dense squares do not occupy every intersection off
112 the diagonal (Figure 1A, red outlines). The first two LCRs are acidic in composition, while the
113 third is an RGG domain which plays a role in RNA-binding (Kim et al., 2019a). The presence of
114 these compositionally distinct LCRs are critical for the ability of G3BP1 to form stress granules,
115 as the acidic LCRs interact with and inhibit the RGG domain, preventing it from interacting with
116 RNA, a necessary step of stress granule assembly (Guillén-Boixet et al., 2020; Yang et al.,
117 2020). Thus, by highlighting the relationships between different LCRs, dotplots can provide key
118 insights relevant to protein function.

119 While not all proteins have LCRs, some proteins almost entirely consist of LCRs and
120 exhibit diverse LCR relationships and organization. For example, ACTB and SYTC lack LCRs,
121 which is reflected by the lack of dense squares in their respective dotplots (Figure 1 - figure
122 supplement 1A, B). Other examples, such as SMN, a component of nuclear gems (Liu and
123 Dreyfuss, 1996), and the nucleolar protein KNOP1 (Grasberger and Bell, 2005; Larsson et al.,
124 1999) have more complex architectures, with multiple copies of similar LCRs, which appear with
125 roughly equal spacing (Figure 1 - figure supplement 1C, D). On the other hand, Nucleolin has
126 multiple types of LCRs, one of which occurs in several copies. Interestingly, these LCR types
127 are spatially segregated in the protein, with one set of LCRs in the N-terminal region, and a

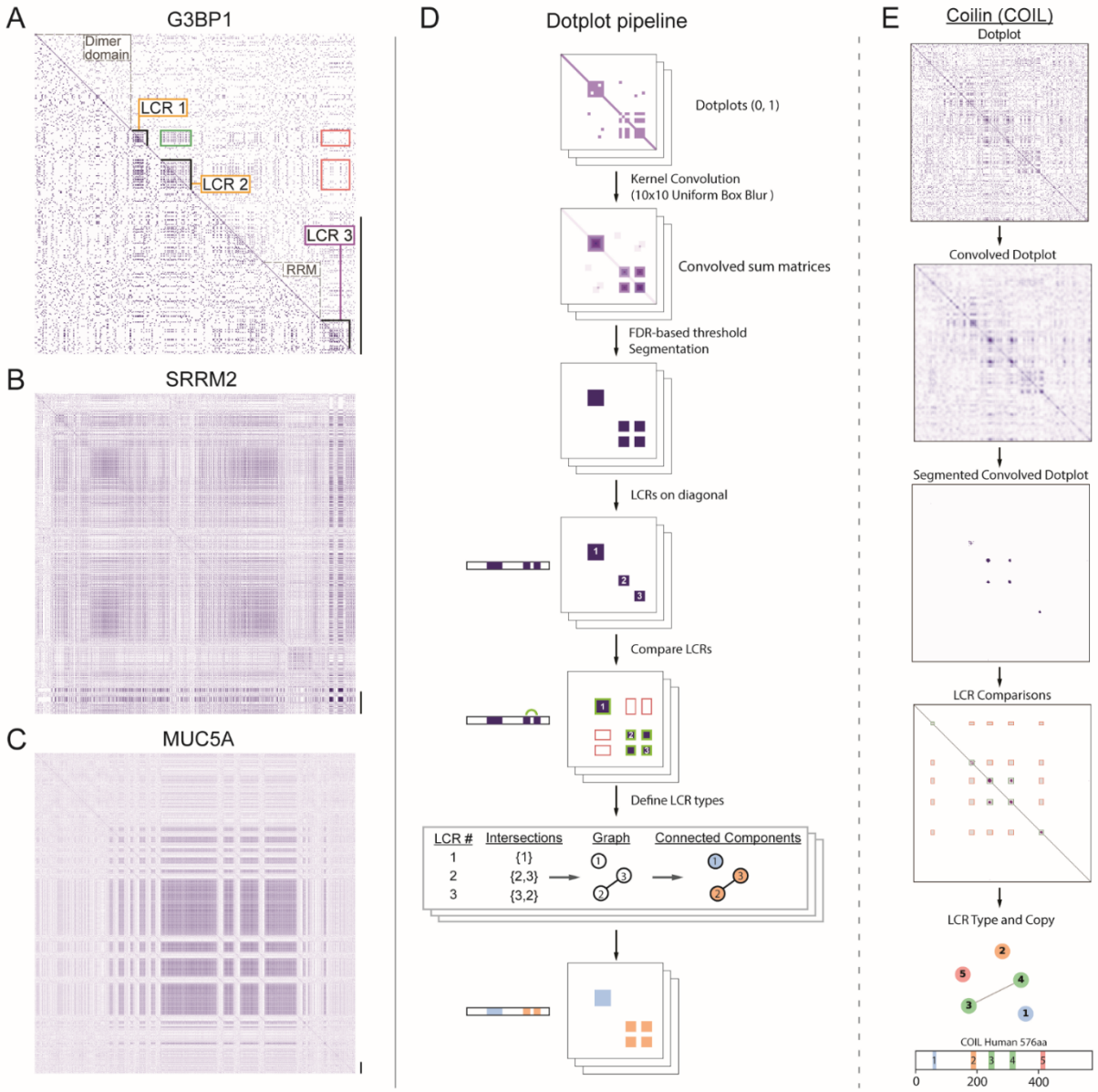


Figure 1: A systematic dotplot approach to reveal the relationships between low complexity regions (LCRs) in proteins

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.

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

130 As can be seen for SRRM2 and MUCIN5A, a large area in their dotplots consist of LCR
131 signatures off the diagonal (Figure 1B, C), indicating that each of these proteins consist of long
132 stretches of similar LCR sequences. For example, the dotplot of SRRM2 contains multiple
133 regions of low complexity which cover an area corresponding to hundreds of amino acids
134 (Figure 1A, B). SRRM2 and another LCR-containing protein SON (Figure 1 - figure supplement
135 1H) were recently found to act as essential scaffolds for formation of nuclear speckles (Sharma
136 et al., 2010; Fei et al., 2017; Ilik et al., 2020), suggesting that proteins which each contain long
137 stretches of similar LCR sequences could play important roles in certain higher order
138 assemblies. In fact, many such proteins have been found to be essential for various higher
139 order assemblies. These include UBP2L and PRC2C (Figure 1 - figure supplement 1F, G),
140 which were only recently discovered to be essential for the formation of stress granules (Youn et
141 al., 2018; Sanders et al., 2020). UBP2L was found in some conditions to be upstream of G3BP1
142 for stress granule formation (Cirillo et al., 2020).

143 Other proteins with long stretches of similar LCR sequences included mucins (MUC5A
144 shown, Figure 1C), collagens and DSPP (Figure 1 - figure supplement 1I), proteins which are
145 essential to the formation of extracellular assemblies with a diverse variety of physical
146 properties. Mucins are key components of mucus, a liquid/gel-like assembly of glycoproteins
147 (reviewed in (Lai et al., 2009)), while DSPP codes for a protein which scaffolds the
148 mineralization of teeth (Stetler-Stevenson and Veis, 1986; Saito et al., 2000; Sreenath et al.,
149 2003). Although proteins which each contain such long stretches of similar LCRs, such as
150 SRRM2, UBP2L, MUCIN5A, and DSPP, are involved in such diverse biological processes, a
151 commonality among them is their scaffolding roles. The fact that these proteins exhibit similar
152 LCR relationships and roles in their respective assemblies suggests that the LCR relationships
153 revealed by dotplots can inform how we understand protein functions.

154 The examples of dotplots make clear that functional information about LCR type and
155 copy number can be extracted from dotplot matrices. LCR type has been shown in a handful of
156 examples to be important for how these proteins, such as G3BP1 (Guillén-Boixet et al., 2020;
157 Yang et al., 2020), function biologically. Additionally, copy number is likely an important
158 contributor to valency for proteins which contain such a large proportion of LCRs, such as
159 SRRM2 or MUC5A. However, there currently is not an approach to assess the global
160 relationship between these features of LCRs and their functions. While several methods exist
161 for identifying LCRs (Wootton and Federhen, 1993; Promponas et al., 2000; Albà et al., 2002;
162 Harrison, 2017), these methods are unable to determine LCR types and their respective copy
163 numbers. As a consequence, we have not been able to systematically understand how LCR
164 sequence and organization influence their function. The ability of dotplots to both identify LCRs
165 and provide information on LCR type and copy number presents an opportunity to develop a
166 comprehensive and systematic tool to identify these features of proteins.
167
168 **A systematic dotplot approach to identify and characterize LCRs proteome-wide**
169 We developed a computational pipeline to extract both the positions and spatial
170 relationships of LCRs using the 2D signature of LCRs in dotplots (Figure 1D, Methods).
171 Because LCRs present themselves as dense squares in a 2D matrix, the identification of LCRs
172 in dotplots was similar to the extraction of features from an image, allowing us to take
173 advantage of image processing tools.
174 Specifically, we computationally extracted the LCRs of any protein by identifying high
175 density regions in its dotplot through classic image processing methods, such as kernel
176 convolution, thresholding, and segmentation (Figure 1D). To identify high density regions in
177 dotplots, we performed kernel convolution on the dotplots with a uniform 10x10 kernel, which
178 calculates a convolved pixel intensity value from 0 to 100 based on the number of dots in that

179 window. Regions of high density will have higher convolved pixel intensities, while regions of
180 low density will have lower convolved pixel intensities.

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

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

204 Finally, when compared to two commonly used LCR-callers, SEG (Wootton and
205 Federhen, 1993) and fLPS (Harrison, 2017), our approach achieves a comparable or better
206 performance in minimizing LCR entropy while maximizing total LCR sequence in the human
207 proteome (Figure 1 - figure supplement 3A-D). Furthermore, we call LCRs in regions of proteins
208 similar to those called by other methods, as can be seen in examples CO1A1 and ZN579
209 (Figure 1 - figure supplement 3E, F). While other approaches (Wootton and Federhen, 1993;
210 Harrison, 2017) are more efficient at identifying the presence of LCRs, our approach allows for
211 proteome-wide identification of LCRs without losing information about LCR type and copy
212 number within proteins. Thus, by making 2D comparisons of LCRs within proteins across the
213 proteome, our systematic dotplot approach provides more information on the relationship
214 between LCRs within proteins, allowing us to ask deeper questions about the role of these
215 features in protein function.
216
217 **Comparison of LCRs defines type and copy number of LCRs across the proteome**
218 The relationship between LCRs within LCR-containing proteins has not been studied on
219 a proteome-wide scale, despite being important in the cases where it has been studied (Hebert
220 and Matera, 2000; Mitrea et al., 2016; Yang et al., 2020). These relationships can now be
221 studied systematically using our dotplot approach. To this end, we compared all LCRs within
222 each protein for the human proteome. The relationship between two LCRs in a protein is
223 determined by whether or not a segmented region of the dotplot exists off the diagonal in the
224 region corresponding to the intersection of those two LCRs. If so, we designate these two
225 compositionally similar LCRs as the same ‘type’ (green boxes in Figure 1D , Methods). The
226 relationships between LCRs can be summarized as a graph where each LCR is a node, and off-
227 diagonal intersections between pairs of LCRs are represented as edges (Figure 1D, Methods).
228 Within each of these graphs, individual connected components are LCRs of the same type. The
229 number of nodes is the total number of LCRs, while the number of connected components

230 defines the number of distinct LCR types. This graph-based visualization is helpful for seeing
231 LCR relationships within proteins, particularly for proteins with many LCRs (Figure 1 - figure
232 supplement 4), and contributes to our understanding of potential valency provided by LCRs to
233 natural proteins.

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

240 With this, we compared the number of total and distinct LCRs for each protein (Figure
241 2A). We can see from this analysis that the range in combinations of total and distinct LCRs is
242 diverse across proteins in the human proteome (Figure 2A, Figure 2 - figure supplement 1A),
243 enabling different combinations of LCRs and functions. Based on the number of total and
244 distinct LCRs in a given protein, proteins can be broadly categorized into four groups (Figure
245 2B), which each make a sizable fraction of the proteome and uniquely contribute to our
246 understanding of how LCRs affect protein-function. We will refer to these groups as ‘single’,
247 ‘multiple-same’, ‘multiple-distinct’, and ‘multiple-mixed’ to reflect the number of total and distinct
248 LCRs that a protein possesses. The single LCR group, which lies in the bottom left corner
249 (Figure 2A), corresponds to proteins with only a single LCR, in which we may assess the
250 isolated function of an LCR. The multiple-same group lies along the vertical axis and
251 corresponds to proteins with multiple LCRs, all of which are the same type (total LCRs >1,
252 distinct LCRS=1). Since all of the LCRs for a given protein in this group are the same, this group
253 is particularly useful for understanding the contribution of LCR copy number to the function of a
254 protein. The multiple-distinct group lies along the diagonal, and corresponds to proteins with
255 multiple LCRs, all of which are distinct from each other (total LCRs=distinct LCRs >1). This

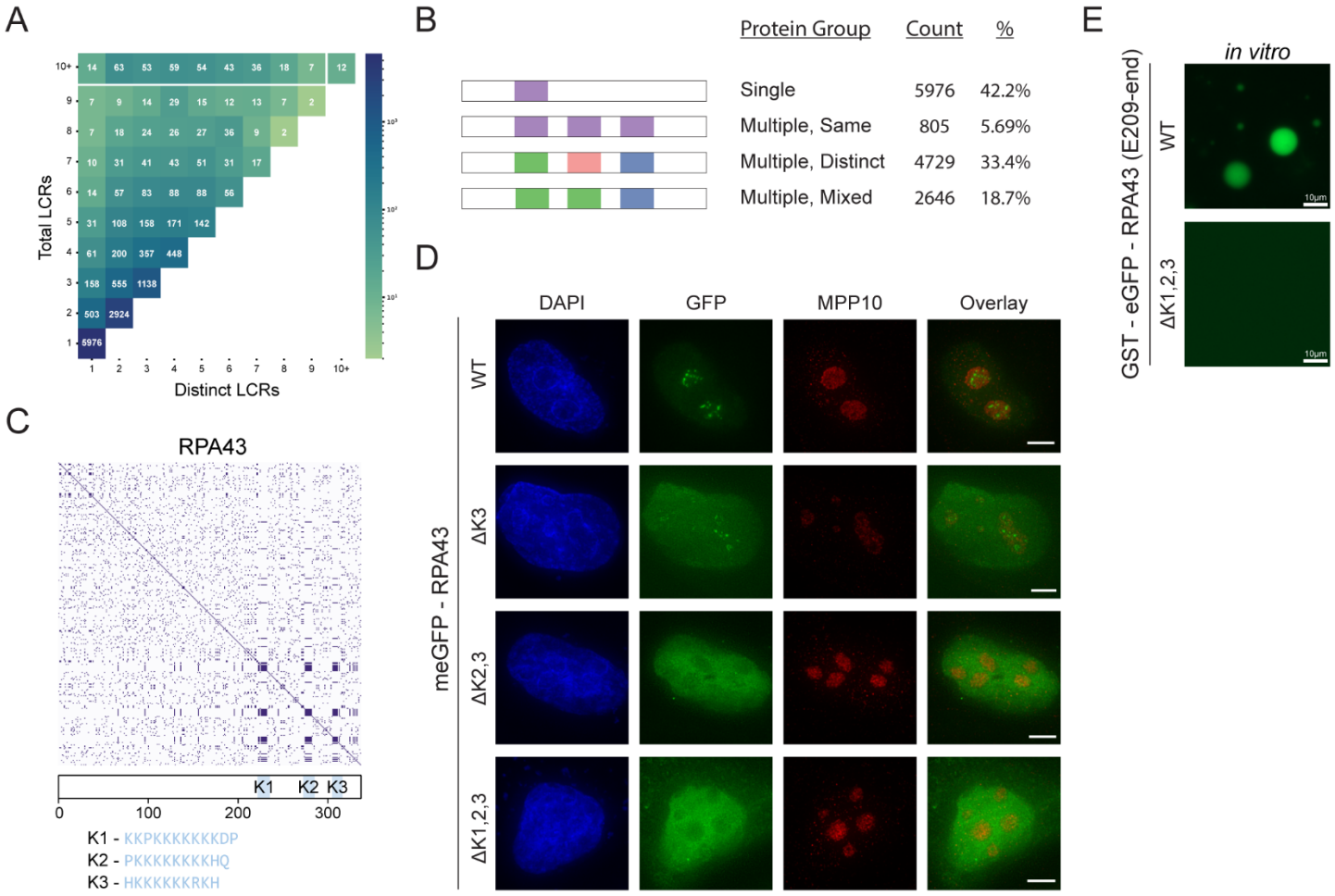


Figure 2: Proteome-wide definition of LCR type and copy number reveals copy number requirements for nucleolar integration of RPA43

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 1 - figure supplement 1,2,3,4.

256 group allows for in-depth study of the relationships between different LCRs. Finally, the multiple-
257 mixed group, which occupies the rest of the graph, corresponds to proteins with multiple LCRs
258 of mixed types, where at least one type is present in at least two copies. This group likely
259 corresponds to more complex proteins which may be affected by both the copy number and
260 type of LCRs they contain. By characterizing the copy number and type of LCRs across the
261 proteome, our approach allows for proteins to be selected on the basis of these features for
262 further study.
263
264 **LCR copy number impacts protein function**
265 The group of proteins which have multiple LCRs of the same type presents an
266 opportunity to specifically understand the role of LCR copy number in natural proteins. To
267 highlight how these groups could inform us on LCR function, we sought to study the role of LCR
268 copy number on higher order assembly by studying a protein in the 'multiple same' group.
269 We chose to study the RNA Polymerase I component RPA43, which localizes to the
270 nucleolus (Dundr et al., 2002), a multi-component higher order assembly. RPA43 has three
271 LCRs in its C-terminus which are all the same type (Figure 2C, Figure 1 - figure supplement
272 4A). To understand the common sequences in this LCR type, we manually checked the
273 sequences determined by our systematic analysis. All three LCRs of RPA43 contained a 10-12
274 amino acid block of mostly K-residues (Figure 1 - figure supplement 4A, bottom row), which
275 were the primary contributor to off-diagonal intersections between these LCRs and thus defined
276 this LCR type. In order to test the importance of LCR copy number in RPA43 function, we chose
277 to focus on the sequences in its three LCRs which make them the same type, the blocks of K-
278 residues. We will refer to these K-rich blocks as K-rich LCRs of RPA43 (K1, K2, and K3
279 respectively).
280 While GFP-fused WT RPA43 localized correctly to the fibrillar center of the nucleolus,
281 deletion of all three of its K-rich LCRs (ΔK1,2,3) led to its exclusion from the nucleolus,

282 confirming that these LCRs are important for its higher order assembly (Figure 2D). It is
283 important to note that this mutant retains the predicted disordered nature of RPA43s C-terminal
284 region (Figure 2 - figure supplement 1B), showing that this phenotype cannot be explained by a
285 predicted change in the solvent accessibility of this region. The fact that all of the LCRs of
286 RPA43 are the same type, and that they together are required for nucleolar integration allows us
287 to specifically study the role of LCR copy number in RPA43 higher order assembly.

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

297 Consistent with these results, while the recombinant GFP-fused RPA43 C-terminus
298 phase separated into liquid droplets *in vitro*, the GFP-fused RPA43 C-terminus with its three K-
299 rich LCRs specifically deleted did not (Figure 2E). Thus, the RPA43 C-terminus contains the
300 sequences sufficient for higher order assembly, and the K-rich LCRs are necessary for this
301 assembly. This result suggests that the K-rich LCRs are not merely linkers between other self-
302 interacting elements, as deletion of such linkers tends to alter the physical properties of the
303 assembly but not its presence (Martin et al., 2020). Rather, these results suggest that these K-
304 rich LCRs either self-interact (perhaps mediated by solvent anions), or interact with a
305 complementary element in the RPA43 C-terminus. Moreover, the observation that *in vivo*
306 nucleolar localization and *in vitro* phase separation require the same sequences suggest that
307 the interactions mediating RPA43's ability to form a higher order assembly are similar to those

308 mediating its nucleolar integration. Together, these results not only highlight the ability of K-rich
309 LCRs to mediate higher order assembly of RPA43, but highlight the importance of
310 understanding LCR copy number more generally. Thus, our approach allows for targeted
311 experiments to arrive at principles by which LCR copy number affects protein function.
312
313 **A map of LCRs**
314 In order to relate the sequences, features, relationships, and functions of LCRs across
315 the proteome, we wanted to understand the full breadth of LCR sequences. By using a
316 sequence map as a foundation to integrate these aspects of LCRs, we could begin to
317 understand how differences in sequence correspond to differences in the features,
318 relationships, and functions of LCRs. As such we took an unbiased approach to visualize the
319 sequence space occupied by LCRs in the human proteome.
320 Using the LCRs identified by dotplots, we represented the amino acid composition of
321 each LCR as a 20-dimensional vector where each dimension corresponds to the frequency of a
322 different amino acid. Thus, each LCR will map to a point in 20-dimensional sequence space. To
323 visualize LCR occupancy in this sequence space, we used Uniform Manifold Approximation and
324 Projection (UMAP) (McInnes et al., 2020) to generate a 2-dimensional map of all LCRs in the
325 human proteome (Figure 3A, Figure 3 - figure supplement 1).
326 We can immediately see from this map that LCRs in the human proteome exhibit a rich
327 diversity of sequence compositions, and do not fall into a handful of isolated groups. Generally,
328 we see that this LCR space has many highly occupied regions (Figure 3A). Using Leiden
329 clustering (Traag et al., 2019), we identified 19 clusters of LCRs in this space which mostly
330 corresponded to high frequency of an amino acid, and serve as useful guides when referring to
331 different regions of the map (Figure 3A). While most of these clusters correspond to LCRs with
332 large contributions from a single amino acid, these clusters still have a substantial presence of
333 many other amino acids. For example, the serine-rich cluster has regions within it that are also

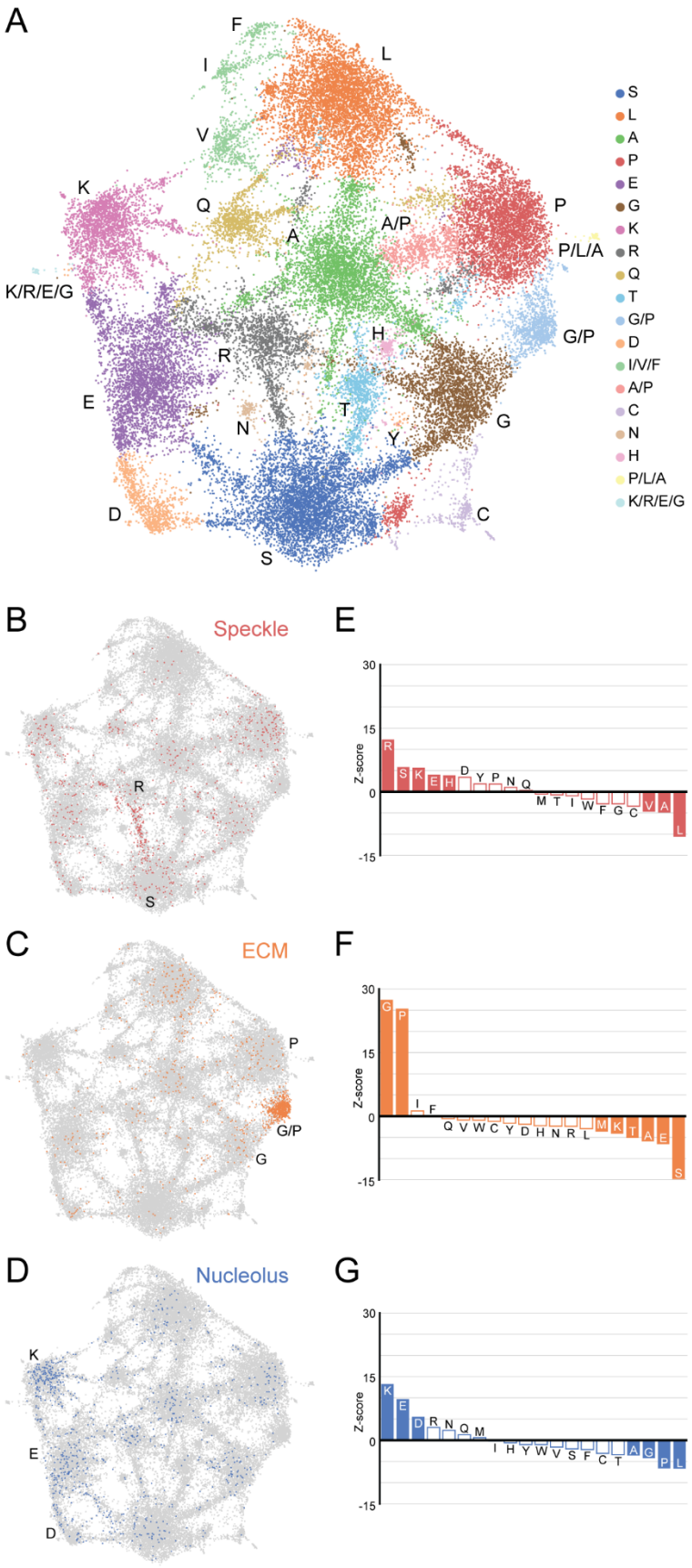


Figure 3: A map of LCRs captures known differences in higher order assemblies

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.

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

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

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

354

355 **Higher order assemblies map to specific regions in LCR sequence space**

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

359 To do this, we mapped annotations of known higher order assemblies to the LCR map.
360 Nuclear speckle proteins, which are commonly localized by LCRs known as RS-domains
361 (Cáceres et al., 1997; Boucher et al., 2001), populated a bridge between the R and S clusters in
362 LCR sequence space (Figure 3B), and were significantly enriched in both of these amino acids
363 (Figure 3E). LCRs of extracellular matrix (ECM) proteins were heavily concentrated in a G/P-
364 rich region (Figure 3C, F), reflecting the many, long collagen proteins in humans. LCRs of
365 nucleolar proteins mapped to the K-rich cluster in LCR sequence space (Figure 3D, G),
366 consistent with nucleolar localization signals possessing K-rich sequences (Scott et al., 2010).
367 Other higher order assemblies also mapped to specific regions in the LCR sequence space,
368 including the centrosome and nuclear pore complex (Figure 3 - figure supplement 3A, B).
369 Wilcoxon rank-sum tests for each of the 20 amino acids confirmed that these spatial biases in
370 the LCR map corresponded to actual differences in LCR composition, independent of Leiden
371 cluster assignment (Figure 3E-G, Figure 3 - figure supplement 3E, F). Conversely, some higher
372 order assemblies are known to not have many individual proteins which share a specific type of
373 LCR, including stress granules for which RNA is a major contributor (Guillén-Boixet et al., 2020;
374 Sanders et al., 2020), and PML bodies which depend on SUMOylation of non-LC sequences
375 (Shen et al., 2006). As expected for these cases, there was neither a spatial bias in the LCR
376 map, nor significantly enriched amino acids (Figure 3 - figure supplement 3C, D, G, H).
377 The ability of the map to highlight the biased LCR compositions of certain higher order
378 assemblies demonstrates that we can capture how differences in sequence correspond to
379 known differences in function. Thus, the LCR map allows us to interrogate the relationship
380 between less understood regions of LCR space and protein function.
381
382 **A unified view of LCRs reveals the prevalence of E-rich LCRs in nucleolar proteins**
383 When examining the distribution of nucleolar protein LCRs across the LCR map, we
384 found that in addition to the K-rich cluster, the E-rich cluster was significantly occupied (Figure

385 3D, G). While there is evidence for the importance of acidic stretches in phase separated bodies
386 (Hebert and Matera, 2000; Mitrea et al., 2016; Guillén-Boixet et al., 2020; Yang et al., 2020),
387 nucleolar LCRs were significantly more likely than LCRs of speckle proteins to have a high
388 frequency of E residues, but not D residues (Figure 4A). This observation suggested that E-rich
389 LCRs may play a nucleolus-specific role.

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

400 By looking at the regions in the LCR map occupied by K- and E-rich LCRs, we noticed
401 that while both K-rich and E-rich LCRs are predicted to be similarly disordered (Figure 3 - figure
402 supplement 4A), the E-rich LCRs had higher ANCHOR scores overall (Figure 3 - figure
403 supplement 4B). To determine if this trend held true for K-rich and E-rich LCRs of nucleolar
404 proteins in particular, we analyzed LCRs in the top 25th percentile of K or E frequency in the
405 nucleolus, which we refer to as K-enriched and E-enriched respectively (Figure 4B). The
406 distributions of IUPred2 scores among these K- and E-enriched nucleolar LCRs are only subtly
407 different (Figure 4C), consistent with what was gleaned from the LCR map. However, while K-
408 enriched nucleolar LCRs exhibited a unimodal distribution of ANCHOR scores centered below
409 0.5, E-enriched LCRs exhibited a bimodal distribution (Figure 4D). One peak of this distribution
410 was at approximately the same ANCHOR score as nucleolar K-rich LCRs (Figure 4D).

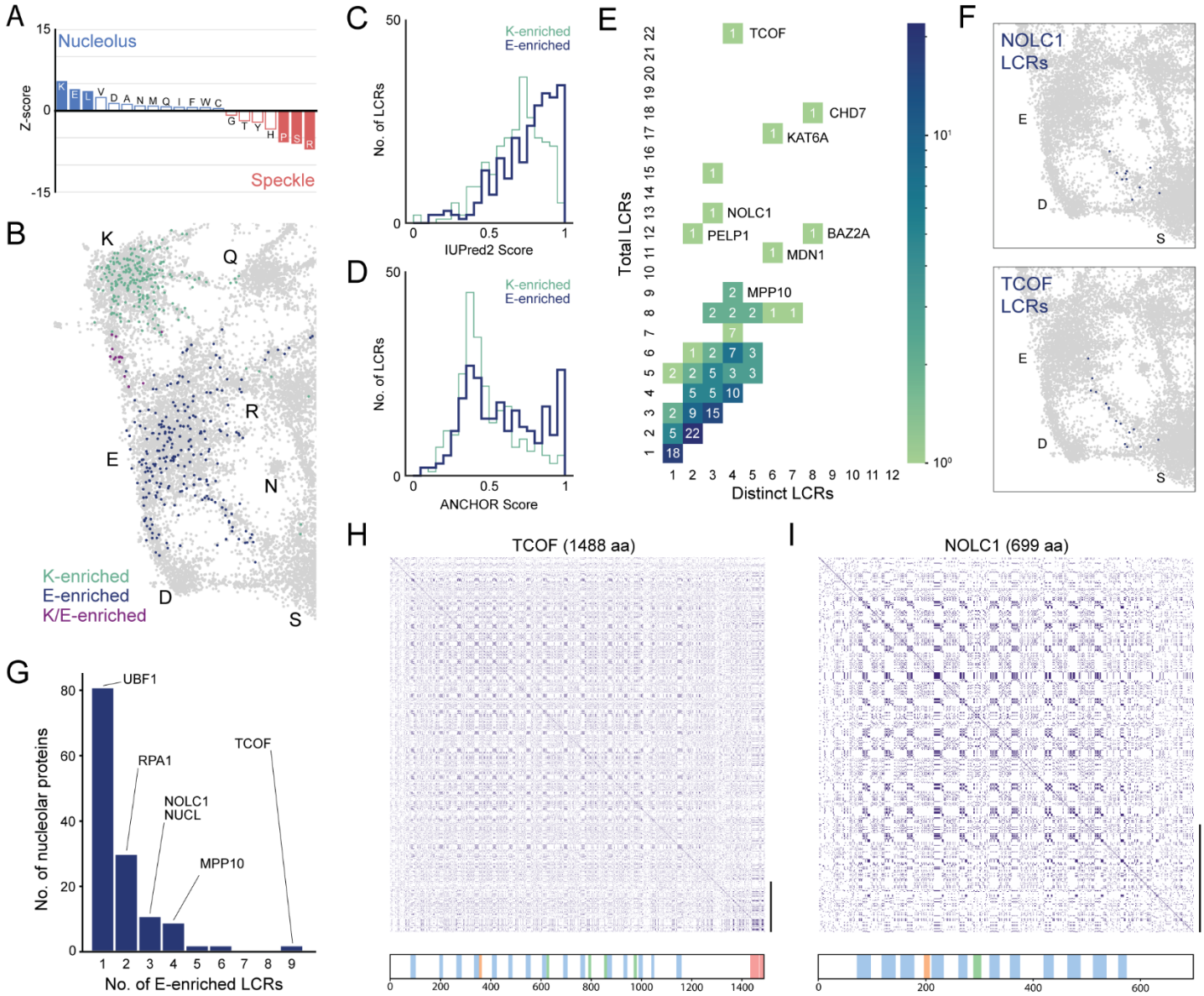


Figure 4: An integrated LCR map reveals the distribution and prevalence of E-rich LCRs among nucleolar proteins

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.

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

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

426 Proteins such as NOLC1 and TCOF have a high number of S and E-rich LCRs. In the
427 total vs. distinct LCR comparison across the proteome, they were also high in total LCRs, while
428 low in distinct LCRs (Figure 4E, G), suggesting that they might be scaffold-like proteins. NOLC1
429 and TCOF have a striking pattern of several evenly spaced E-rich LCRs, illustrated by their
430 dotplots and UMAPs (Figure 4F, H, I, Figure 1 - figure supplement 4D, Figure 4 - figure
431 supplement 1). Furthermore, these regions are required for NOLC1 and TCOF to interact with
432 each other (Werner et al., 2015, 2018). Consistent with their scaffold-like property, NOLC1 and
433 TCOF have been shown to recruit proteins important for key nucleolar functions, including key
434 components of the Pol I transcription machinery (Werner et al., 2015, 2018). We find that most
435 nucleolar E-enriched LCR-containing proteins contain only one or two of these LCRs (Figure
436 4G), and may act as clients in the nucleolus with respect to interactions mediated by E-rich

437 LCRs. These proteins include the catalytic component of RNA Pol I (RPA1) and the primary
438 transcription factor for rDNA transcription (UBF1). The implication of E-rich LCRs in TCOF-
439 NOLC1 interaction suggests that proteins with a high number of E-rich LCRs might enable
440 nucleolar assembly through the recruitment of other E-rich containing proteins.
441 Thus, by integrating the sequences, features, and relationships of LCRs, such a unified
442 view provides a framework for understanding of the role of LCRs in higher order assembly.
443
444 **An expanded map of LCRs across species**
445 The spatial relationship between regions of LCR space and higher order assemblies
446 raises questions about if there is a conserved relationship between LCR sequence composition
447 and the functions and physical properties of their parent proteins. In species where the
448 existence of a given assembly such as the nucleolus is conserved, is occupancy of the
449 sequence space also conserved? Similarly, does emergence of a certain higher order assembly
450 across evolution such as the extracellular matrix correlate with the occupancy of a certain region
451 of sequence space? Conversely, many species have distinct higher order assemblies with
452 different functions and physical properties from those in humans, such as plants and fungal cell
453 walls. Do these assemblies occupy a region of sequence space that is distinct from that
454 occupied in humans, or do they use a sequence space that is occupied in humans?
455 To answer these questions, we wanted to capture the entire breadth of LCR sequence
456 space across species, so that we could concurrently compare how they occupy this sequence
457 space. We applied our dotplot and dimensionality reduction approach to the proteomes of *E.*
458 *coli*, *S. cerevisiae*, *A. thaliana*, *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, and *H.*
459 *sapiens*. This allowed us to simultaneously compare between prokaryotes and eukaryotes,
460 among fungi, plants, and animals, and across metazoans.
461 We first generally compared the number of LCRs and LCR-containing proteins between
462 species. At our stringent false discovery rate of 0.002, *E. coli* did not have any LCRs, while

463 other species had many LCRs. So, we relaxed the FDR for *E. coli* to 0.05 to allow further
464 analysis. While E. coli and S. cerevisiae had fewer LCRs, multicellular species had more, with
465 both the number of LCRs and LCR-containing proteins increasing across metazoans (Figure 5 -
466 figure supplement 1A, B). In particular, *D. melanogaster* had the most LCRs and highest
467 proportion of LCR-containing proteins of the species analyzed Figure 5 - figure supplement 1A,
468 B), consistent with previous studies (Huntley and Clark, 2007). The distribution of entropies for
469 LCRs varied between species, but were all significantly lower than that of length-matched,
470 randomly-sampled sequences from their respective proteomes (Figure 5 - figure supplement
471 1C).

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

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

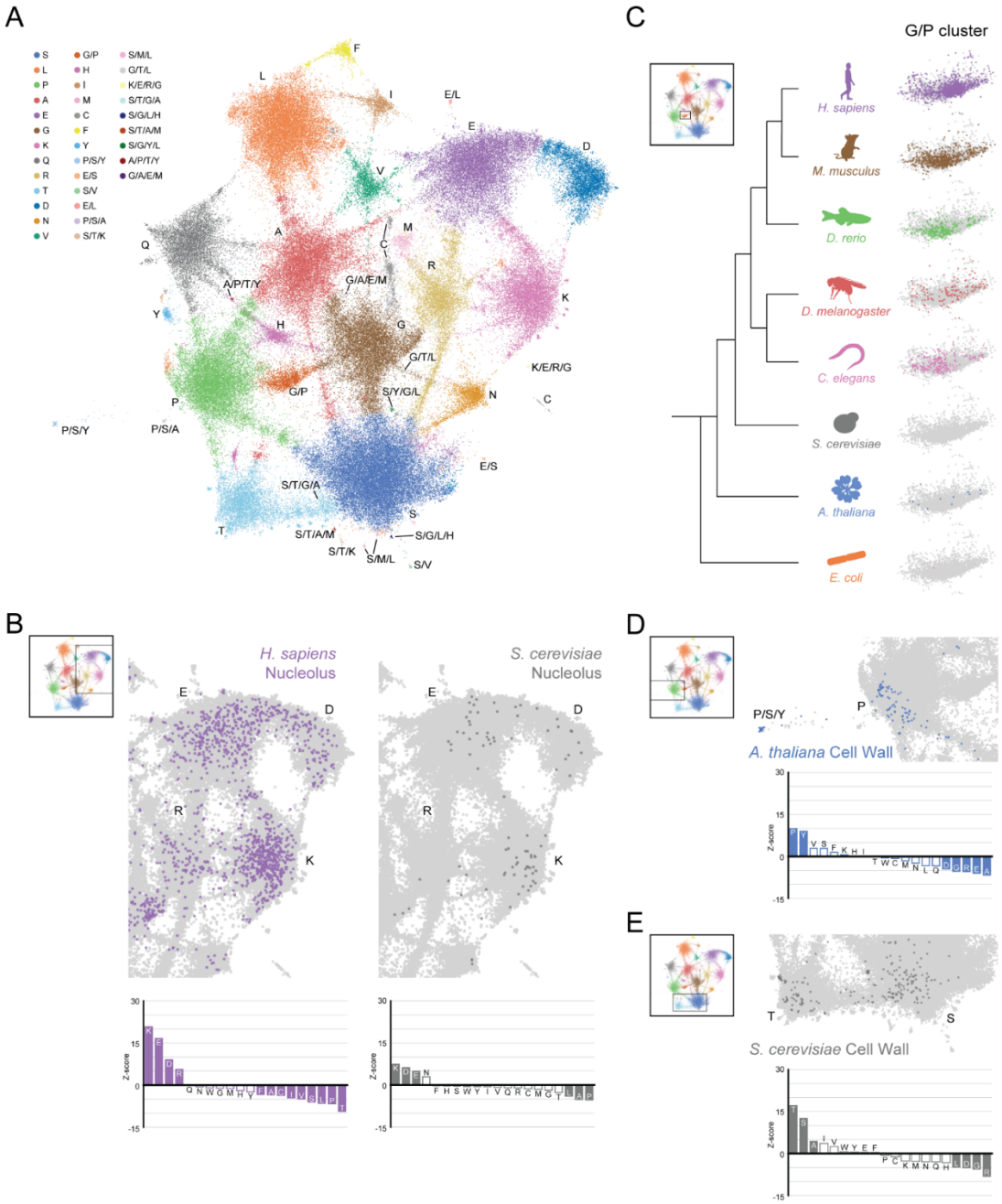


Figure 5: The conservation and emergence of higher order assemblies is captured in an expanded LCR map across species

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 3 - figure supplement 1,2,3,4,5,6.

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

495 Many of the spatial relationships between clusters were maintained in this expanded
496 LCR sequence space. For example, certain bridges were observed in several species, while a
497 few bridges were predominantly occupied in specific species, indicating that amino acid
498 combinations in LCRs differed between species. For example, the H-rich cluster and its bridge
499 connecting to the Q-rich cluster was expanded in *D. melanogaster* (Figure 5 - figure supplement
500 2, 3, 4C). LCRs in the Q/H-rich bridge have generally higher ANCHOR scores than those in the
501 main Q-rich cluster (Figure 5 - figure supplement 5B), suggesting that H residues are combined
502 with varying degrees of Q residues to allow for more nuanced differences in LCR properties.

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

506

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

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

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

520 Furthermore, changes in the LCR sequence space corresponded to differences in higher
521 order assemblies, such as the extracellular matrix which occupied the G/P cluster in humans.
522 While *E. coli*, *S. cerevisiae*, and *A. thaliana* had nearly no LCRs in the G/P cluster, this cluster
523 was much more occupied in metazoans (Figure 5C), corresponding with the emergence of
524 collagens, a hallmark of the metazoan lineage (Hynes, 2012). This difference in G/P occupancy
525 could not be explained by differences in the total number of LCRs in these species, since *A.*
526 *thaliana* had more LCRs than *C. elegans* but much lower occupancy in the G/P cluster. When
527 looking across metazoans, although this cluster was occupied in *C. elegans* and *D.*
528 *melanogaster*, it was more heavily occupied in vertebrates. Many LCRs existed in this cluster in
529 *D. rerio*, and even more in *M. musculus*, and *H. sapiens*, which spanned most of the space in
530 this G/P cluster. Again, the difference in G/P cluster occupancy could not be explained by the
531 total number of LCRs in each species, as *D. melanogaster* had more LCRs than all of the
532 vertebrates, but lower G/P occupancy (Figure 5C, Figure 5 - figure supplement 3). The gradual
533 differences in occupancy of the G/P cluster between the metazoan species (Figure 5C)
534 correlated with the expansion of the extracellular matrix across metazoans (reviewed in (Hynes,
535 2012)), highlighting that the LCR map traces the progression of a higher order assembly across
536 evolution.

537 Across longer evolutionary distances, different species-specific higher order assemblies
538 mapped to unique regions of LCR sequence space. LCRs of cell wall proteins of *A. thaliana*, for
539 example, primarily mapped to the P-rich cluster and a nearby P/S/Y-rich cluster (Figure 5D,
540 Figure 5 - figure supplement 6G), reflecting the set of hydroxyproline-rich cell wall proteins

541 which include extensins, arabinogalactan proteins (AGPs) and proline-rich proteins (PRPs).
542 Extensins, which have SPPPP motifs, are known to be important scaffolds for the assembly of
543 the cell wall, in which they are thought to form self-assembling networks to organize pectin
544 (Cannon et al., 2008; Sede et al., 2018). In *S. cerevisiae*, LCRs of cell wall proteins mapped to
545 the S-rich and T-rich clusters (Figure 5E, Figure 5 - figure supplement 6H), which included
546 flocculation proteins. These S- and T-rich LCRs are often sites for O-mannosylation in
547 mannoproteins, which is crucial for the integrity of the cell wall (Gentzsch and Tanner, 1996;
548 González et al., 2012; Neubert et al., 2016).
549 Our approach allowed us to answer several general questions about the relationships
550 between the sequence space occupied by LCRs and their functions. Firstly, we show that when
551 a given assembly is conserved, occupancy of the corresponding LCR sequence space is also
552 conserved. Secondly, the emergence of a higher order assembly can correspond to the
553 population of a previously unoccupied sequence space. Finally, higher order assemblies with
554 different physical properties occupy different regions of sequence space, even when they fulfill
555 similar roles in their respective species. While these principles may not always hold for every
556 sequence space or assembly, they may guide how we interpret the spaces and assemblies
557 which have yet to be explored.
558
559 **A teleost-specific T/H cluster contains scaffold-like proteins**
560 Given the relationship between some regions of LCR sequence space and higher order
561 assemblies, we looked for previously undescribed regions of sequence space which are
562 differentially occupied across species.
563 We found various species-specific regions which lacked detailed annotations, one of
564 which was the T/H-rich cluster specific to *D. rerio* (Figure 6A, Figure 5 - figure supplement 3).
565 Many of the LCRs in this cluster included direct TH repeats (Figure 6A), making this a cluster of
566 particular interest because these amino acid residues may have properties of mixed-charge

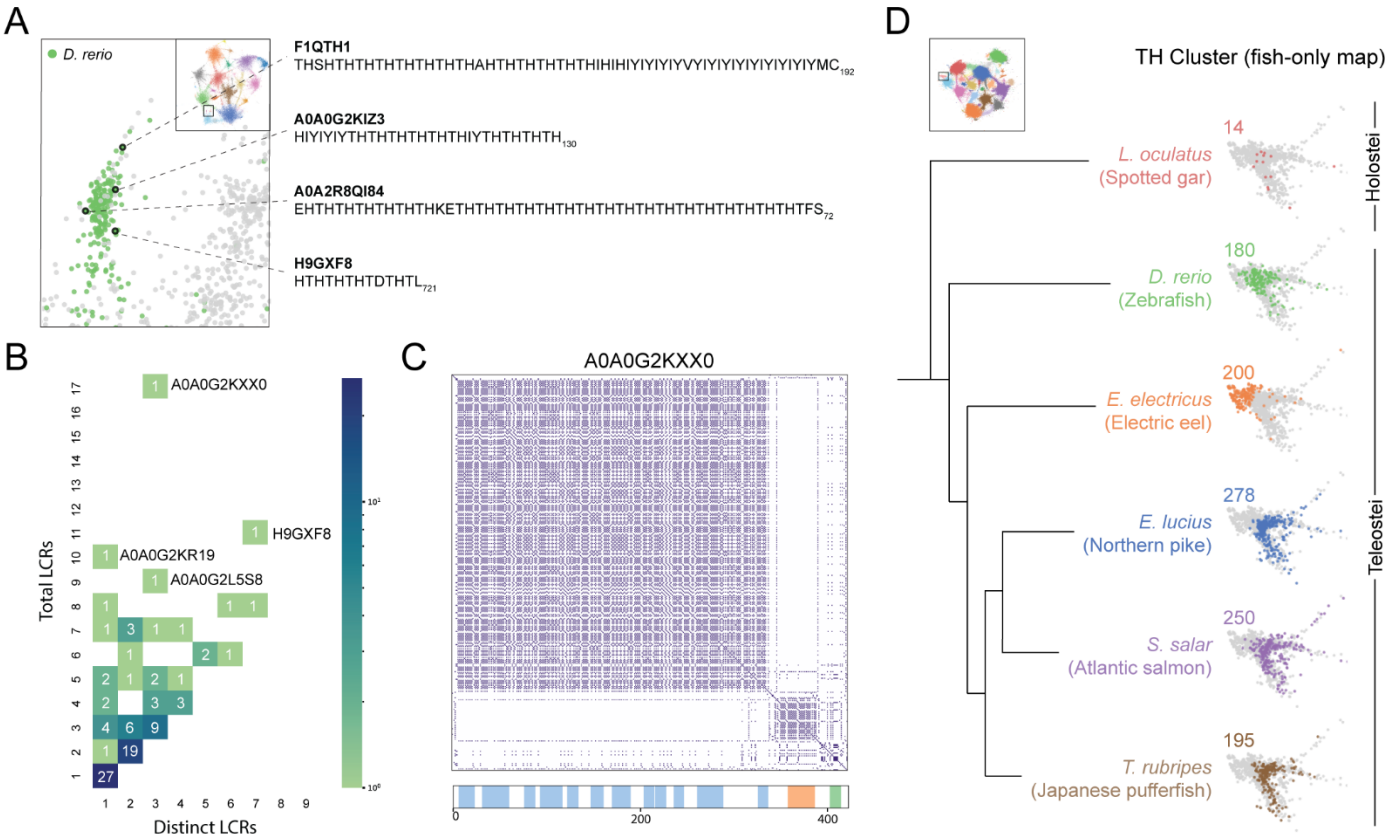


Figure 6: A conserved, teleost-specific T/H rich cluster exhibits signatures of higher order assemblies

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.

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

585 Given that we identified the T/H-rich LCR in zebrafish, we wanted to see if this
586 composition of LCRs was generally conserved in fishes. To test this, we used our dotplot and
587 UMAP approach to identify and cluster LCRs from a range of fishes from the clade
588 Actinopterygii (Hughes et al., 2018), to which zebrafish belongs. The six species we analyzed
589 were zebrafish, electric eel, northern pike, Atlantic salmon, Japanese pufferfish, and spotted gar
590 (*D. rerio*, *E. electricus*, *E. lucius*, *S. salar*, *T. rubripes*, and *L. oculatus*, respectively). Of these
591 fishes, all but the spotted gar substantially occupied the T/H-rich cluster. The fishes which
592 heavily occupied the T/H-rich cluster each contained ~200 T/H LCRs, while the spotted gar was

593 only lightly occupied with 14 LCRs (Figure 6D, Figure 6 - figure supplement 1A). Even when
594 accounting for the total number of LCRs of each species, the spotted gar had the lowest
595 percentage of LCRs in the T/H cluster (Figure 6 - figure supplement 1A, B). This difference in
596 occupancy of the T/H cluster correlated exactly with evolutionary relationships between these
597 fishes. Those containing T/H-rich LCRs belonged to Teleostei, while the spotted gar, which did
598 not, belonged to Holostei, a group which diverged from Teleostei in Actinopterygii. Moreover,
599 the seven species we analyzed outside of Actinopterygii did not have T/H-rich LCRs either
600 (Figure 6A, Figure 5 - figure supplement 3), which strongly suggests that T/H-rich LCRs are
601 indeed teleost specific. The finding that T/H-rich LCRs are conserved in teleosts, one of the
602 largest clades of fishes, suggests that these conditionally mixed-charge LCRs may play a
603 conserved, important role in these species.
604 Our approach was able to unearth conserved LCR compositions, with scaffold-like
605 distributions within their parent proteins. These results not only demonstrate the existence of
606 unexplored LCRs with signatures of higher order assemblies, but also highlight the ability of our
607 approach to systematically explore the vast diversity of proteins across species.
608
609 **DISCUSSION**
610 Here, we have established a systematic approach to study LCRs, providing a unified
611 view of how the sequences, features, relationships and functions of LCRs relate to each other.
612 This unified view enabled us to gain insight into the role of LCRs in multivalent interactions,
613 higher order assemblies, and organismal structures. Moreover, this framework for
614 understanding LCRs raises fundamental questions about how LCRs encode their functions.
615
616 **How can low complexity sequences capture the diversity of LCR function?**
617 While the functions of proteins are encoded in their sequence, it has been difficult to
618 assign functions to LCRs. Any mapping between LCR function and sequence space presents a

619 question of how the many disparate functions of LCRs can exist in a space which only employs
620 a few amino acids at a time.

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

633 Differences in composition also imply differences in sequence. It follows that functional
634 consequences downstream of sequence, such as post-translational modifications, can be
635 affected by differences in composition. We have shown that several bridge-like connections
636 exist between the clusters for serine and other amino acids in the LCR map. One well
637 understood kinase, CK2, binds and phosphorylates serines in acidic contexts (Rusin et al.,
638 2017), changing the physical properties of this sequence by making it more negatively charged.
639 Interestingly, bridge-like connections exist between both S and D, and S and E in the LCR map,
640 raising the possibility that their physical properties can be regulated to different extents by CK2.
641 Notably, assembly of NOLC1 and TCOF, which have many LCRs in the bridge between S and
642 E, is known to be regulated by CK2 phosphorylation (Werner et al., 2015, 2018). Similarly, a
643 bridge-like connection exists between E and K, two oppositely charged residues. Lysine
644 acetylation eliminates the positive charge in lysine, and has been shown *in vivo* to

645 predominantly occur in K/E/D-rich contexts (Lundby et al., 2012). Interestingly, K/E/D-rich
646 sequences are prevalent in nucleolar LCRs, and evidence points to lysine acetylation preventing
647 nucleolar integration (Fantini et al., 2010; Lirussi et al., 2012). Together, these observations
648 suggest that different nucleolar LCRs along the K/E bridge may be differentially regulated by
649 lysine acetylation, potentially affecting the structure of the nucleolus. Thus, differences in post-
650 translational modifications may provide an additional layer by which LCRs can encode biological
651 functions.

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

655

656 **Implications of bridges between certain amino acids in LCR space**

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

661 Some of these bridges represent mixtures of similar residue properties, such as
662 hydrophobic or negatively-charged amino acids. These findings are consistent with the
663 hypothesis that some sets of amino acids with similar physical properties may be redundant,
664 and thus varying combinations of them are not selected against. Interestingly, while R and K are
665 both positively-charged, basic residues, the region between these clusters was poorly
666 populated, suggesting that these residues may not always be interchangeable in LCRs. This is
667 consistent with known differences between R and K, such as the ability of R to participate in
668 stacking interactions. In fact, recent evidence showed that the physical properties of R and K
669 substantially differ, while the difference between D and E is much more subtle (Fossat et al.,
670 2021; Greig et al., 2020; Wang et al., 2018). Thus, while co-occurrence of similar amino acids

671 may not be entirely surprising, a lack of co-occurrence between seemingly similar amino acids
672 may point towards interesting differences between them.
673 Likewise, while dissimilar amino acids may not often co-occur in LCRs, the presence of
674 bridges with dissimilar amino acids may represent combinations which have emergent
675 functions. It has not escaped our notice that R/S and G/P are combinations of dissimilar amino
676 acids which all correspond to functional, conserved higher order assemblies--the speckle and
677 extracellular matrix. In these cases, it is known how the combinations of amino acids may
678 enable emergent properties, such as mixed charge domains (Greig et al., 2020) or tight-packing
679 polyproline helices (Ramachandran and Kartha, 1955; Rich and Crick, 1955; Cowan and
680 McGAVIN, 1955). However, certain combinations exist in which the properties are not well
681 understood, such as H/Q, N/S or T/H, among others. Thus, we hypothesize that the existence of
682 bridges between dissimilar amino acids may correspond to LCRs with specific emergent
683 properties. These types of LCRs represent open, unexplored regions of LCR space for which
684 the relationship between sequence and function has yet to be determined.
685
686 **A unified LCR map relates disparate higher order assemblies across species**
687 The ability of the LCR map to capture certain higher order assemblies raises questions
688 of what the LCRs in other parts of the map may tell us about their functions. While we do not
689 interpret that all LCRs must be involved in higher order assembly, the observation that LCRs of
690 certain higher order assemblies populated different regions of same sequence space allows us
691 to consider if there are similarities among these assemblies which give us insight into LCR
692 function. For example, the nucleolus is a liquid assembly of protein, RNA, and DNA essential for
693 ribosome biogenesis, while the extracellular matrix is a solid/gel-like assembly of glycoproteins
694 scaffolded by long collagen fibers. Despite the physical differences in each of these assemblies,
695 the LCRs are essential for their higher order assembly. The K-rich LCRs of nucleolar proteins
696 such as RPA43 are required for their higher order assembly and integration into the nucleolus,

697 while the G-P-P motif-containing LCRs in various collagens form key assemblies in the ECM
698 (Timpl et al., 1981; Mould and Hulmes, 1987; Hansen and Bruckner, 2003) and the G/V/P-rich
699 LCRs of elastin assemble to provide ECM elasticity (Urry et al., 1974; Rauscher et al., 2006).
700 Although these examples are vastly different in physical properties, a common theme is that the
701 LCRs enable the integration and assembly of various biomolecules in biological structures.
702 If this is the case, we may gain insight into the structures and organizations of species
703 by comparing the differences in the sequence space occupied by their LCRs. One fruitful
704 comparison was between the human extracellular matrix and plant cell wall, which greatly affect
705 cellular organization in their respective species. Each of these have taken a role in the
706 extracellular space, yet they have different chemical compositions, structures, and proteins. The
707 LCR spaces occupied by these proteins are unique for each extracellular assembly,
708 corresponding to differences in the specific interactions and processes required for their
709 formation. While human ECM and plant cell wall proteins both occupy spaces which have a
710 substantial presence of prolines, the specific differences in the regions they occupy give insight
711 into their unique properties. For example, LCRs of ECM proteins occupy the G/P-rich cluster
712 and the presence of glycines in ECM collagen proteins is crucial for tight packing of helices to
713 form the collagen triple helix (Beck et al., 2000), which is the basis for higher order assembly of
714 most of the tissues in the human body. On the other hand, while plant cell wall proteins also use
715 polyproline II helices, these P-rich LCRs occupy a different region in the map from LCRs in the
716 ECM. Moreover, plant cell wall proteins contain different P-rich LCR compositions which
717 delineate between extensins and other proline-rich cell wall proteins. Such differences, in
718 whether or not the contiguous prolines are interrupted, have been proposed to explain the
719 origins of plant cell wall proteins with different properties (Kieliszewski and Lamport, 1994;
720 Lamport et al., 2011), supporting the idea that functional divergence of LCRs can occur through
721 relatively local differences in sequence space. Our map of LCR sequence space not only shows
722 differences in the LCRs of ECM and cell wall proteins, but also shows that two different groups

723 of LCRs exist among cell wall proteins. Thus, this view of LCR sequence space captures key
724 sequence determinants of higher order assemblies, even when the differences between them
725 are subtle.

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

733

734 **ACKNOWLEDGMENTS**

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

742

743 **AUTHOR CONTRIBUTIONS**

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

747

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.

762

763 Cell Culture

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

769

770 Protein purification

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

783 Pellets were thawed and lysed on ice in 15 mL lysis buffer containing freshly added
784 lysozyme and benzonase prepared according to manufacturer instructions (Qiagen Qproteome
785 Bacterial Protein Prep Kit, Cat. No. 37900), 1mM PMSF (ThermoFisher Scientific, 36978), and
786 1.5 cOmplete mini EDTA-free protease inhibitor cocktail tablets (Millipore Sigma, 11836170001)
787 per 250 mL culture. Lysates were incubated on ice for 20 minutes with occasional inversion, and

788 sonicated for 5 cycles (30 secs on, 30 secs off, high intensity) on a Bioruptor 300 at 4-6°C.
789 Cellular debris and unlysed cells were pelleted by centrifugation for 30 minutes at 12,000 x g at
790 4°C.
791 Cleared lysates were syringe filtered (Pall Life Sciences, Product ID 4187) and added to
792 0.625 mL of glutathione-sepharose beads (GE Healthcare, GE17-0756), which were pre-
793 equilibrated in equilibration buffer (1X PBS, 250mM NaCl, 0.1% Tween-20) by performing four
794 10 mL washes for 5 minutes each with end-over-end rotation at 4°C. After addition of filtered
795 lysates, beads were incubated for 2 h at 4°C on an end-over-end rotator. Beads were
796 centrifuged at 500 x g for two minutes and unbound lysate was removed. Beads were washed
797 three times for 10 minutes with 10 mL cold wash buffer (150mM NaCl, 10mM $MgCl_2$, 10mM
798 Na_2HPO_4, 2mM ATP) at 4°C with end-over-end rotation. Three to five 0.5mL elutions were
799 performed at 4°C on a nutator, with freshly prepared elution buffer (100mM TRIS pH 8, 20mM
800 reduced glutathione, 5mM EDTA pH 8, 2mM ATP), each for 10 minutes. Elutions were
801 collected, concentrated, and subsequently buffer exchanged into protein storage buffer (25 mM
802 Tris pH 7.5, 150 mM KCl, 0.5 mM EDTA, 0.5 mM DTT freshly added, 10% glycerol) using
803 Amicon Ultra-0.5 centrifugal filter units with a 10kDa cutoff (Millipore Sigma, UFC5010). Protein
804 concentrations were determined, after which proteins were diluted to 100 μM in protein storage
805 buffer, aliquoted, and stored at -80°C.
806
807 Droplet formation assays
808 Droplet formation assays were performed in droplet formation buffer (50 mM Tris pH 7.0,
809 150 mM NaCl), in the presence of a final concentration of 10% PEG-8000 (New England
810 Biolabs, B1004), in a total volume of 12 μL. Droplet formation was initiated by the addition of 1
811 μL of purified protein (in protein storage buffer) to 11 μL of pre-mixed Droplet formation buffer
812 and PEG-8000 on ice (8.6 μL of Droplet formation buffer + 2.4 μL 50% PEG-8000). The final
813 protein concentration in the reaction was 8.3 μM. After the addition of purified protein, the

814 reaction was mixed by pipetting, 10 µL was loaded onto a microscope slide (Fisher Scientific,
815 12-544-2), and droplets were immediately imaged using a fluorescent microscope (Evos FL) at
816 40X magnification. Representative images were chosen for Figure 2.
817 Droplet formation assays were repeated over the course of about 6 months, with each
818 replicate corresponding to the same experiment carried out on different days, using the same
819 preparation of purified protein.
820
821 Immunofluorescence
822 Glass coverslips (Fisherbrand, 12-545-80) were placed in 24-well plates (Genesee
823 Scientific, 25-107) and coated in 3 µg/mL of fibronectin (EMD Millipore, FC010) for 30 minutes
824 at room temperature. HeLa cells were seeded in each well at 50,000 cells per well. 24 hours
825 after seeding, the cells were transfected with GFP-tagged protein plasmids using Lipofectamine
826 2000 (Invitrogen, 11668027). Each well was transfected using 100 ng of plasmid and 1 µL of
827 Lipofectamine 2000 in a total of 50 µL of OptiMEM (Gibco, 31985070) according to the
828 Lipofectamine 2000 instructions. Cells on glass coverslips were collected for
829 immunofluorescence 48 hours after transfection. Cells were collected by washing with 1x PBS
830 (Genesee Scientific, 25-508) and fixation in 4% paraformaldehyde (PFA) for 15 minutes at room
831 temperature, followed by another 3 washes with 1x PBS. Cells were permeabilized and blocked
832 by incubation in blocking buffer (1% BSA (w/v), 0.1% Triton X-100 (v/v), 1x PBS) for 1 hour at
833 room temperature. Coverslips were then incubated overnight at 4°C in a 1:100 dilution of
834 primary antibody (anti-MPP10, Novus Biologicals, NBP1-84341) in blocking buffer. After 3
835 washes with blocking buffer, coverslips were incubated for 2 hours in a 1:1000 dilution of
836 secondary antibody (anti-rabbit, Invitrogen, 32260). Coverslips were washed 3 times with
837 blocking buffer, then once with 1x PBS, and mounted on glass slides using ProLong Diamond
838 antifade mountant with DAPI (Invitrogen, P36962). Slides were sealed using clear nail polish,
839 allowed to dry, and stored at 4°C. Slides were imaged on a DeltaVision TIRF microscope using

840 100X oil immersion objective lens. The same set of exposure conditions (one exposure per
841 channel) was used across all slides. Raw images were deconvoluted, from which a max
842 projection image was generated. Deconvolution and max projection were performed using
843 Deltavision SoftWoRx software. Displayed images were scaled such that the distribution of
844 signal was representative. Image analysis was performed using Fiji
845 (https://imagej.net/software/fiji/). For each transfected construct, representative cells were
846 chosen. Cells that were excluded were cells that were not appreciably transfected, and cells
847 that highly overexpressed the transfected constructs.
848 The immunofluorescence experiment was performed more than three times over the
849 course of about 1 year, with each replicate corresponding to the same experiment carried out
850 on different days.
851
852 **External data**
853 Proteome Datasets
854 Proteomes were downloaded from UniProt for all species analyzed (see table below).
855 Every proteome was greater than 90% complete based on Benchmarking Universal Single-
856 Copy Ortholog (BUSCO) assessment score for proteome completeness. One protein sequence
857 was downloaded per gene in FASTA format. Thus, all protein names used in the manuscript are
858 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

859

860 Higher order assembly annotations

861 Annotations for higher order assemblies were downloaded from Uniprot, based on their
862 subcellular location annotations. Only entries which were Swiss-Prot reviewed (i.e. entry
863 belongs to the Swiss-Prot section of UniProtKB) were included in the annotations. Annotations
864 were accessed in FASTA format. Annotations for stress granule were taken from a published
865 experiment (Jain et al., 2016). Stress granule protein sequences from the “Tier1” list of stress
866 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

867

868 **Core approach**

869 See Figure 1D for overview and flowchart. All code was written in Python 3. Run on
870 Google Colaboratory or the Luria server at MIT. Python modules used were NumPy (1.20.1),
871 BioPython (1.78), Pandas (1.2.3), Mahotas (1.4.11), SciPy (1.6.2), Scanpy (1.7.2), AnnData
872 (0.7.5), NetworkX (2.3), Matplotlib (3.4.1), Seaborn (0.11.1). Code, dotplot module outputs, and
873 other relevant files can be found on zenodo (https://doi.org/10.5281/zenodo.5555373).

874

875 Dotplot generation (Module 1)

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

889

890 LCR calling (Module 2, part 1)

891 LCRs were called by identifying high density regions in protein dotplots through classic
892 image processing methods, such as kernel convolution, thresholding, and segmentation (Figure
893 1D).
894 To identify high density regions in dotplots, we performed kernel convolution on the
895 dotplots with a uniform 10x10 kernel, which calculates a convolved pixel intensity value from 0
896 to 100 based on the number of dots in that window. This kernel relates to the minimum length of
897 an LCR.
898 We used the convolved dotplots to determine this “high density” cutoff to define LCRs.
899 Specifically, we used a false discovery rate (FDR)-based approach to threshold the convolved
900 pixel intensities in a way that reliably identifies high density regions and treats the same
901 sequence similarly regardless of the proteome it comes from. For a given proteome, we
902 generated a background model by simulating an equally sized, length-matched ‘null proteome’,
903 whose sequences were generated from a uniform amino acid distribution. Using a uniform
904 amino acid distribution for the null proteome minimizes proteome-specific effects on whether a
905 sequence is considered to contribute to a region of high density in a dotplot. Moreover,
906 matching the lengths of the proteomes accounts for differences in the length distributions of
907 proteins in different proteomes. We compared the distribution of convolved pixel intensities from
908 the across all convolved dotplots of proteins in the real proteome with those from the null
909 proteome and identified the lowest convolved pixel intensity which satisfied a stringent FDR of
910 0.002 (Figure 1D, Figure 5 - figure supplement 1). FDR was defined by the number of pixels
911 from the null set which pass the threshold divided by the total number of pixels which pass the
912 threshold (from the real and null sets combined). This threshold was then applied to every
913 protein in the proteome to generate segmented dotplots, in which high-density regions (referred
914 to as segmented regions) had values of 1 while other regions had values of 0. The positions
915 from -4 and +5 of the boundaries of the segmented regions were included as the start and stop
916 of the LCR to account for the convolution kernel size. The exception to this was LCRs which

917 existed within that distance from the start or stop of a protein, in which the protein start or stop
918 was designated the start or stop accordingly. Only segmented regions which intersected with
919 the diagonal were called as LCRs.
920
921 LCR type and copy number determination (Module 2, part 2)
922 To computationally determine the types of LCRs and the copy number for each type, we
923 determined the presence of segmented regions at the intersection between called LCRs in the
924 segmented dotplot (Figure 1D, E). For each protein, we represented the LCRs as a network in
925 which the LCRs were nodes and intersections between LCRs were edges (Figure 1D, E, Figure
926 1 - figure supplement 4). The total number of nodes equals the total number of LCRs in the
927 protein. The number of connected components of this network equals the number of distinct
928 LCR types in the protein. Therefore, the number of nodes within a given connected component
929 equals the number of LCRs of that type. NetworkX (version 2.3) was used to calculate these
930 values, and plot the network representation of LCR relationships within proteins.
931
932 Entropy calculation, random length-matched sequence sampling
933 Shannon entropy was calculated for each LCR sequence and a length-matched
934 sequence which was randomly sampled from the respective proteome. Random length-matched
935 sequence sampling was done by indexing the position of all proteins in the proteome from 1 to
936 the length of the proteome (i.e. the sum of lengths of all proteins), and randomly selecting a
937 position between 1 and the length of the proteome minus the length of the sequence of interest.
938 The randomly sampled sequence was the sequence of the matched length, starting at the
939 selected position. Shannon entropy for both the LCR and randomly sampled sequence was
940 calculated using Scipy's implementation.
941
942 Other LCR calling methods (SEG/fLPS)

943 LCRs were called with other methods, SEG (Wootton and Federhen, 1993) and fLPS
944 (Harrison, 2017) for comparison.
945 SEG was run on the human proteome using 'default', 'intermediate', and 'strict' settings,
946 as defined by the PLAtform of TOols for LOw COmplexity (PlaToLoCo) (Jarnot et al., 2020).
947 Settings used from PlaToLoCo (http://platoloco.aei.polsl.pl/#!/help, accessed May 20, 2021) are
948 restated here for completeness. 'Default': *W* = 12, *K1* = 2.2, *K2* = 2.5; 'Intermediate': *W* = 15, *K1*
949 = 1.9, *K2* = 2.5 (Huntley and Golding, 2002); 'Strict': *W* = 15, *K1* = 1.5, *K2* = 1.8 (Radó-Trilla and
950 Albà, 2012). From the output, we extracted the LCR coordinates for use in downstream entropy
951 calculations. SEG was downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/seg/seg/ on May 20,
952 2021.
953 fLPS was run on the human proteome using 'default' and 'strict' settings, as defined by
954 the PLAtform of TOols for LOw COmplexity (PlaToLoCo) (Jarnot et al., 2020), with a uniform
955 background amino acid composition. Settings used from PlaToLoCo
956 (http://platoloco.aei.polsl.pl/#!/help, accessed May 20, 2021) are restated here for
957 completeness. 'default: m = 15, M = 500, t = 0.001, c = equal; 'strict': m = 5, M = 25, t =
958 0.00001, c = equal. From the output of fLPS, 'whole' rows were dropped in order to remove LCR
959 calls covering the full length of a protein, which obscured LCR calls of subsequences of
960 proteins. We then extracted the LCR coordinates for use in downstream entropy calculations.
961 fLPS was downloaded from: https://github.com/pmharrison/flps/blob/master/fLPS.tar.gz on May
962 20, 2021.
963
964 Generation of LCR maps (UMAP dimensionality reduction, Leiden clustering)
965 LCR maps contained a 2 dimensional representation of different LCR amino acid
966 compositions. For each LCR in the proteome, the amino acid composition was calculated as the
967 frequency of each amino acid in the LCR, and was represented as a vector in 20-dimensional
968 space. The 20-dimensional vectors of all LCRs were saved in AnnData format as an array in

969 which rows were LCRs and columns were the amino acid frequencies. LCR maps were
970 generated by dimensionality reduction from 20 to 2 dimensions using Scanpy's implementation
971 of UMAP (random_state=73, n_components=2; n_neighbors=200 for Figure 5A and
972 n_neighbours=default for Figures 3A and 6D; (McInnes et al., 2020; Wolf et al., 2018)). Amino
973 acid distributions on the LCR map were generated by coloring each point on a color scale
974 corresponding to the frequency of the amino acid represented. Leiden clustering was performed
975 using Scanpy (random_state=73), and the most represented amino acids in each cluster was
976 determined by looking at the amino acid distributions in each cluster by eye.
977
978 Annotation of LCR maps (higher order assemblies, biophysical predictions)
979 Annotation of LCRs belonging to higher order assemblies (see table above) was done by
980 adding annotations to the AnnData object and coloring the LCRs using Scanpy's plotting
981 implementation. Wilcoxon rank sum (MannWhitneyU) tests for amino acid enrichment in LCRs
982 of higher order assemblies were performed using Scanpy. For the Wilcoxon rank sum tests
983 comparing one annotation against all other LCRs, default settings were used. For the Wilcoxon
984 rank sum tests comparing between two annotation sets of LCRs, one annotation set was set as
985 the reference.
986 Biophysical predictions were calculated and mapped for all LCRs. For IUPred2A and
987 ANCHOR2 predictions, which are context dependent, the scores at each position were
988 calculated for full-length proteins in the proteome using a modified version of the official python
989 script ((Mészáros et al., 2018); https://iupred2a.elte.hu/download_new, accessed May 31, 2021)
990 to allow for batch predictions. LCR positions identified by our dotplot approach were used to
991 extract the corresponding ANCHOR and IUPred2A scores for each position in each LCR. The
992 mean ANCHOR and IUPred2A scores for each LCR were calculated and used to color the
993 UMAP plot. The IUPred2A and ANCHOR2 scoring was run with the default 'long' setting and '-a'
994 to include ANCHOR predictions. Kappa scores (Das and Pappu, 2013; Holehouse et al., 2017)

995 for mixed-charge distribution were calculated for each LCR using the localCIDER package

996 (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.

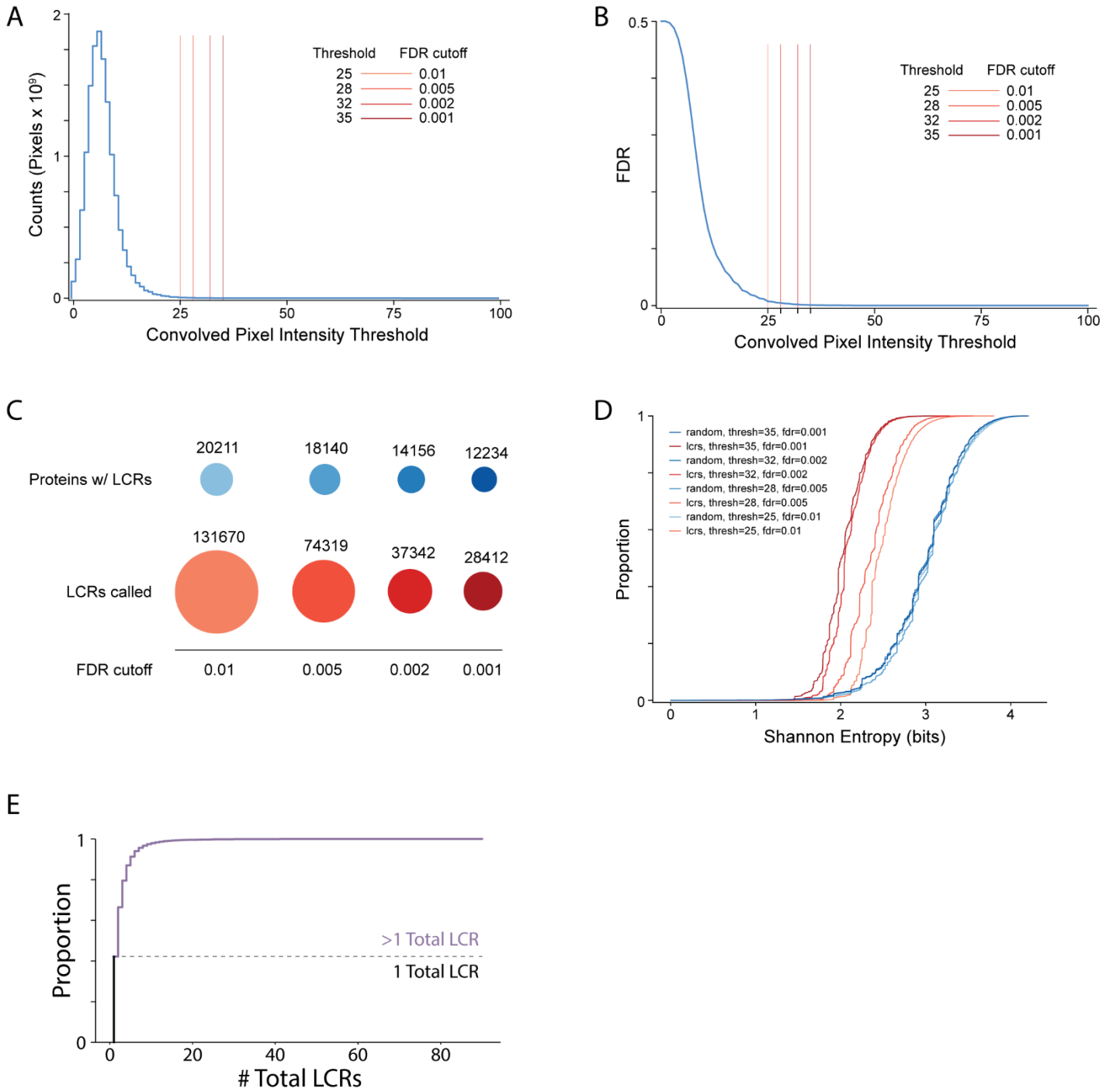


Figure 1 - figure supplement 2: Summary statistics from systematic dotplot analysis of human proteome

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.

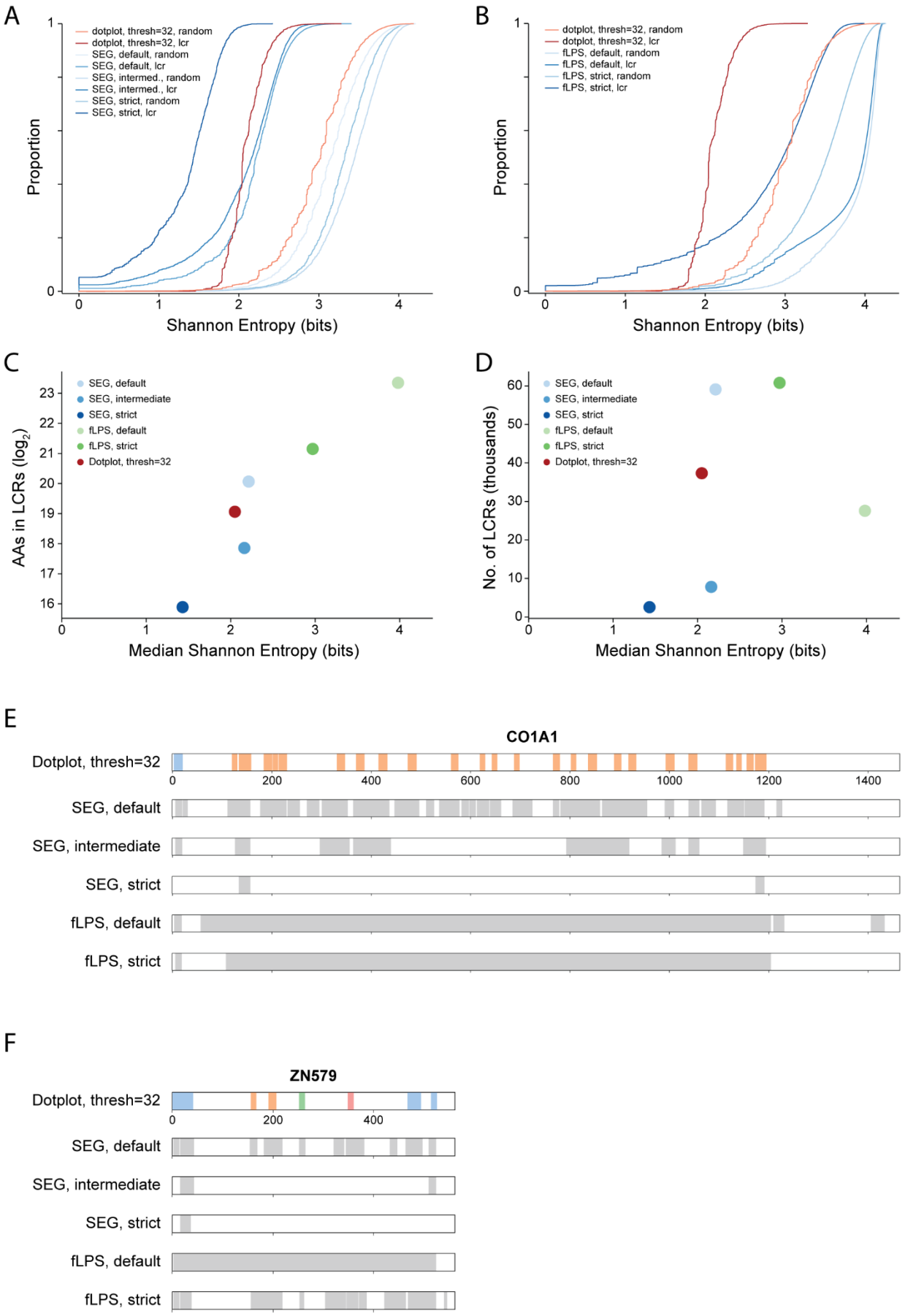

A
dotplot, thresh=32, random
dotplot, thresh=32, lcr
SEG, default, random
SEG, default, lcr
SEG, intermed., random
SEG, intermed., lcr
SEG, strict, random
SEG, strict, lcr
Proportion
Shannon Entropy (bits)
B
dotplot, thresh=32, random
dotplot, thresh=32, lcr
fLPS, default, random
fLPS, default, lcr
fLPS, strict, random
fLPS, strict, lcr
Proportion
Shannon Entropy (bits)
C
SEG, default
SEG, intermediate
SEG, strict
fLPS, default
fLPS, strict
Dotplot, thresh=32
AAs in LCRs (log2)
Median Shannon Entropy (bits)
D
SEG, default
SEG, intermediate
SEG, strict
fLPS, default
fLPS, strict
Dotplot, thresh=32
No. of LCRs (thousands)
Median Shannon Entropy (bits)
E
CO1A1
Dotplot, thresh=32
SEG, default
SEG, intermediate
SEG, strict
fLPS, default
fLPS, strict
F
ZN579
Dotplot, thresh=32
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.

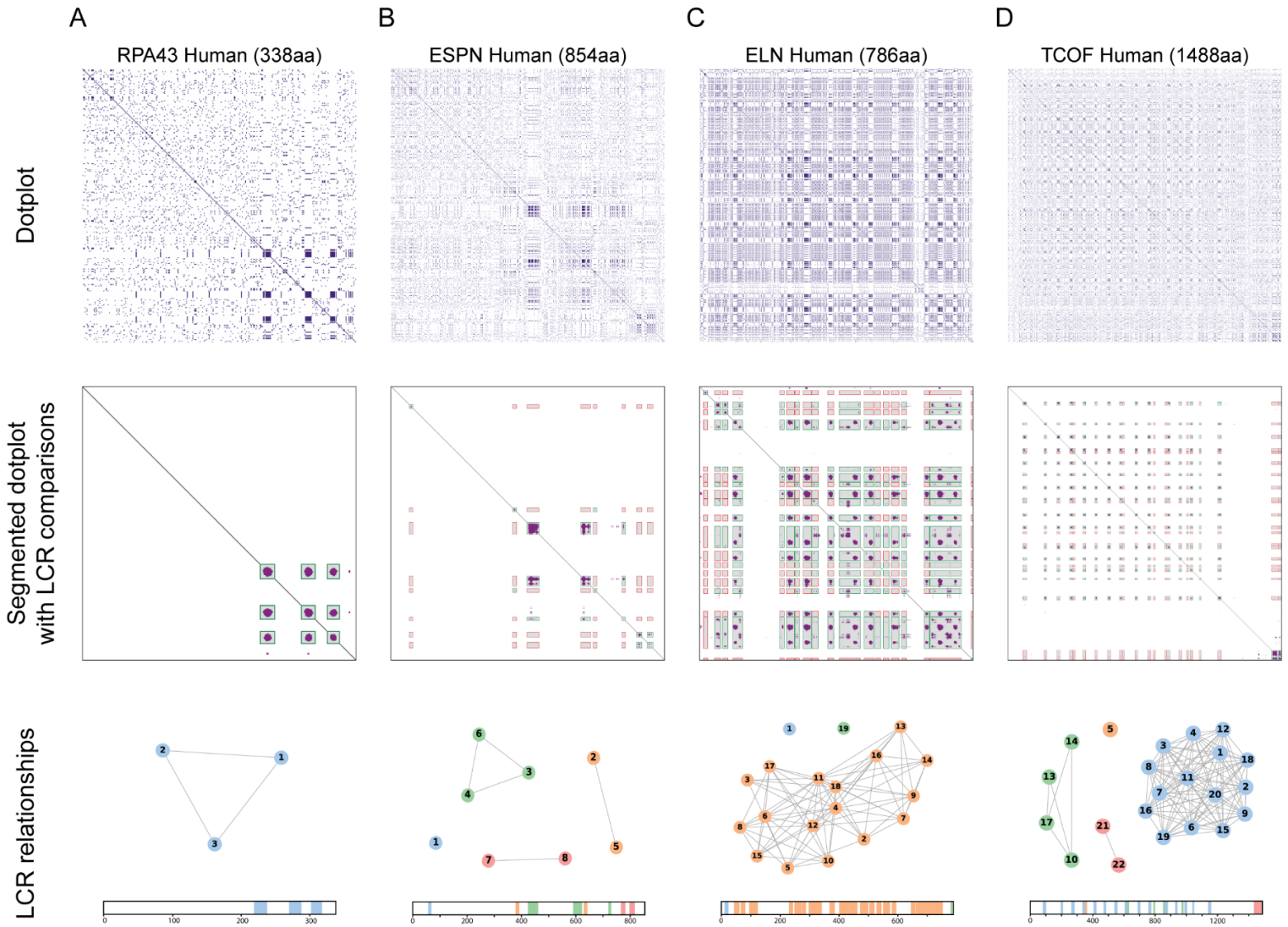
B) Same as A) but using fLPS (default or strict, see Methods for details).

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.



LCR Sequences

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- VGTPAAAAAAAAAAAKAAKYGAAAGL
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- NPKSKKEKKKSDKRKKDKEKKEK-
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.

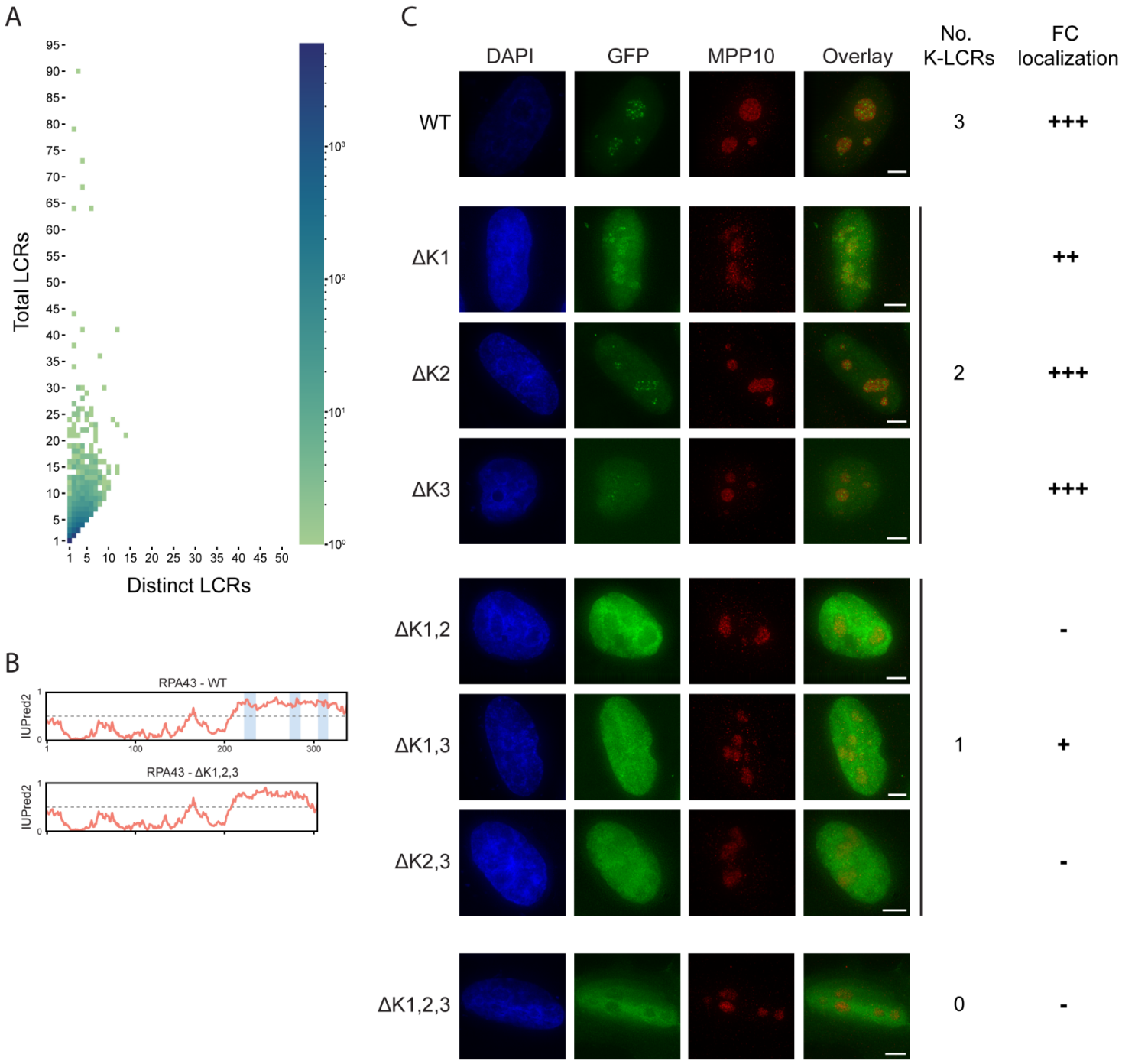


Figure 2 - figure supplement 1: Supplementary information for LCR type and copy number

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).

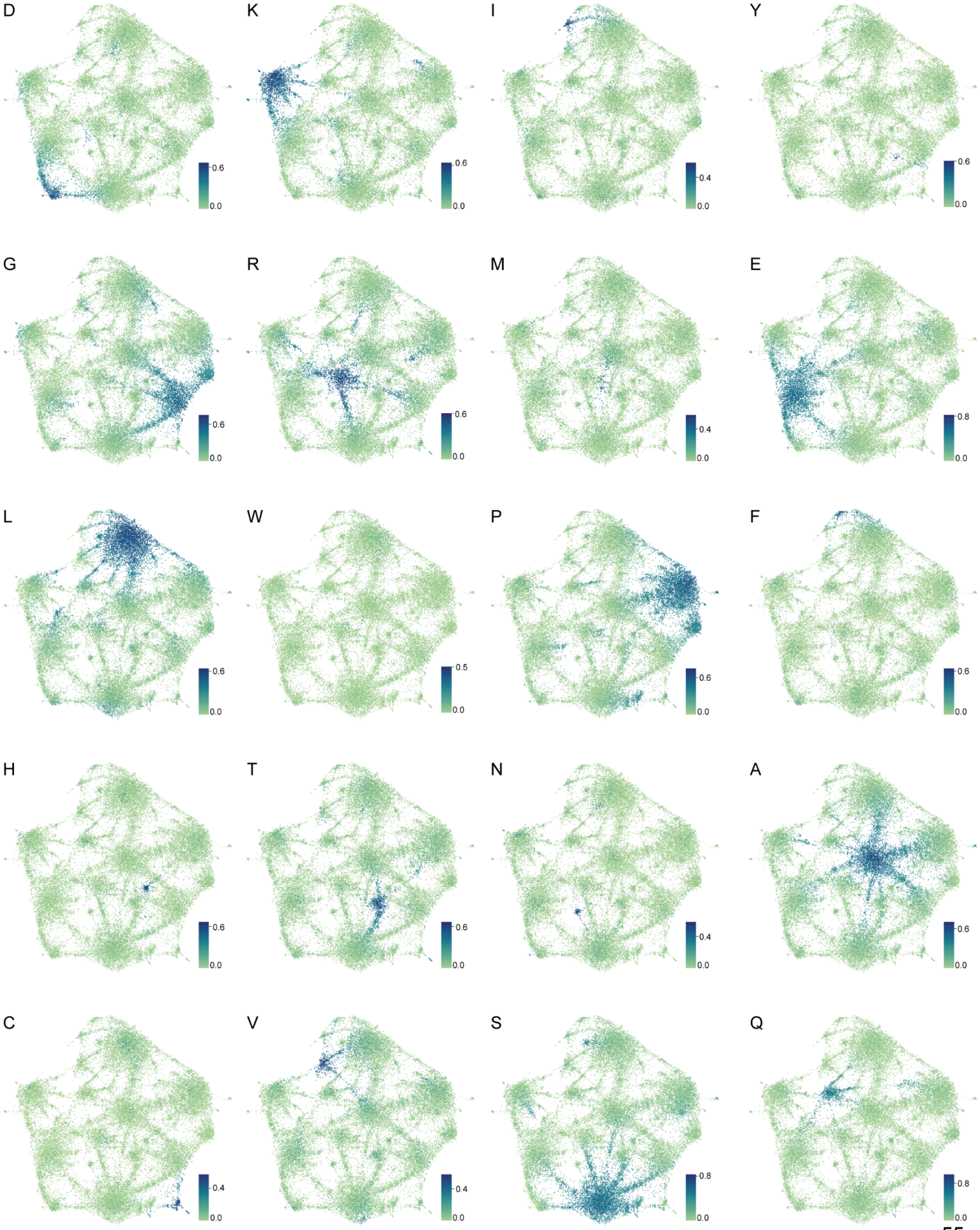

D
0.6
0.0
K
0.6
0.0
I
0.4
0.0
Y
0.6
0.0
G
0.6
0.0
R
0.6
0.0
M
0.4
0.0
E
0.8
0.0
L
0.6
0.0
W
0.5
0.0
P
0.6
0.0
F
0.6
0.0
H
0.6
0.0
T
0.6
0.0
N
0.4
0.0
A
0.6
0.0
C
0.4
0.0
V
0.4
0.0
S
0.8
0.0
Q
0.8
0.0

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.

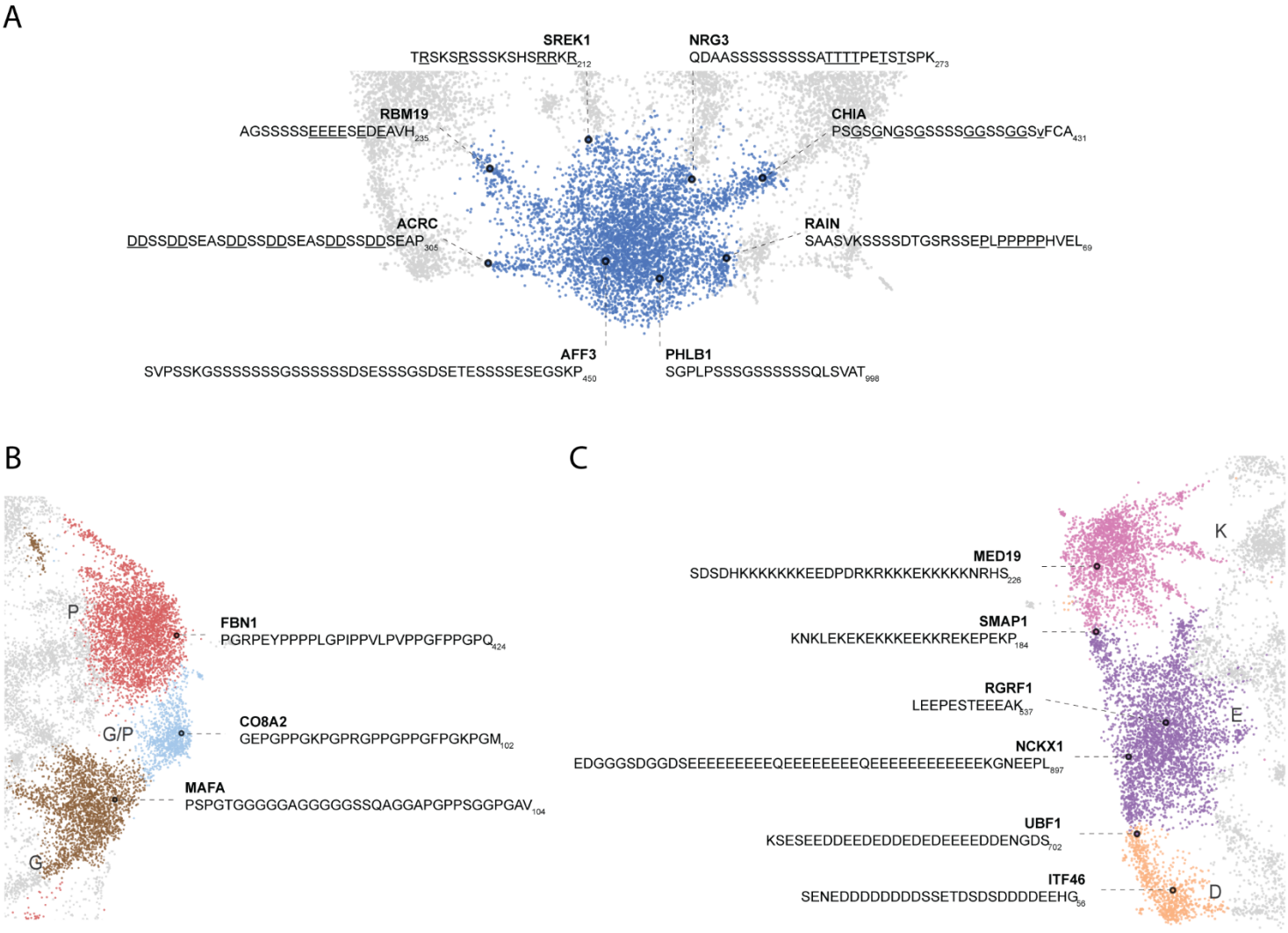


Figure 3 - figure supplement 2: Nuanced sequence differences among LCRs correspond to their positions in the UMAP
Close up view of specific clusters in human proteome UMAP (shown in Figure 3A), with several LCR sequences and their parent proteins annotated. For all LCRs shown the subscript at the end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

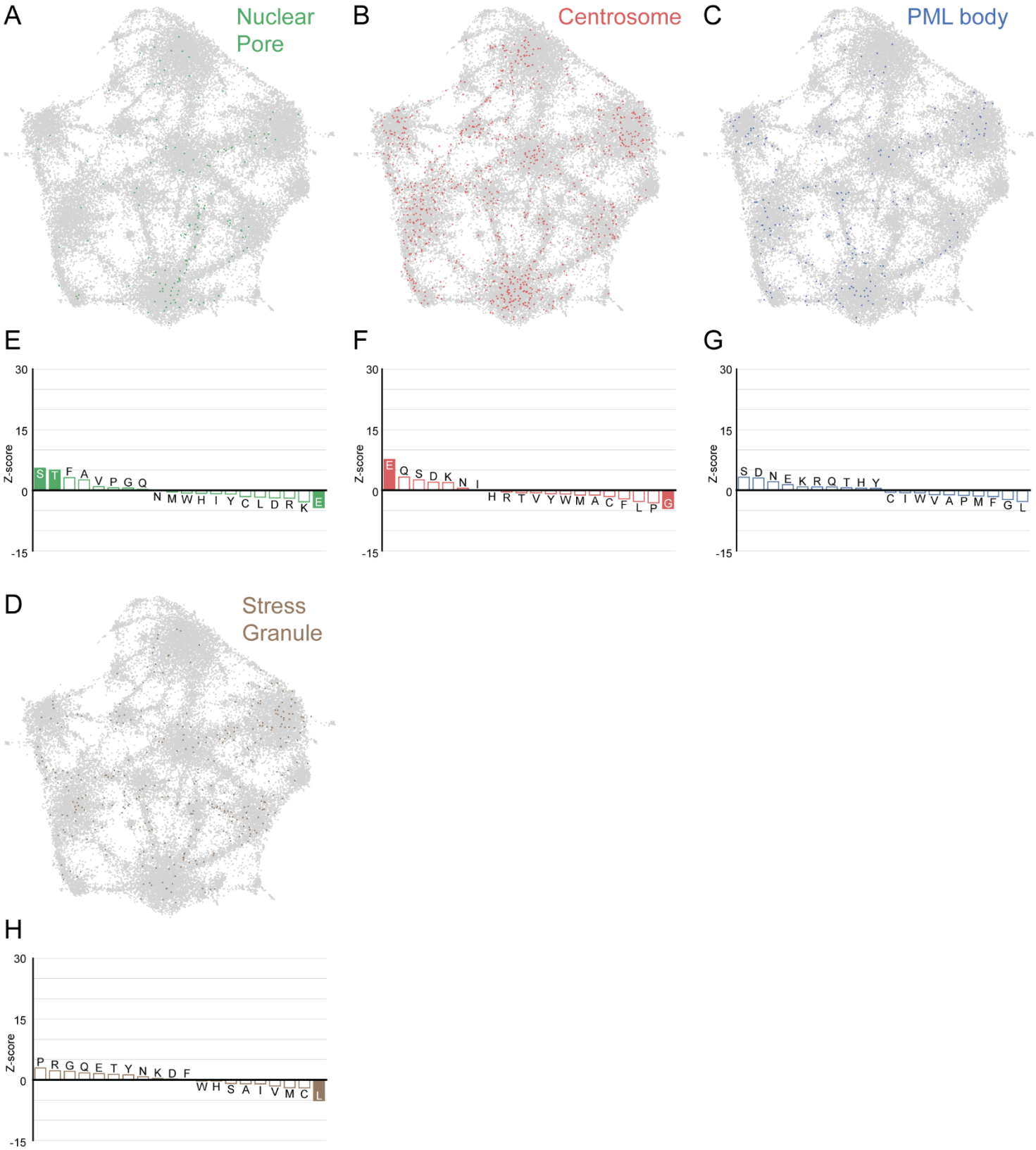


Figure 3 - figure supplement 3: LCRs of known higher order assemblies annotated on onto human proteome UMAP from Figure 3A

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.

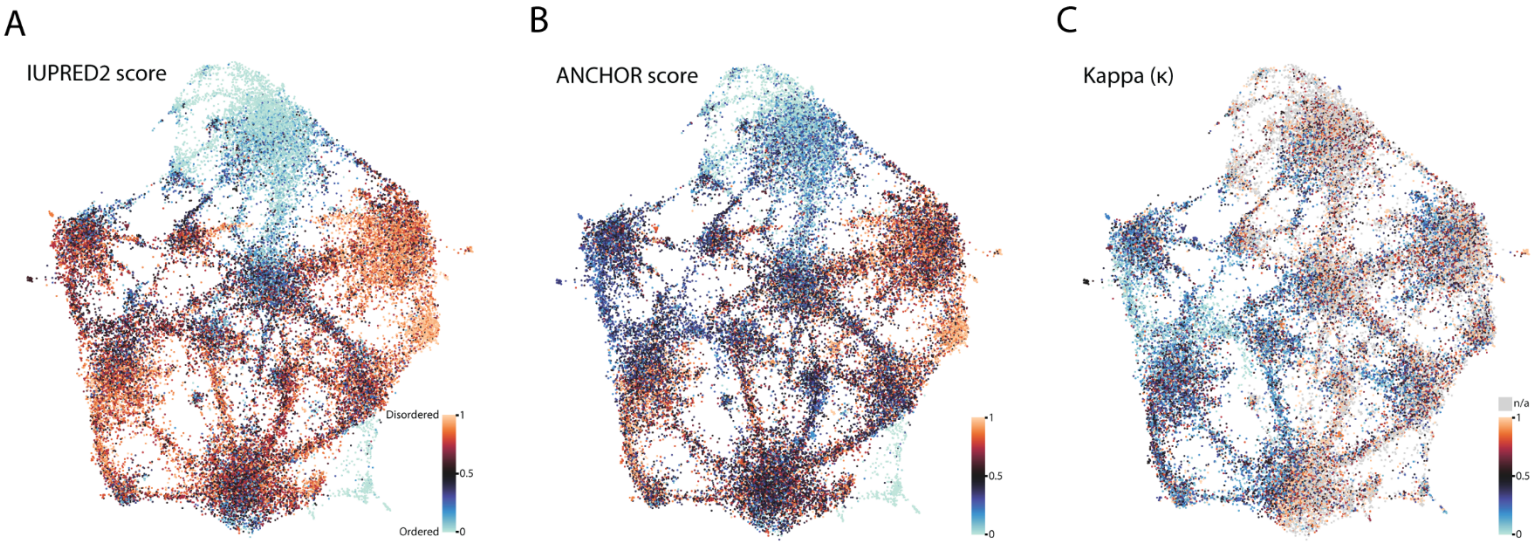


Figure 3 - figure supplement 4: Biophysical predictions of LCRs mapped onto human proteome UMAP from Figure 3A.

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.

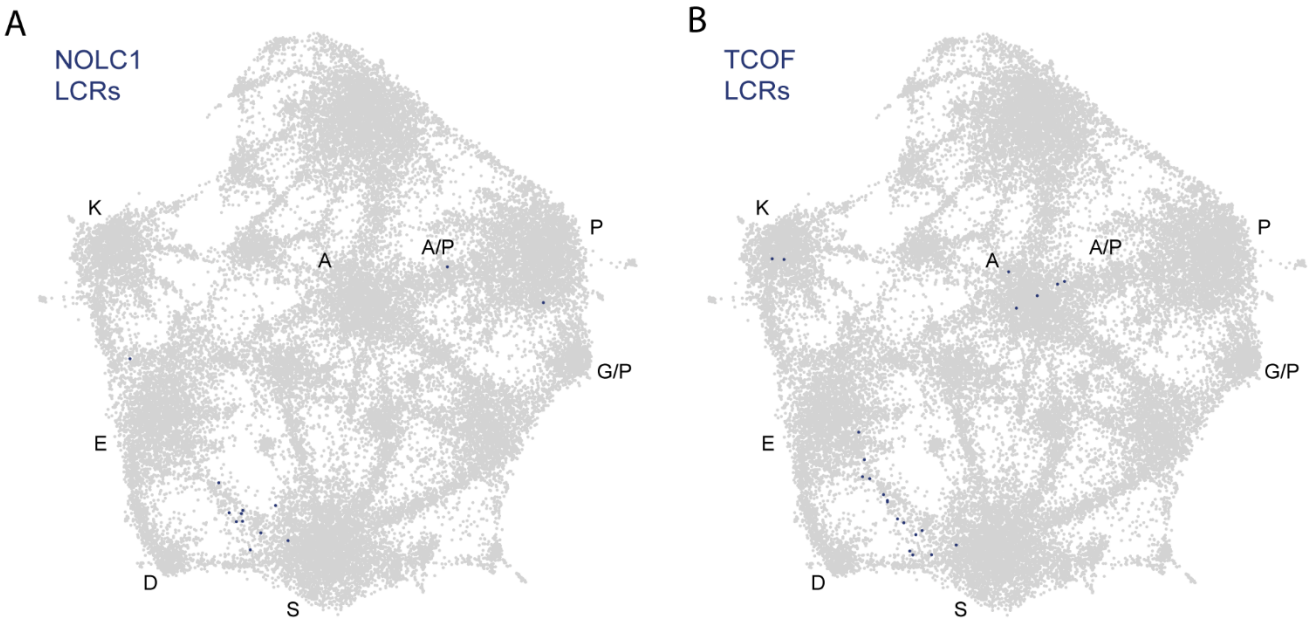


Figure 4 - figure supplement 1: LCRs of NOLC1 and TCOF mapped onto human proteome UMAP from Figure 3A.

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.

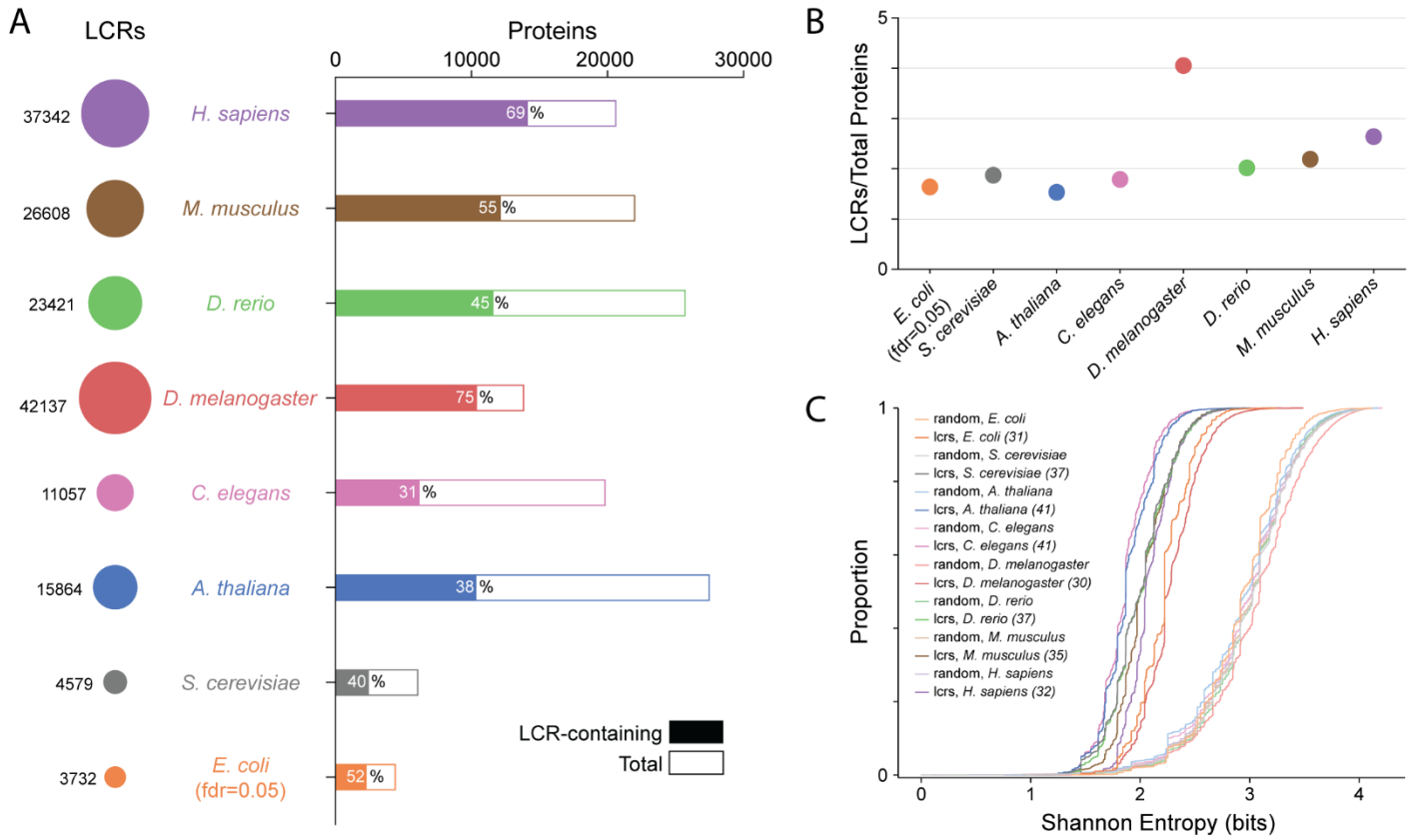


Figure 5 - figure supplement 1: Summary statistics from systematic dotplot analysis across species

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.

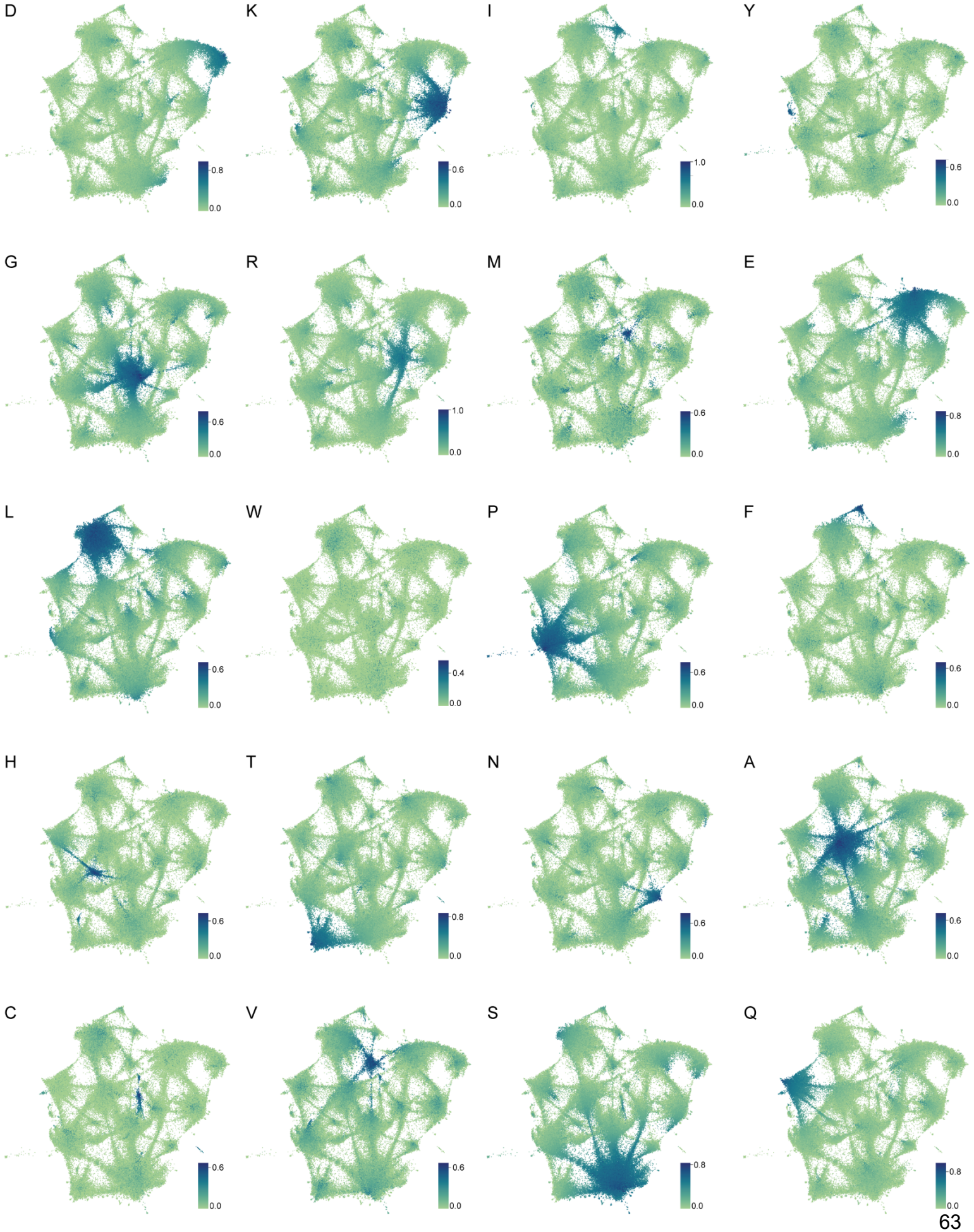

D
0.8
0.0
K
0.6
0.0
I
1.0
0.0
Y
0.6
0.0
G
0.6
0.0
R
1.0
0.0
M
0.6
0.0
E
0.8
0.0
L
0.6
0.0
W
0.4
0.0
P
0.6
0.0
F
0.6
0.0
H
0.6
0.0
T
0.8
0.0
N
0.6
0.0
A
0.6
0.0
C
0.6
0.0
V
0.6
0.0
S
0.8
0.0
Q
0.8
0.0

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.

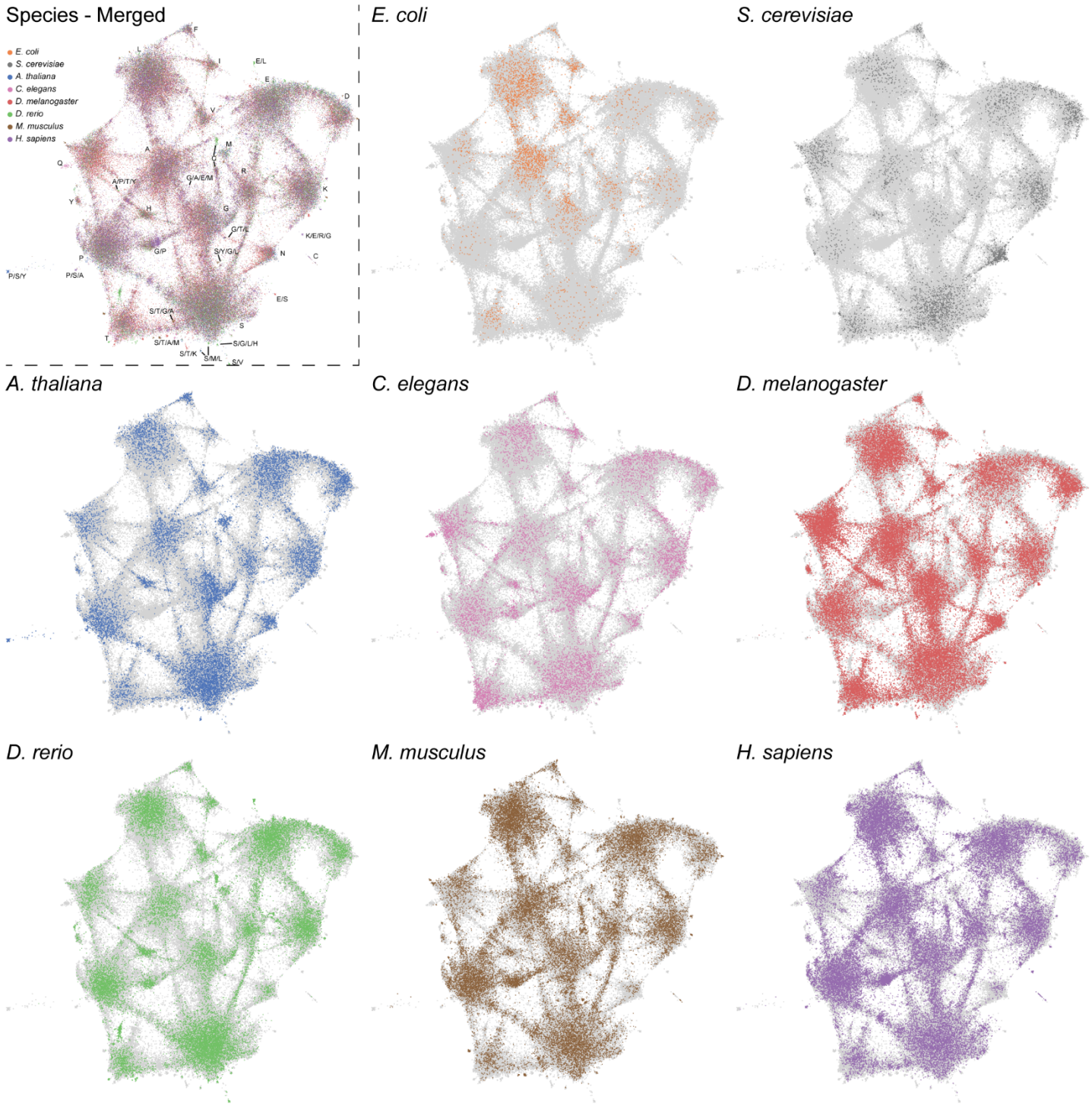


Figure 5 - figure supplement 3: LCRs of individual species mapped onto expanded UMAP from Figure 5A
UMAPs of LCRs in proteomes of human (*H. Sapiens)*, mouse (*M. musculus*), zebrafish (*D. rerio*)*, fruit fly* (*D. melanogaster*)*,* worm (*C. elegans*)*,* Baker's yeast (*S. cerevisiae*)*, A. thaliana, 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.

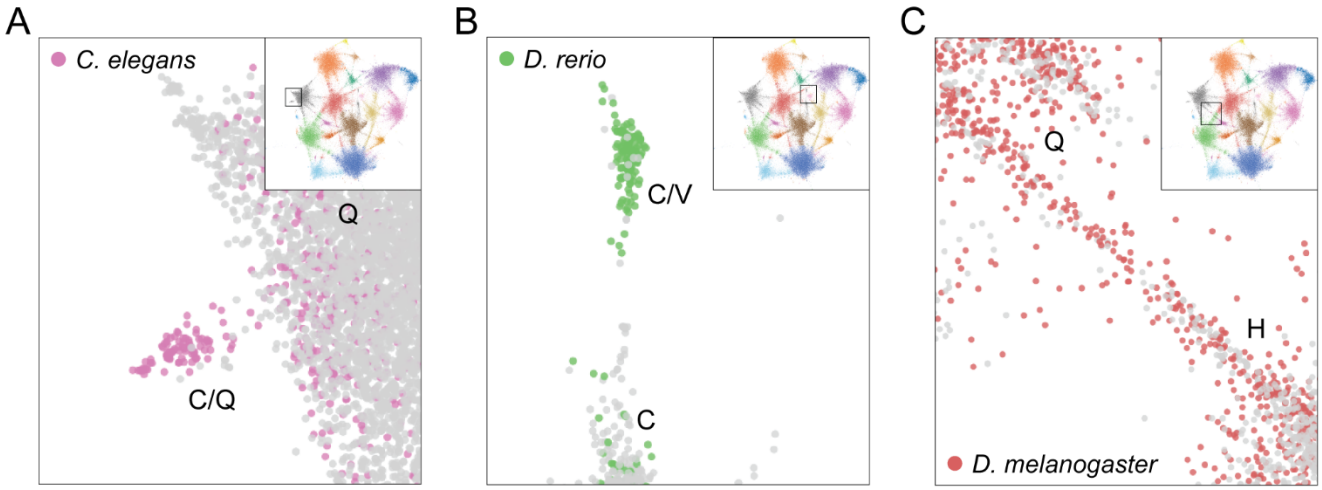


Figure 5 - figure supplement 4: Examples of species-specific clusters in the expanded UMAP from Figure 5A

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.

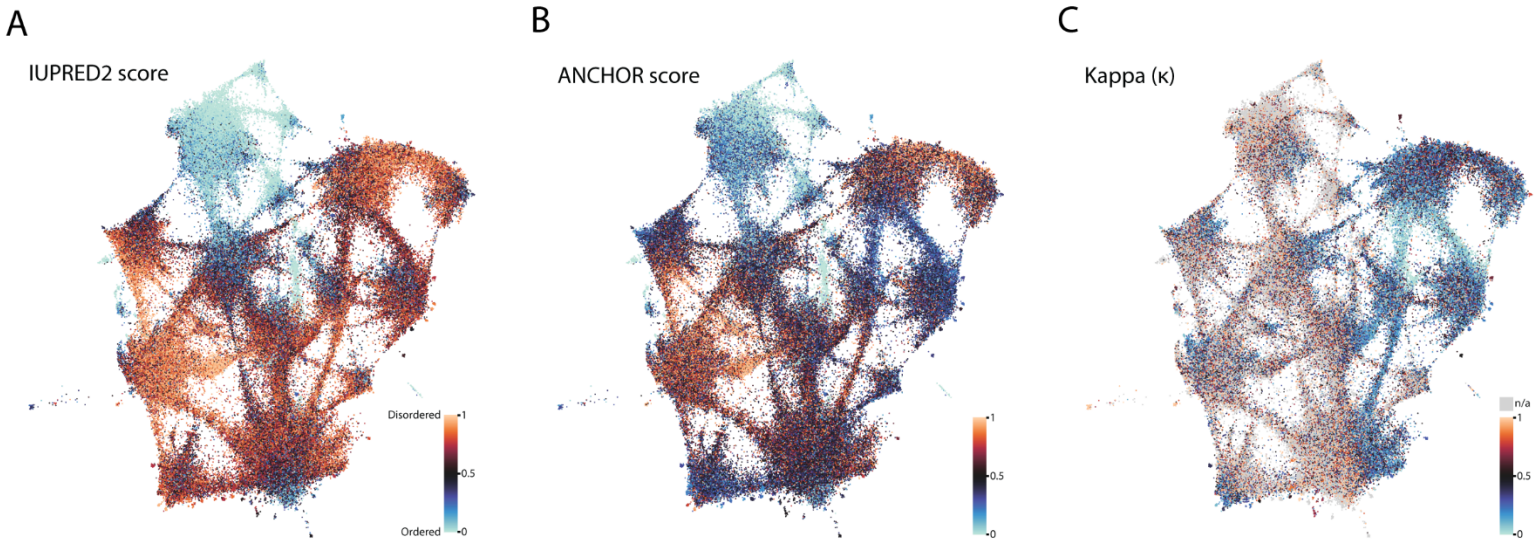


Figure 5 - figure supplement 5: Biophysical predictions of LCRs mapped onto the expanded UMAP from Figure 5A.
Mapping biophysical predictions of LCRs onto UMAP of LCRs from proteomes of human (*H. Sapiens)*, mouse (*M. musculus*), zebrafish (*D. rerio*)*, fruit fly* (*D. melanogaster*)*,* worm (*C. elegans*)*,* 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.

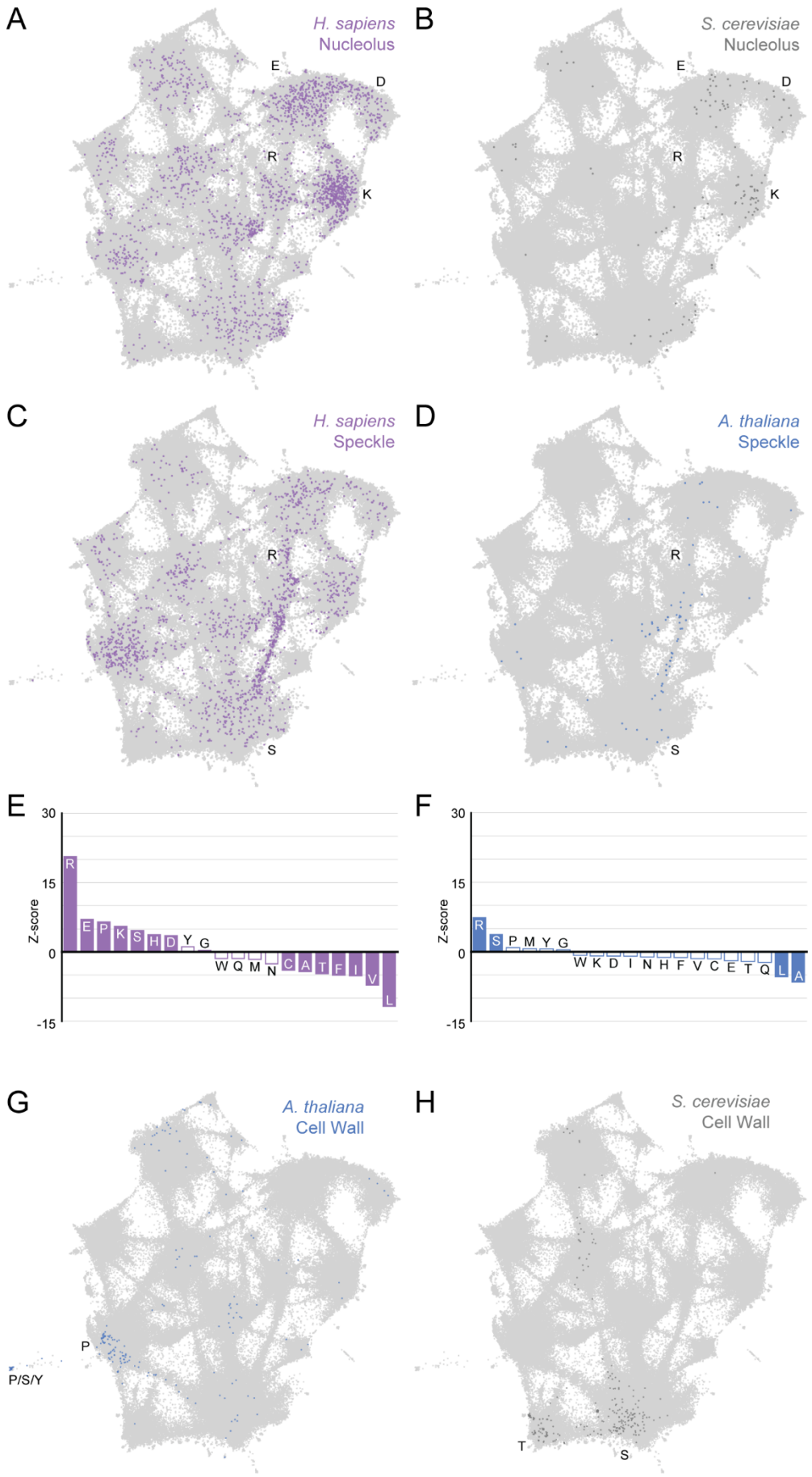

A
H. sapiens
Nucleolus
E
D
R
K
B
S. cerevisiae
Nucleolus
E
D
R
K
C
H. sapiens
Speckle
R
S
D
A. thaliana
Speckle
R
S
E
30
15
0
-15
Z-score
R E P K S H D Y G W Q M N C A T F I V L
F
30
15
0
-15
Z-score
R S P M Y G W K D I N H F V C E T Q L A
G
A. thaliana
Cell Wall
P
P/S/Y
H
S. cerevisiae
Cell Wall
T
S

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.

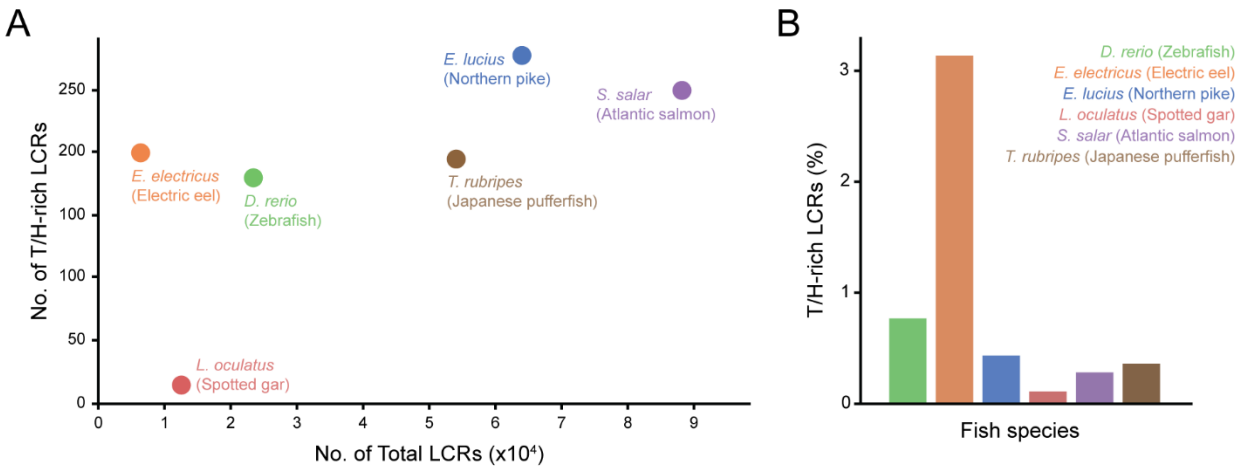


Figure 6 - figure supplement 1: Number and proportion of T/H-rich LCRs across fish species

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.