PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Proteome-scale relationships between local amino acid composition and protein fates and functions

Sean M. Cascarina*, Eric D. Ross*

Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, CO, United States of America

* Sean.Cascarina@colostate.edu (SMC); Eric.Ross@colostate.edu (EDR)

OPEN ACCESS

Citation: Cascarina SM, Ross ED (2018) Proteome-scale relationships between local amino acid composition and protein fates and functions. PLoS Comput Biol 14(9): e1006256. https://doi.org/10.1371/journal.pcbi.1006256

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: May 30, 2018

Accepted: August 16, 2018

Published: September 24, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the National Science Foundation (MCB-1517231; https://www.nsf.gov/div/index.jsp?div=MCB) to EDR and by the National Institute of General Medical Sciences (GM105991; https://www.nigms.nih.gov/) to EDR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Proteins with low-complexity domains continue to emerge as key players in both normal and pathological cellular processes. Although low-complexity domains are often grouped into a single class, individual low-complexity domains can differ substantially with respect to amino acid composition. These differences may strongly influence the physical properties, cellular regulation, and molecular functions of low-complexity domains. Therefore, we developed a bioinformatic approach to explore relationships between amino acid composition, protein metabolism, and protein function. We find that local compositional enrichment within protein sequences is associated with differences in translation efficiency, abundance, half-life, protein-protein interaction promiscuity, subcellular localization, and molecular functions of proteins on a proteome-wide scale. However, local enrichment of related amino acids is sometimes associated with opposite effects on protein regulation and function, highlighting the importance of distinguishing between different types of low-complexity domains. Furthermore, many of these effects are discernible at amino acid compositions below those required for classification as low-complexity or statistically-biased by traditional methods and in the absence of homopolymeric amino acid repeats, indicating that thresholds employed by classical methods may not reflect biologically relevant criteria. Application of our analyses to composition-driven processes, such as the formation of membraneless organelles, reveals distinct composition profiles even for closely related organelles. Collectively, these results provide a unique perspective and detailed insights into relationships between amino acid composition, protein metabolism, and protein functions.

Author summary

Low-complexity domains in protein sequences are regions that are composed of only a few amino acids in the protein "alphabet". These domains often have unique chemical properties and play important biological roles in both normal and disease-related processes. While a number of approaches have been developed to define low-complexity domains, these methods each possess conceptual limitations. Therefore, we developed a

Competing interests: The authors have declared that no competing interests exist.

complementary approach that focuses on local amino acid composition (i.e. the amino acid composition within small regions of proteins). We find that high local composition of individual amino acids is associated with pervasive effects on protein metabolism, subcellular localization, and molecular function on a proteome-wide scale. Importantly, the nature of the effects depend on the type of amino acid enriched within the examined domains, and are observable in the absence of classically-defined low-complexity (and related) domains. Furthermore, we define the compositions of proteins involved in the formation of membraneless, protein-rich organelles such as stress granules and P-bodies. Our results provide a coherent view and unprecedented resolution of the effects of local amino acid enrichment on protein biology.

Introduction

Low-complexity domains (LCDs) in proteins are regions enriched in only a subset of possible amino acids. LCDs can be composed of homopolymeric repeats of a single amino acid, short tandem repeats consisting of only a few different amino acids, or aperiodic stretches with little amino acid diversity [1]. Proteins containing LCDs are relatively common among organisms from all domains of life, and are particularly common among eukaryotes [2–4]. For example, approximately 70% of genes in the *Saccharomyces cerevisiae* genome possess at least one classically-defined LCD [3]. Furthermore, the total number of LCDs far exceeds the total number of yeast genes (~2-fold more LCDs than genes), indicating that many genes contain multiple distinct LCDs.

Various methods have been developed to assess biopolymer sequence complexity [1,5–9]. One of the most commonly employed methods to define LCDs is the SEG algorithm [1], which scans protein (or nucleic acid) sequences using a short sliding window, and calculates the local Shannon entropy for each window (see [10] for a detailed description). Subsequences with a Shannon entropy value below a pre-determined "trigger" threshold are classified as LCDs. LCD boundaries are later extended and refined by merging overlapping LCDs and calculating combinatorial sequence probabilities. Another metric commonly used to assess relative sequence complexity is compositional bias, which involves determining the statistical probability of a sequence given whole-proteome frequencies of the individual amino acids [11,12]. These approaches (or closely-related approaches) have been used extensively to examine LCDs on a proteome-wide scale [1,3,12–17].

LCD-containing proteins have been implicated in a variety of normal and pathological cellular processes. For example, Q/N-rich yeast proteins often play a role in transcription regulation, endocytosis, and cell cycle regulation, among other functions [11,18]. Many proteins containing Q/N-rich LCDs, or LCDs of related types (Q/N/G/S/Y-rich LCDs) have been linked to prion or prion-related processes [11,18–21]. Additionally, many prion-like LCDs, which are often composed of short tandem repeats of low-complexity [22], have been linked to stress granules and processing bodies (P-bodies) in eukaryotes (see [23] for recent review). The amino acid composition of these LCDs confers unusual biophysical properties to these domains [24], which likely relates to their unique behavior *in vitro* and *in vivo* [25–30]. However, these unusual characteristics appear to be inseparably linked to pathological processes as well. For example, genetic expansion of regions encoding homopolymeric glutamine repeats (the simplest type of LCD) in various proteins can lead to a multitude of neurodegenerative disorders, including Huntington's Disease and spinocerebellar ataxias (for review, see [31]). Furthermore, mutations in the LCDs of stress granule proteins can alter stress granule

dynamics and lead to degenerative diseases [26,28,30,32,33]. The importance of LCDs extends well beyond Q/N-rich LCDs, as LCDs of other compositions have also been linked to normal and pathological cellular processes [12,14,17,34,35].

Although LCDs can clearly impact protein regulation and function, a number of challenges have thus far limited a proteome-scale understanding of these relationships. One major challenge lies in defining LCDs. Current approaches use statistically-defined thresholds for sequence complexity or compositional bias [1,11], or arbitrarily-chosen repeat lengths for proteins with homopolymeric repeats [34–41]. Although these definitions of LCDs, compositionally biased sequences (herein referred to as "statistically-biased domains" to avoid later confusion), or homopolymeric repeats have facilitated important discoveries, the biological relevance of these thresholds has not been rigorously examined. Furthermore, these proteins are often grouped into a single class even though their compositions, and therefore physical properties, can differ dramatically (a limitation that was appreciated in a recent review [42]).

To address these limitations, we have developed an alternative approach to infer relationships between amino acid composition and protein metabolism and function. By focusing on amino acid composition, which is the fundamental feature underlying both sequence complexity and statistical amino acid bias, we examined links between local compositional enrichment and various aspects of protein regulation and function without appealing to pre-defined sequence complexity or statistical bias thresholds. We find that local compositional enrichment correlates with differences in nearly all core aspects of a protein's tenure in the cell, including translation efficiency, abundance, half-life, protein-protein interaction promiscuity, subcellular localization, and function. However, enrichment for different amino acids is associated with different effects, even for residues often grouped based on physicochemical similarities, highlighting the importance of distinguishing LCDs of different types. These relationships are discernible at compositions below those required for classification as low-complexity or statistically-biased, suggesting that the thresholds in traditional methods may not be biologically optimized. Finally, analysis of experimentally-defined protein components of stress granules and P-bodies reveals both shared and distinct compositional features associated with these organelles.

Results

Systematic survey of local amino acid composition

Fundamentally, both sequence complexity and statistical amino acid bias are indirect measures of local amino acid composition. Since composition is a more direct indication of overall protein domain properties, we sought to examine whether composition alone could be used to infer residue-specific relationships between local amino acid composition and protein regulation and function. We first developed an algorithm to partition the yeast proteome on the basis of maximum local composition for each amino acid using a series of scanning window sizes (Fig 1; see Methods). For all amino acids, the majority of proteins are partitioned into composition bins of $\leq$ 25% (Fig 2 and S1 Table). However, the number of proteins achieving higher local compositions, indicated by a right-hand shoulder or tail in the distribution, were strongly residue-dependent. For example, proteins containing local enrichment of highly hydrophobic residues (I, L, M, and V), aromatic residues (F, W, and Y), or cysteine are almost exclusively limited to composition bins of $\leq$ 45% for the smallest window size, whereas alanine and proline distributions extend to slightly higher composition ranges (up to 60–65%). Proteins containing local enrichment of polar (G, N, Q, S, and T) or charged (D, E, and K) residues in composition bins of $\geq$ 40% are relatively common even among larger window sizes (albeit to differing degrees), whereas histidine and arginine rich regions are relatively rare.

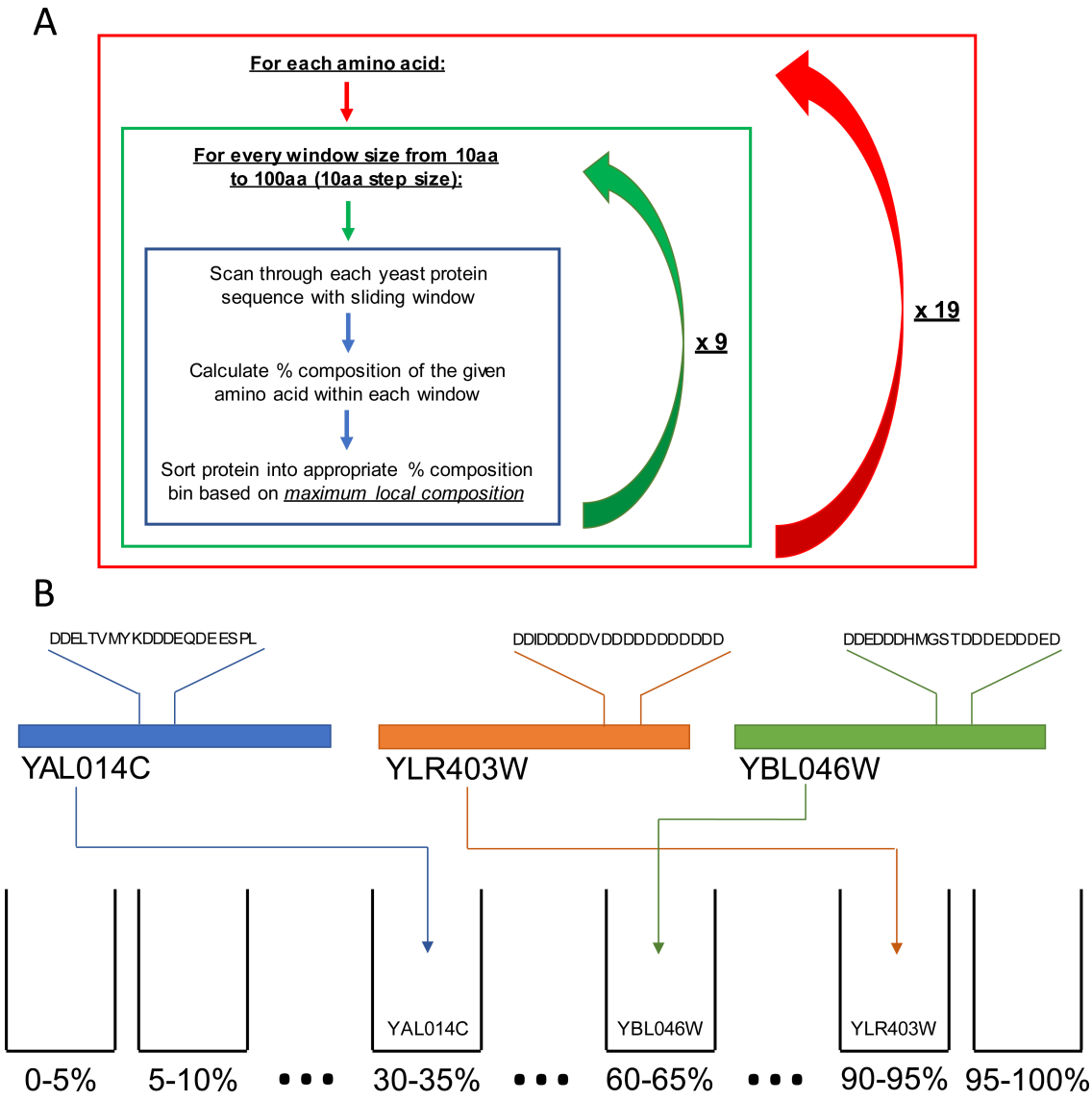

Fig 1. Depiction of proteome sorting on the basis of maximum local composition. (A) For each amino acid and window size combination, each yeast protein is sorted into percent composition bins based on the maximum local composition of the amino acid within the given sliding window size. This effectively sorts the yeast proteome 200 distinct ways (20 amino acids x 10 different sliding window sizes). (B) Visual representation of proteins sorted based on maximum local aspartic acid composition with a 20 amino acid sliding window.

https://doi.org/10.1371/journal.pcbi.1006256.g001

These data indicate that relatively high local enrichment is tolerated for some amino acids, while compositional enrichment for other amino acids appears to be restricted in yeast.

Residue-specific relationships between local compositional enrichment and protein metabolism

While the origins and evolution of LCDs have been extensively explored [3,4,14,38,43,44], the regulation and metabolism of LCD-containing proteins remain poorly-understood. Proteins with intrinsically disordered segments, which often qualify as LCDs [45,46], have been associated with lower protein half-lives [47]. However, not all intrinsically disordered regions lead to

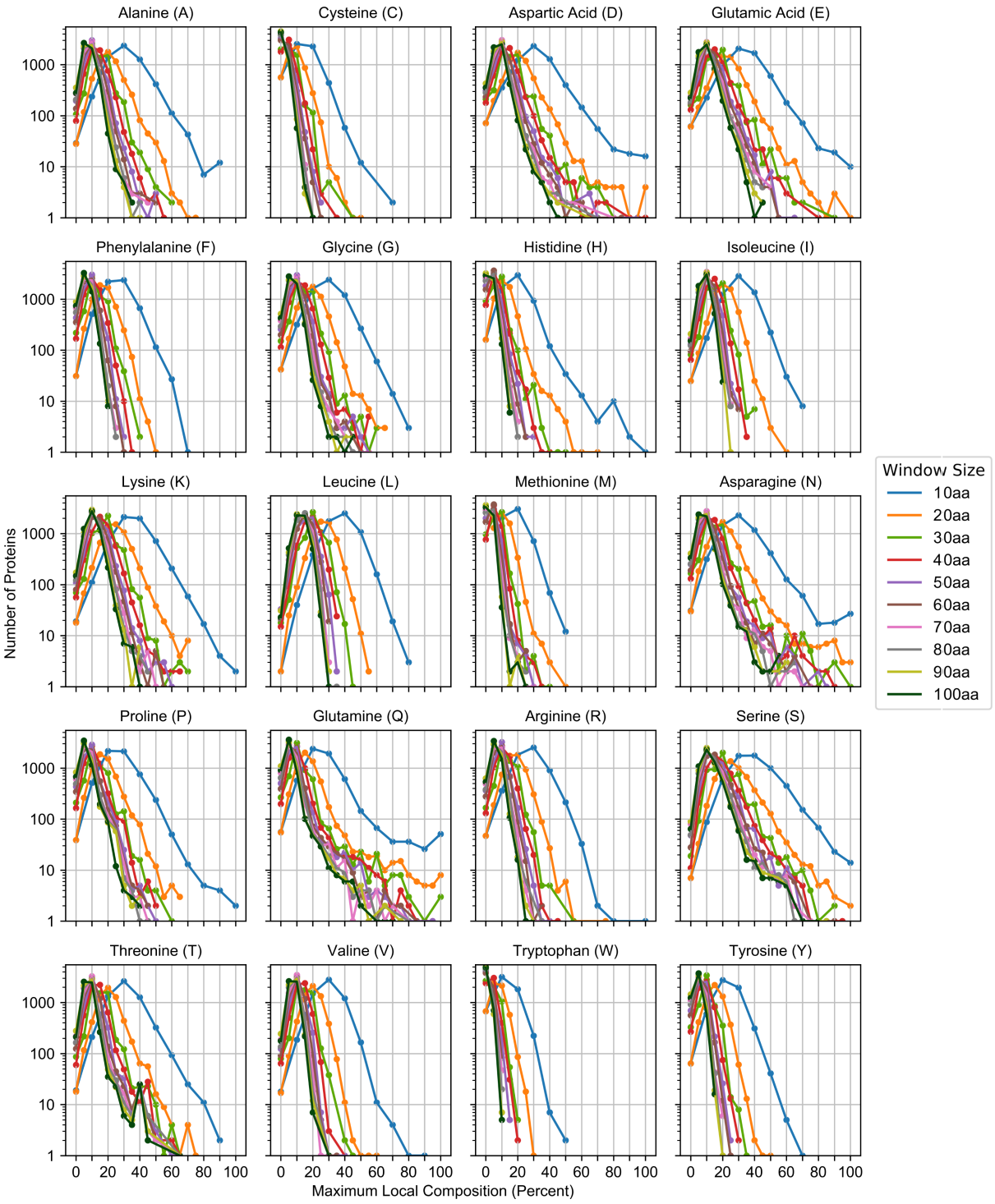

Fig 2. Distribution of the yeast proteome based on maximum local amino acid composition. The number of proteins partitioned into each window size/percent composition bin for each of the 20 canonical amino acids are plotted as a function of maximum local composition for each window size. Scatter points are connected by line segments for visual clarity only.

https://doi.org/10.1371/journal.pcbi.1006256.g002

short protein half-lives, and not all LCDs are intrinsically disordered [15]. Additionally, proteins with homopolymeric repeats, when considered as a single class, are associated with lower

translation efficiency, lower protein abundance, and lower protein half-life compared to proteins lacking homopolymeric repeats [37]. However, the regulation and structural properties of proteins with LCDs or homopolymeric repeats is likely strongly dependent on the predominant amino acids within the domain of interest [42].

To explore relationships between local compositional enrichment and protein metabolism, we first examined possible links between local compositional enrichment and protein abundance. Recent advances in proteomic methods have facilitated remarkable proteome coverage for both protein abundance [48] and protein half-life [49] measurements in yeast. At each window size/percent composition bin, the distribution of protein abundance values for all proteins partitioned into that bin was compared to the protein abundance distribution for all other yeast proteins (Mann-Whitney *U* test). Transitions from significantly lower median abundance to significantly higher median abundance or vice versa are observed upon enrichment for many amino acids individually (Fig 3). However, the direction of the trends upon progressive compositional enrichment are dependent on amino acid type. For the majority of amino acids (C, D, F, H, I, L, M, N, P, Q, R, S, T, W, or Y) compositional enrichment is associated with lower median protein abundance. However, compositional enrichment of A, G, or V is associated with higher median protein abundance. Two very similar transitions are observed for both E-rich and K-rich sequences: as compositional enrichment increases, the relative median protein abundance transitions from high to low, then back to high. Collectively, these trends are consistent with, yet much stronger than, previously observed correlations between protein abundance and whole-protein composition [50,51]. This suggests that the trends observed previously may actually reflect the effects of local compositional enrichment, which would increase apparent whole-protein composition for the enriched amino acid yet be dampened by confounding effects from the remainder of the protein sequence.

Similar trends are observed when compositional enrichment is compared to protein half-lives (Fig 4). Compositional enrichment for the majority of amino acids (C, H, K, M, N, P, S, or T) is associated with lower protein half-life, whereas enrichment for A, G, I, or V is associated with higher protein half-life. Enrichment for F leads to an initial transition from lower to higher half-lives, while further enrichment leads to a transition back to lower half-lives. It is worth noting that similar trends were observed in an independent protein half-life dataset when the proteins were analyzed based on whole-protein amino acid composition [52], suggesting that maximum local composition is sufficient to detect associations between amino acid composition and half-life. Although for many amino acids the trends are readily apparent, the strength of the association between compositional enrichment and protein half-life appears to be slightly weaker than the association between compositional enrichment and protein abundance. This is likely due, at least in part, to limited proteome coverage (relative to the protein abundance dataset). However, a recent study also suggested that protein half-life is strongly affected by factors other than sequence characteristics [53], which would likely further dampen relationships between compositional enrichment and protein half-life. Finally, protein half-life is generally less-conserved than protein abundance [54], perhaps suggesting that specific relationships between conserved sequence features and protein half-life may not be particularly strong. Therefore, it is rather surprising that we observe the indicated trends in spite of these limitations, and could suggest that half-life is more strongly influenced by local composition than particular primary sequence motifs.

Direct measurement of protein synthesis rates is more experimentally challenging. Consequently, proteome-wide coverage for experimentally-derived translation efficiency remains substantially lower than coverage for protein abundance and half-life. The normalized translation efficiency (nTE), a reported metric of translation elongation efficiency [55], is based on codon usage frequencies and tRNA gene copy numbers, allowing for calculation of translation

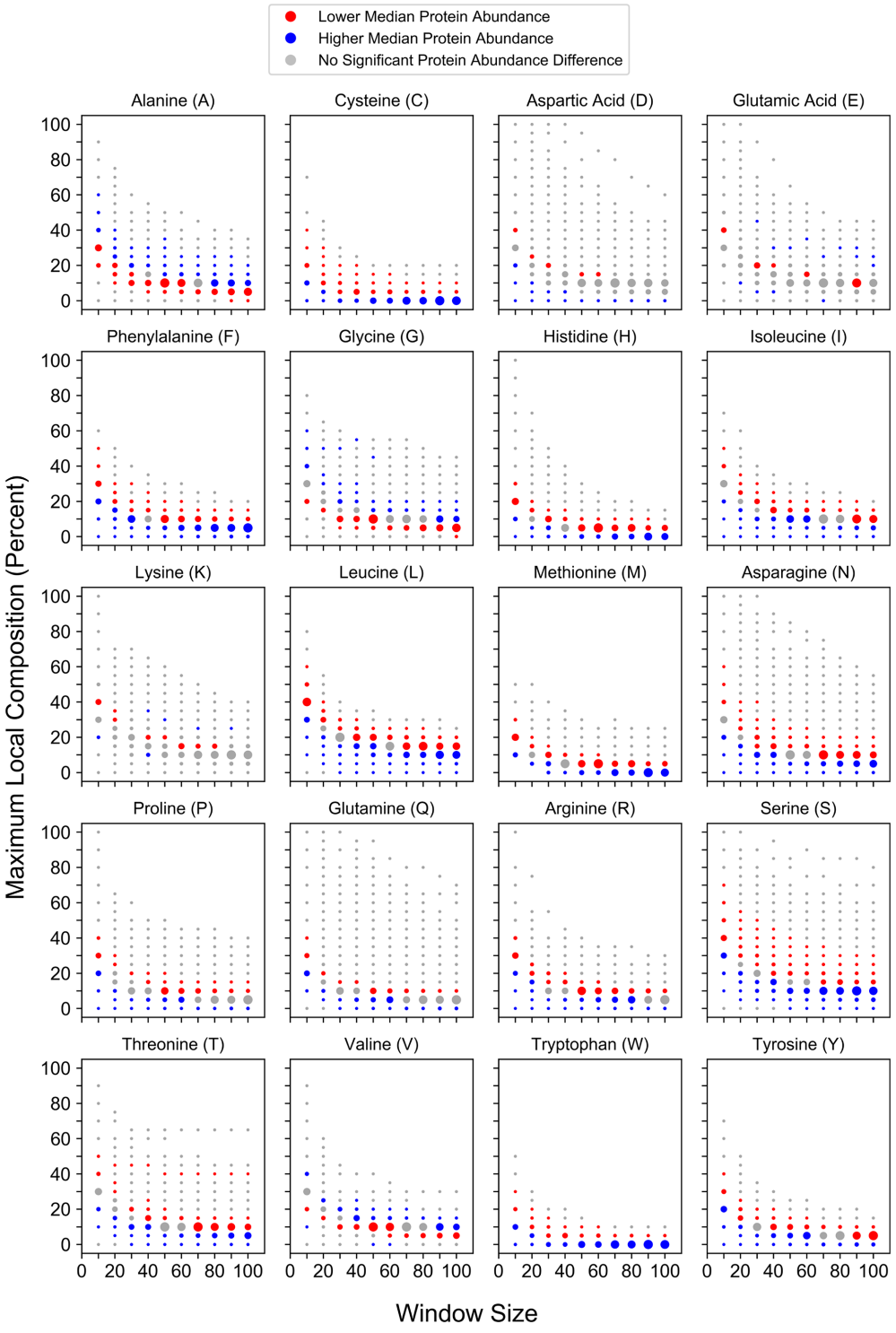

Fig 3. Maximum local amino acid composition is associated with residue-specific differences in protein abundance. For each amino acid, protein abundance values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length $\geq$ the corresponding window size that were excluded from the bin. For this figure and related subsequent figures, red and blue points indicate bins for which the distribution of protein abundance values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicate bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.

https://doi.org/10.1371/journal.pcbi.1006256.g003

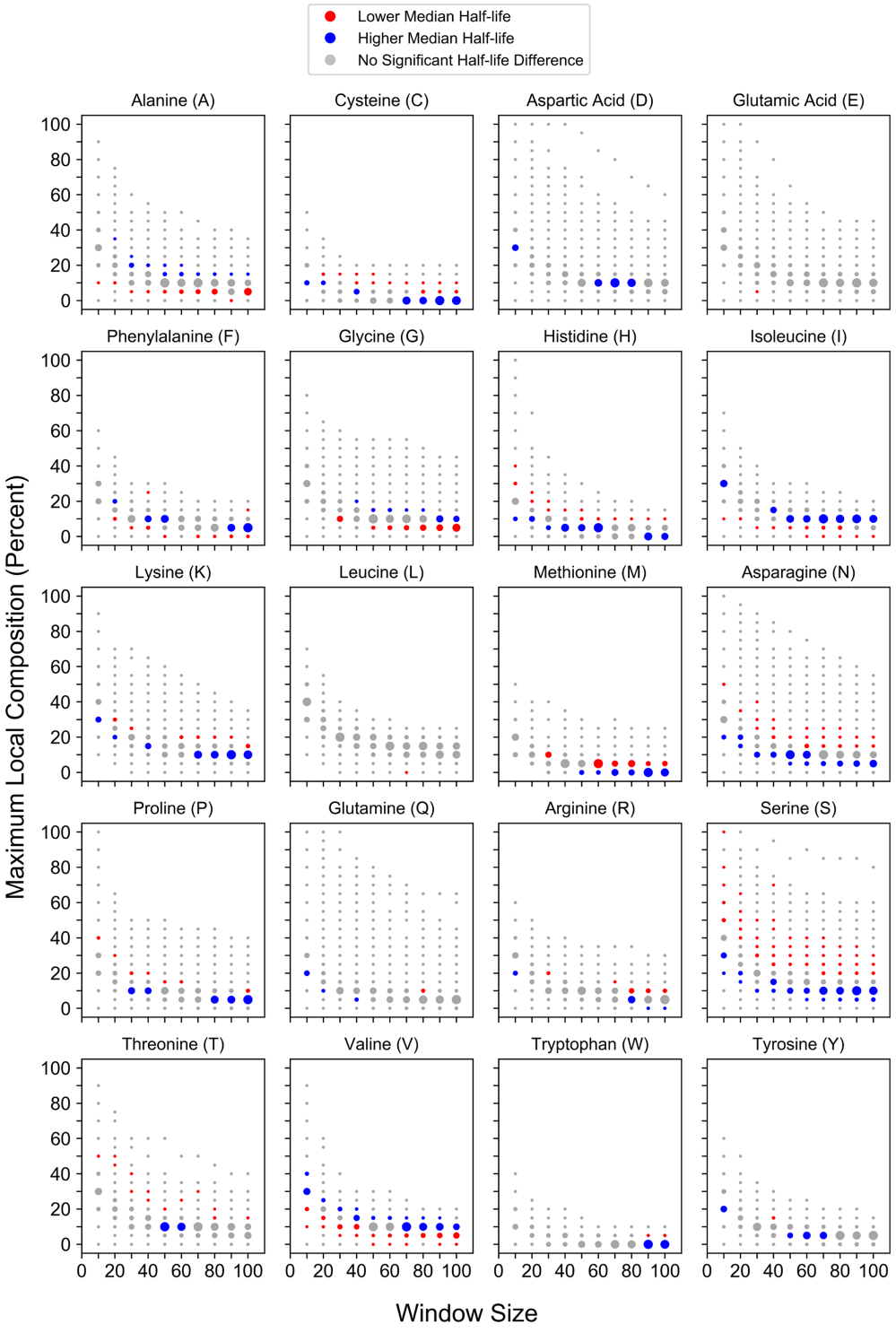

Fig 4. Maximum local amino acid composition is associated with residue-specific differences in protein half-life. For each amino acid, protein half-life values corresponding to proteins partitioned into a given window size and percent composition bin were compared to half-life values for all proteins of length ≥ the corresponding window size that were excluded from the bin.

https://doi.org/10.1371/journal.pcbi.1006256.g004

efficiency for the entire proteome. Therefore, we first examined relationships between local compositional enrichment and calculated translation elongation efficiency. nTEs were calculated for whole-protein sequences using the corresponding coding region on mRNA transcripts (see Methods). Translation efficiency is strongly dependent on the locally-enriched amino acid (Fig 5). For the majority of amino acids (C, D, E, F, H, I, K, L, M, N, P, Q, R, or Y), local enrichment is associated with significantly lower median nTEs suggesting that, as a single class, proteins with local compositional enrichment tend to be translated relatively inefficiently. Proteins with domains enriched in S, T, or W are generally associated with significantly lower median nTEs, although proteins with very high S, T, or W enrichment are associated with significantly higher median nTEs. However, proteins with domains enriched in A, G, or V residues are consistently associated with significantly higher median nTEs, suggesting that these proteins may be translated relatively efficiently. Remarkably, nearly identical trends are observed between local compositional enrichment and the experimentally-derived protein synthesis rates reported for a limited proteome (S1 Fig) despite a substantial reduction in sample size (n = 1115; [56]), suggesting that nTE can serve as a good surrogate for overall protein synthesis efficiency. Collectively, these results indicate that local amino acid enrichment is associated with differences in protein production rates in a composition-dependent manner.

For most amino acids, we noticed a remarkable correspondence in the trends for translation efficiency, protein abundance, and protein half-life, despite the fact that these values are derived from entirely different methods and experiments. For example, local enrichment for many amino acid types is associated with low nTE values, low protein abundance, and low protein half-life (Table 1). While translation efficiency and protein degradation rate are largely functionally independent in cells, protein abundance depends, at least in part, on both translation efficiency and protein half-life [49]. This may suggest that protein abundance for these proteins is limited in cells by a combination of poor translation efficiency and rapid degradation rate. In contrast, local enrichment for some amino acids is associated with high protein abundance also tended to have higher nTE values and higher half-lives, perhaps suggesting that high protein abundance for these proteins is achieved by a combination of efficient translation and poor degradation.

Nearly-identical, residue-specific relationships between local compositional enrichment and protein abundance are observed in *C. elegans*

As a model eukaryotic organism, *S. cerevisiae* provides a number of important advantages in proteome-scale studies relating protein sequence to protein metabolism and function. In addition to the unmatched proteome coverage in protein abundance and protein half-life datasets, and the availability of yeast-specific tools such as nTE, sequence-function analyses in yeast are further simplified by the absence of tissue-specific effects and limited alternative splicing (only ~4% of yeast genes contain introns and, of those genes, only a small fraction is capable of producing alternative protein products [57,58]).

With these caveats in mind, we sought to examine whether similar relationships between local amino acid composition and protein abundance could be detected in a model multicellular eukaryotic organism. We decided to focus on whole-organism protein abundance measurements in *C. elegans* [59] for four main reasons: 1) due to technical experimental challenges, protein abundance measurements in *C. elegans* are substantially more robust than protein half-life measurements; 2) on a proteome-wide scale, protein abundance is more strongly conserved across yeast species than protein half-life [49], suggesting that final protein levels tend

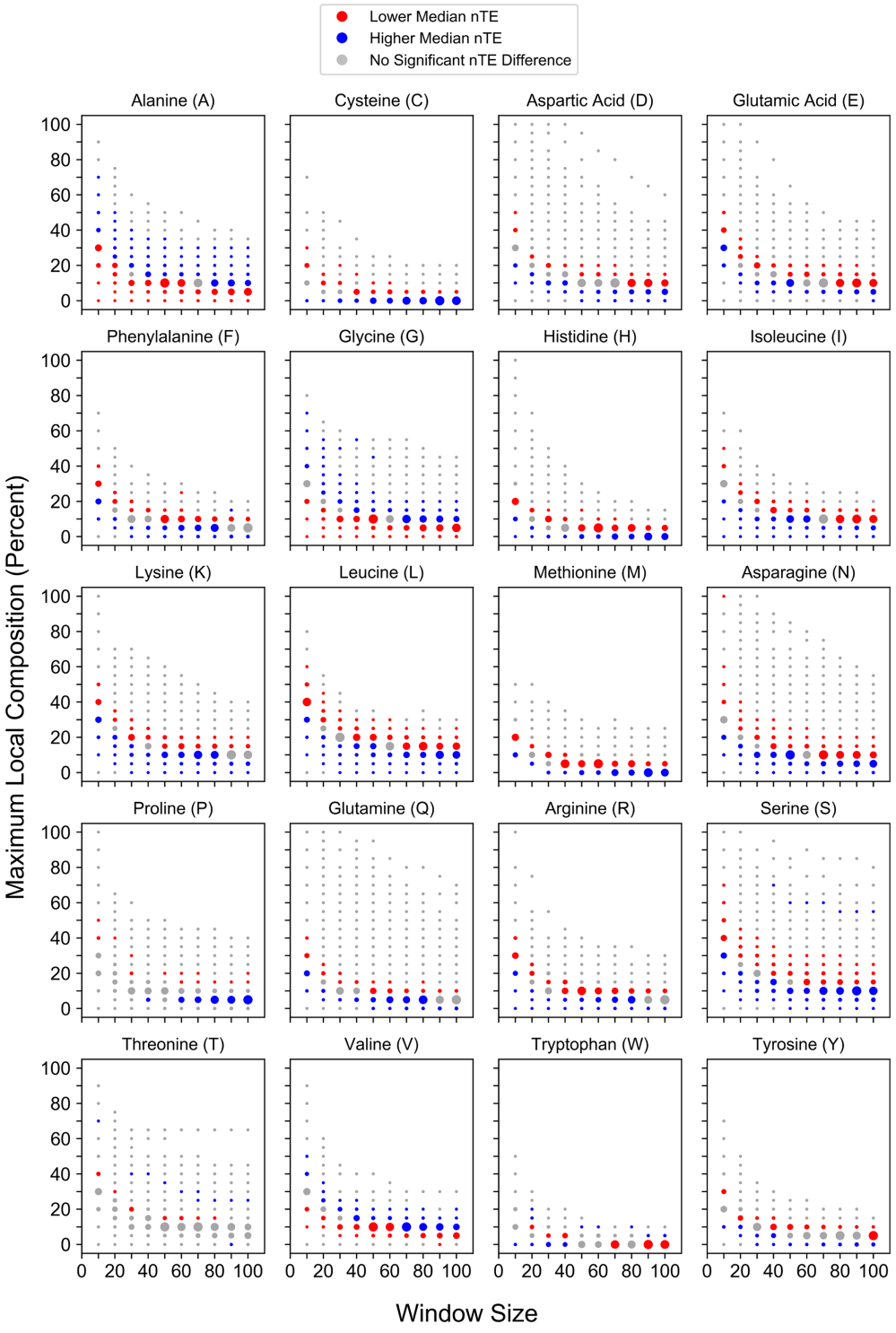

Fig 5. Maximum local amino acid composition is associated with residue-specific differences in nTE. For each amino acid, nTE values corresponding to proteins partitioned into a given window size and percent composition bin were compared to nTE values for all proteins of length $\geq$ the corresponding window size that were excluded from the bin.

https://doi.org/10.1371/journal.pcbi.1006256.g005

Table 1. The life-cycle of proteins with high local composition of individual amino acids involves the coordinated regulation of translation efficiency, protein abundance, and protein half-life. For each amino acid, trends in median values for nTE, protein abundance, and half-life upon enrichment (i.e. approaching higher percent compositions) of the given amino acid are indicated. "Higher" indicates that proteins in larger percent composition bins tend to have a larger median value compared to all other proteins, while "Lower" indicates that proteins in larger percent composition bins tend to have a larger median value compared to all other proteins. "Mixed" indicates amino acids for which multiple transitions are observed upon progressive compositional enrichment. Datasets without clear, statistically significant transition thresholds are also indicated ("n.s."). Colors correspond to the colors used in Figs 3–5.

Amino Acid	nTE	Abundance	Half-life
A	Higher	Higher	Higher
C	Lower	Lower	Lower
D	Lower	Lower	n.s.
E	Lower	Mixed	n.s.
F	Lower	Lower	Mixed
G	Higher	Higher	Higher
H	Lower	Lower	Lower
I	Lower	Lower	Higher
K	Lower	Mixed	Lower
L	Lower	Lower	n.s.
M	Lower	Lower	Lower
N	Lower	Lower	Lower
P	Lower	Lower	Lower
Q	Lower	Lower	n.s.
R	Lower	Lower	n.s.
S	Mixed	Lower	Lower
T	Mixed	Lower	Lower
V	Higher	Higher	Higher
W	Mixed	Lower	n.s.
Y	Lower	Lower	n.s.

https://doi.org/10.1371/journal.pcbi.1006256.t001

to be constrained across organisms, while regulation of the metabolic pathways that contribute to protein abundance may vary; 3) protein abundance is, at least partially, a function of translation efficiency and protein half-life; and 4) the parameters underlying the translation efficiency method (namely the "*s*-vector", or the efficiency of wobble base pairing between tRNA isoacceptors) were optimized for yeast [60]. Therefore, the nTE method may not be amenable to application in other organisms.

In order to examine relationships between maximum local composition and protein abundance, we first determined the proteome distribution of *C. elegans* proteins as a function of maximum local composition for each amino acid. The *C. elegans*-specific proteome distributions (S2 Fig and S2 Table) were overall quite similar to the yeast proteome distributions (Fig 2). However, the maximum local composition for S and N appear to be slightly more constrained in *C. elegans* (indicated by contraction of the shoulder to lower maximum compositions), while G, P, and T achieve slightly higher maximum local compositions, indicating relaxed constraints on local enrichment of these residues. These results are consistent with previous observations noting both shared and organism-specific homopolymeric repeat signatures or bulk proteome compositions across proteomes from different organisms [4,38,40,41,44–46,61].

As observed in yeast, progressive compositional enrichment results in a transition from higher to lower median abundance for the majority of amino acids with a clear trend (C, F, I, M, N, P, S, W, and Y; Fig 6). Furthermore, all three amino acids (A, G, and V) that exhibit a

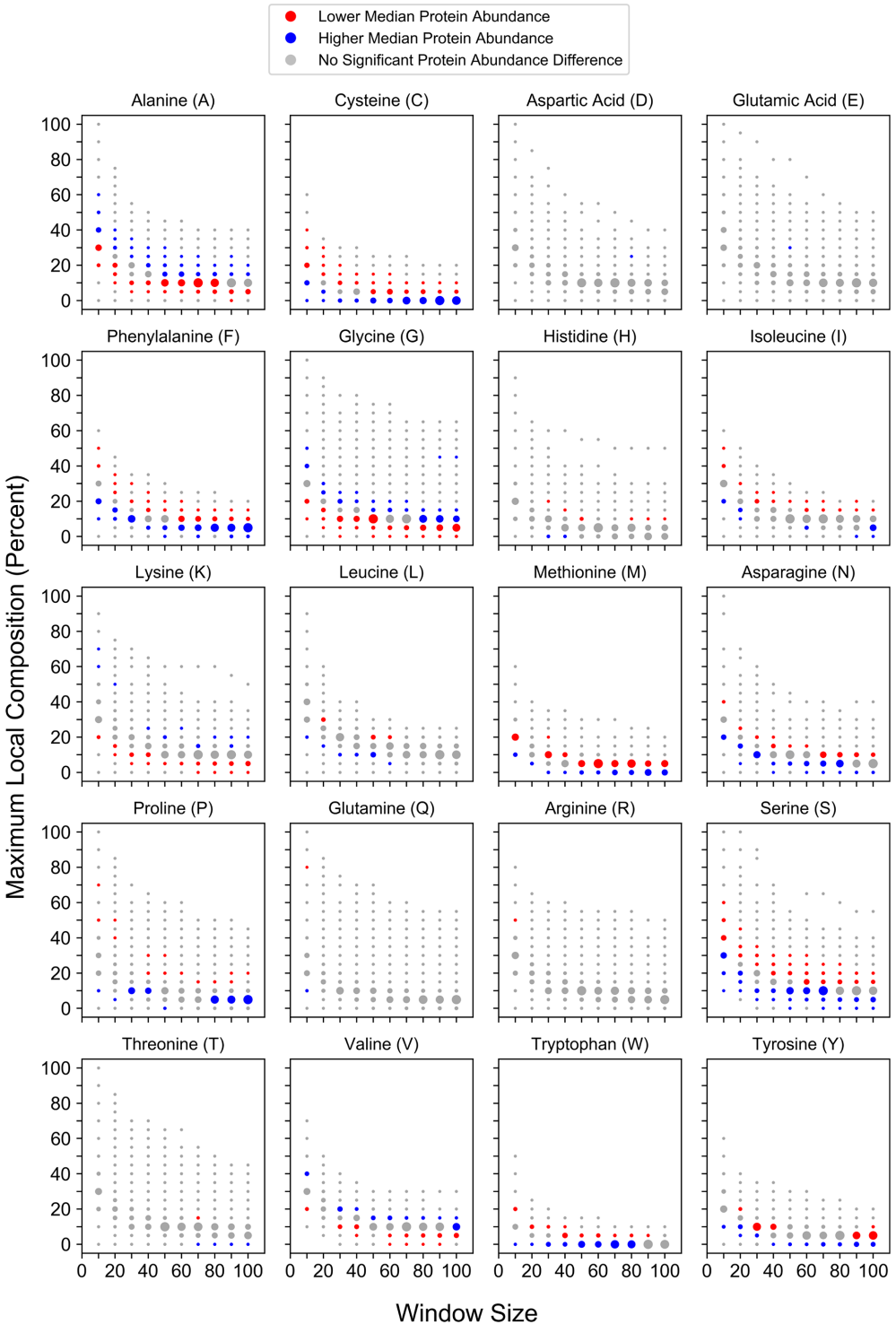

Fig 6. Relationships between maximum local composition and protein abundance in *C. elegans*. For each amino acid, protein abundance values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length ≥ the corresponding window size that were excluded from the bin. Trends in protein abundance as a function of maximum local composition for many amino acids are remarkably similar between yeast and *C. elegans* (compare with Fig 3).

https://doi.org/10.1371/journal.pcbi.1006256.g006

transition from lower to higher median abundance upon progressive enrichment in yeast exhibit the same trend in *C. elegans* as well. Indeed only one amino acid with a clear transition in protein abundance upon local enrichment differs between *C. elegans* and *S. cerevisiae*: in yeast, local K enrichment is associated with mixed effects on protein abundance (depending on the degree of K enrichment), whereas in *C. elegans* local K enrichment is weakly (yet consistently) associated with higher protein abundance. Therefore, nearly identical residue-specific relationships are observed between local amino acid enrichment and protein abundance in a more complex eukaryote.

Compositional enrichment is linked to effects on protein metabolism in the absence of classical low-complexity, statistically-biased, and homopolymeric domains

An important advantage of approaching LCDs from a composition-centric perspective is the ability to examine relationships between amino acid composition and protein outcomes without appealing to pre-defined thresholds of statistical amino acid bias [11] or sequence complexity [1,10], which may not reflect biologically-relevant thresholds. Indeed, the transitions observed in the median translation efficiencies, protein abundances, and protein half-lives often occur at surprisingly mild levels of compositional enrichment, suggesting that these trends may be observed even in the absence of classically-defined statistically-biased or low-complexity domains.

Statistical amino acid bias conceptually parallels our investigation of compositional enrichment, and has been used to investigate the functions of proteins with statistically-biased domains [11,12]. To examine whether compositional enrichment may be linked to biologically-relevant effects on protein metabolism independently of statistically-biased domains, a conservative bias threshold was employed to define statistically-biased domains using previously developed methodology [12] (also, see Methods). Proteins with statistically-biased domains were then filtered from the yeast proteome (n = 866 statistically-biased proteins for the yeast translated proteome of sequences $\geq$ 30 residues in length). However, even in the absence of statistically-biased domains, compositional enrichment resulted in robust trends in translational efficiency, protein abundance, and protein half-life that re-capitulated those originally observed (S3–S5 Figs). This suggests that compositional enrichment affects protein metabolism at thresholds preceding those required for classification as statistically-biased by alternative methods.

The SEG algorithm, by default, employs substantially more relaxed criteria when classifying protein domains as low-complexity [1]. Indeed, of the 5,901 proteins of length $\geq$30 amino acids in the translated ORF proteome, 4,147 proteins contain at least one LCD, which is consistent with previous estimates [3]. Nevertheless, despite a large reduction in proteome size, many of the trends in protein metabolism are discernible even when all proteins with a SEG-positive sequence are filtered from the proteome (S6–S8 Figs). This suggests that compositional enrichment exerts biologically relevant effects even among non-LCD-containing proteins.

Proteins containing homopolymeric amino acid repeats (often defined as five or more identical amino acids in succession), were recently reported to have lower translation efficiency, lower protein abundance, and lower protein half-life when compared to proteins without homopolymeric repeats [37]. Homopolymeric repeats are effectively short sequences of maximum possible single-amino acid density. Therefore, proteins with homopolymeric repeats are expected to be disproportionately common among compositionally enriched domains, raising the possibility that the trends observed in the present study have been mis-attributed to

compositional enrichment alone. To examine this possibility directly, the relationship between compositional enrichment and nTE, abundance, and half-life was re-evaluated for a filtered proteome that excludes all proteins containing at least one homopolymeric repeat (n = 755 proteins excluded). While exclusion of these proteins preferentially reduces the sample sizes at higher compositional enrichment percentages, the absence of homopolymeric repeat proteins has little effect on the trends in nTE, abundance, and half-life as a function of compositional enrichment (S9–S11 Figs). This does not definitively rule out the possibility that homopolymeric repeats may, in some way, specifically affect translation efficiency, abundance, and half-life. However, since homopolymeric repeats *per se* are not absolutely required, the effects of homopolymeric repeats may instead be explained simply by local compositional enrichment.

Collectively, these results suggest that compositional enrichment affects translation efficiency, protein abundance, and protein half-life at thresholds preceding those required for classification as low-complexity or statistically-biased by traditional methods. It is worth noting that in the course of eliminating proteins with classically-defined low-complexity, statistically-biased, or homopolymeric domains, proteins with multiple distinct domains strongly enriched in different amino acid types, or with single domains strongly enriched in more than one amino acid, are eliminated from the proteome before re-evaluation. Therefore, the trends in protein metabolism observed upon enrichment of a given amino acid are not due to confounding effects of domains strongly enriched in other amino acids occurring within the same protein sequences.

Local compositional enrichment influences protein-protein interaction promiscuity in a residue-specific manner

Local enrichment of a single amino acid can dramatically influence the physicochemical properties of a given protein domain [24]. In a cellular context, these physicochemical properties likely influence interactions between proteins and surrounding molecules, including other proteins.

To examine whether local compositional enrichment affects protein-protein interactions, we explored relationships between enrichment for each of the amino acids and protein-protein interaction promiscuity (defined as the number of unique interacting partners per protein). Proteins found in a range of high-percent composition-bins for most amino acids (A, D, E, G, K, N, P, Q, R, and V) are associated with significantly more interacting partners relative to all other proteins (Fig 7), suggesting that these domains are relatively promiscuous. Additionally, proteins with mild enrichment for select hydrophobic residues (I, L, and M) are generally associated with more interacting partners, although fewer comparisons reach statistical significance (blue or red dots). These results are consistent with previous reports that, as a single class, proteins with LCDs or homopolymeric repeats tend to have more protein-protein interaction partners [16,37]. However, proteins in a range of high-percent composition-bins for each of the aromatic residues (F, W, and Y) are associated with significantly fewer interacting partners relative to other proteins, suggesting that aromatic residues tend to lack the interaction promiscuity observed at higher percent compositions for other amino acids. Furthermore, proteins with moderate to high local C content and proteins with extremely high maximum local S or T content are also associated with significantly fewer interacting partners relative to other proteins, suggesting that these domains are relatively non-promiscuous as well. This is particularly interesting, given that these trends were not observed upon enrichment for other polar residues. Again, this highlights the potential pitfall of grouping amino acids with related physicochemical properties into a single category. Collectively, these results indicate that

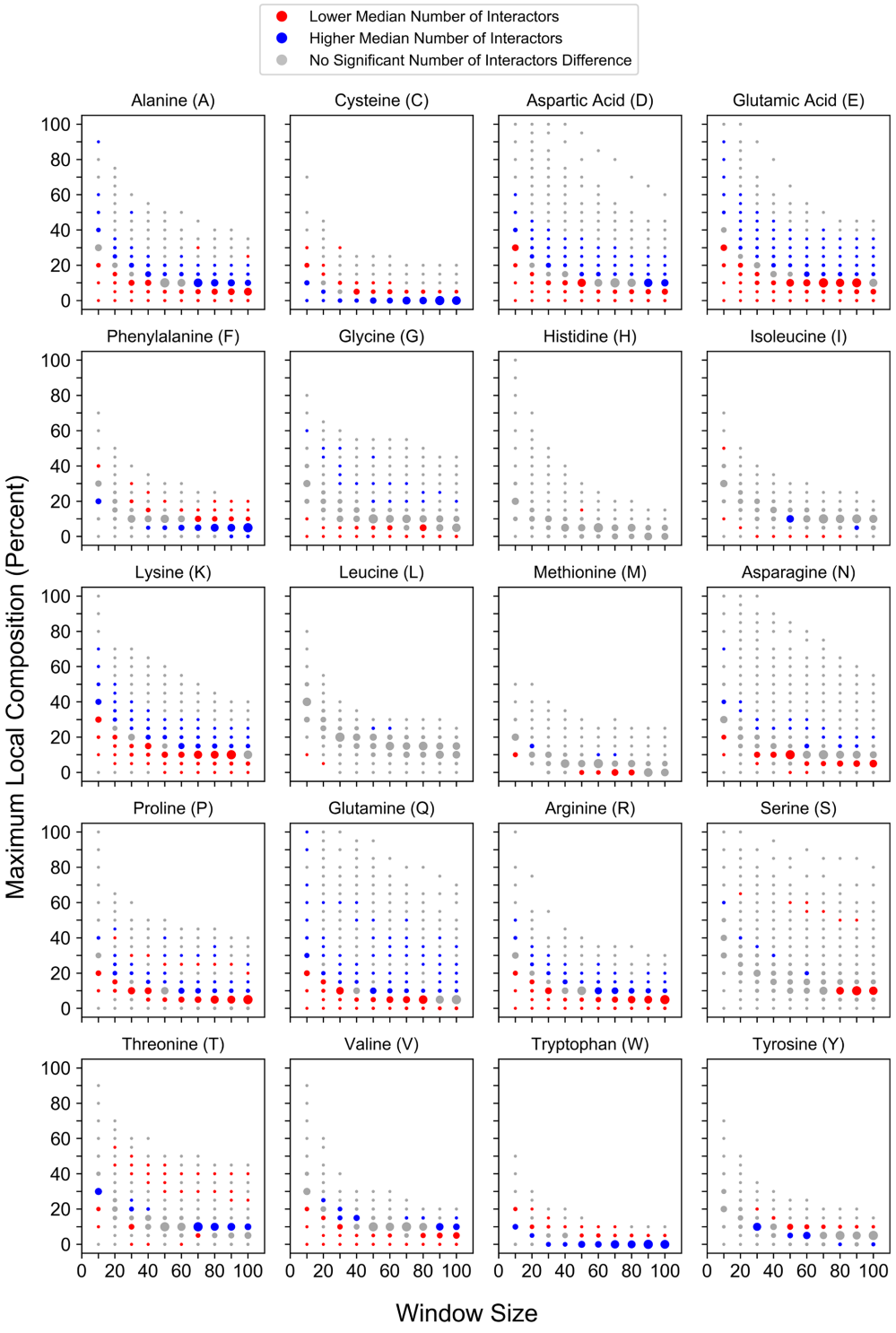

Fig 7. Maximum local amino acid composition corresponds to protein-protein interaction promiscuity in a residue-specific manner. Local enrichment for individual amino acids corresponds to composition-dependent changes in the number of unique protein-protein interaction partners.

https://doi.org/10.1371/journal.pcbi.1006256.g007

protein-protein interaction promiscuity varies for proteins with high compositional enrichment in a residue-specific manner.

Proteins with high compositional enrichment can fulfill overlapping or specialized molecular roles in the cell

Previous studies have attempted to associate proteins containing LCDs, statistically-biased domains, and homopolymeric repeats with particular cellular functions [12,16–18,34,37]. However, one important consideration when inferring relationships between proteins with LCDs and cellular functions, for example, is the prevalence of proteins with multiple LCDs [3], and of LCDs strongly enriched in more than one amino acid type [11,14,18,36]. Therefore, attempts to associate cellular functions to specific LCD types, without controlling for other LCDs within the same protein sequences, risk mis-attributing functions to unrelated protein features [12,14,34,36]. While multiple LCDs within the same protein (or multiple amino acid types enriched within the same LCD) may cooperate to generate novel structures or functions, this complicates interpretation of the role of each individual amino acid type within LCDs. Furthermore, because some types of LCDs are more common than others, general attempts to associate cellular functions with LCDs, statistically-biased domains, or homopolymeric repeats likely reflect the functions associated with only the most common types when considered as a single, unified class [16,37]. Therefore, definitive assignment of cellular functions to each individual class of LCD necessitates exclusion of proteins with other types of LCDs.

In order to minimize possible confounding effects introduced by proteins with multiple regions enriched in different amino acid types, a modified version of the initial calculation performed by the SEG algorithm (namely, the Shannon entropy; see Methods) was employed to define proteins with only a single type of compositionally-enriched domain (CED). In an effort to incorporate our results (which indicate that compositional enrichment may exert biologically-relevant effects at compositions preceding the SEG algorithm threshold) into our definition of single-CED proteins, percent composition bins for which at least 75% of the residing proteins contained a SEG-positive sequence (as defined above) were pooled to generate a single list of CED-containing proteins for each amino acid. Proteins that contain multiple types of CEDs were then removed from the dataset, resulting in a non-redundant set of proteins with only one type of CED. Importantly, this method captures the exclusion of proteins containing more than one type of CED, as well as proteins with CEDs strongly enriched in more than one amino acid type.

Gene Ontology (GO) term analysis was performed separately for each window size within each single-CED category. For each type of CED, there is strong overlap in the enriched GO terms across the range of window sizes, suggesting that the associations between functions and residue-specific CEDs are not strongly length-dependent at this scale. Therefore, for simplicity of interpretation, significantly enriched GO terms for each window size were pooled to generate a single non-redundant list of enriched GO terms for each CED type.

Removal of proteins with multiple types of CEDs reveals a remarkable degree of specialization for CEDs of different types (Fig 8, and S3 Table), which is often not observed for CEDs when considered as a single category or when multi-CED proteins are not excluded. For example, L-rich proteins are predominantly associated with functions at the ER and vacuole membranes, whereas I-rich proteins are more strongly associated with carbohydrate transport at the plasma membrane. A-rich proteins are associated with a variety of processes or cellular components, including translation, protein kinase activity, the cell wall, and carbohydrate/alcohol catabolism. N-rich proteins are strongly (and perhaps exclusively) associated with functions related to transcription, whereas Q-rich proteins appear to be more weakly

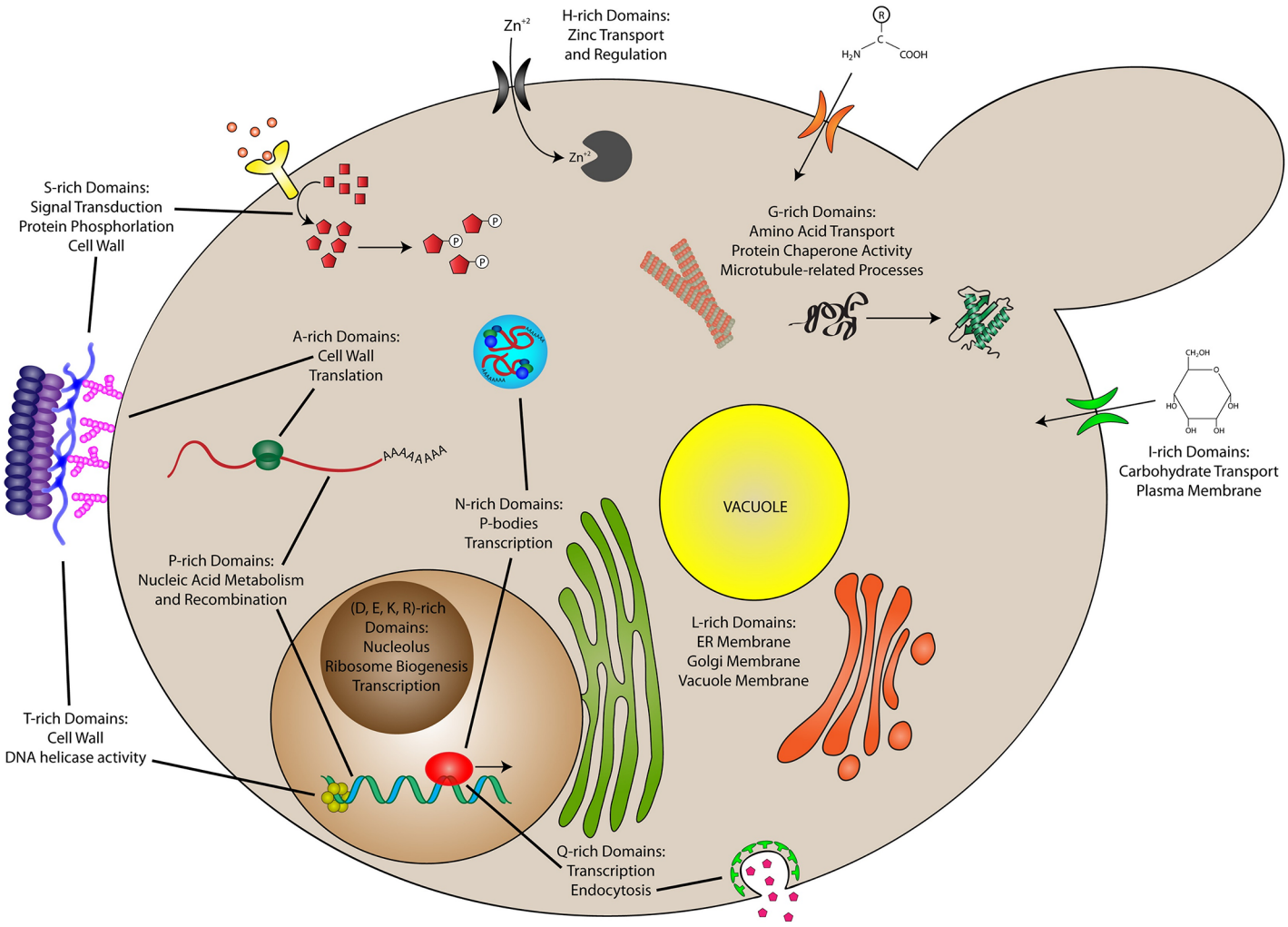

Fig 8. Cell model depicting predominant functions of CEDs. Residue-specific CEDs are associated with both overlapping and distinct functions. Main cellular/molecular processes associated for each type of CED are derived from significantly enriched (Bonferroni-corrected $p \leq 0.05$) GO-terms in S3 Table.

https://doi.org/10.1371/journal.pcbi.1006256.g008

associated with transcription and, instead, are associated with a larger variety of functions including endocytosis, mating projection of the membrane, and response to glucose. Finally, although yeast cell wall proteins are often radically S/T-rich, after controlling for co-enrichment of S and T in the same proteins, S-rich proteins are more strongly associated with membrane-related processes (cell wall, cellular bud tip, cellular bud neck, mating tip projection, etc.), protein kinase activity, and transcription, whereas T-rich proteins tend to be associated with nucleic acid binding and helicase activity, with fewer associations with membrane-related processes. Therefore, after controlling for the presence of multiple CEDs within the same proteins, specialized functions emerge even among commonly grouped amino acids.

Furthermore, CEDs enriched in some amino acids share functions despite the removal of multi-CED proteins, suggesting some degree of co-specialization. For example, D-, E-, and K-rich CEDs were each associated with functions in the nucleus/nucleolus, including ribosomal RNA processing, nucleic acid binding, transcription, and histone/chromatin binding. Intriguingly, intrinsically disordered domains with opposite net charges (along with other charged macromolecules such as nucleic acids and polyADP-ribose) can drive phase separation or

complex coacervation in the nucleus [62–64]. It is possible that these domains, along with nucleic acids and other polyionic molecules, may participate in nuclear processes via dynamic electrostatic association with these or other membraneless assemblies. By contrast, H-rich CEDs are associated with processes related to zinc ion transport and regulation. There were no GO terms significantly associated with R-rich CEDs. However, compositional enrichment for R appears to be constrained, as evidenced by the sharp decline in the number of proteins with R-rich domains toward higher maximum local percent compositions (see Fig 2), which may be further impacted by the removal of proteins with other types of CEDs.

In summary, when examined as separate classes, different types of CEDs can have overlapping or specialized roles in the cell.

Compositional enrichment corresponds to preferential localization to specific subcellular compartments

The molecular specialization observed for CEDs indicates that proteins with enrichment of particular residues may localize to particular subcellular compartments in order to execute their specialized functions. Furthermore, protein quality control factors can differ between subcellular compartments (for review, see [65]), which may contribute to composition-dependent differences in protein metabolism. Therefore, we applied a bottom-up approach to infer the composition profiles associated with the major subcellular compartments (see Methods).

Largely aqueous subcellular compartments are almost exclusively associated with proteins containing domains enriched in charged residues, polar residues, and proline (Fig 9; see also S12 Fig). However, differences in compositional enrichment profiles are apparent even among related aqueous compartments. For example, significant associations with charged, Q, or N residues reach more extreme percent compositions in the nucleus, whereas as significant associations with P enrichment reach higher percent compositions in the cytoplasm. By contrast, the highly membraneous internal organelles (e.g. the endoplasmic reticulum and Golgi apparatus) are predominantly associated with enrichment of hydrophobic or aromatic residues (Figs 9 and S13). The yeast vacuole is also associated with composition profiles resembling those of membraneous compartments, with additional weaker associations with S and C enrichment. Few weak associations are observed for mitochondria. The yeast cell wall is strongly associated with S enrichment (likely related to its ability to be glycosylated), with additional moderate associations with T and A enrichment, and a weak association with mild V enrichment (Figs 9 and S14). As expected, the plasma membrane is associated with enrichment for a variety of hydrophobic and aromatic residues. However, the plasma membrane is also significantly associated with enrichment of a select subset of polar residues (namely C, G, S, and T), further corroborating the specialized roles observed for these CEDs at the outer membrane. Indeed, G-rich CEDs are significantly associated with amino acid transport (see S3 Table), and S- or T-rich CEDs of the plasma membrane could have overlapping functions or interactions with S- and T-rich CEDs of the cell wall. Together, these observations indicate that subcellular compartments may tolerate or prefer proteins with specific types of CEDs.

Components of stress granules and processing bodies possess shared and unique compositional features

Recent observations indicate that a variety of Q/N-rich and G-rich domains can form highly dynamic protein-rich droplets in aqueous environments [25–30], a process referred to as liquid-liquid phase separation. These types of LCDs are prevalent among components of membraneless organelles such as stress granules and P-bodies [23]. Furthermore, stress granules and P-bodies share many properties with protein-rich liquid droplets formed *in vitro*,

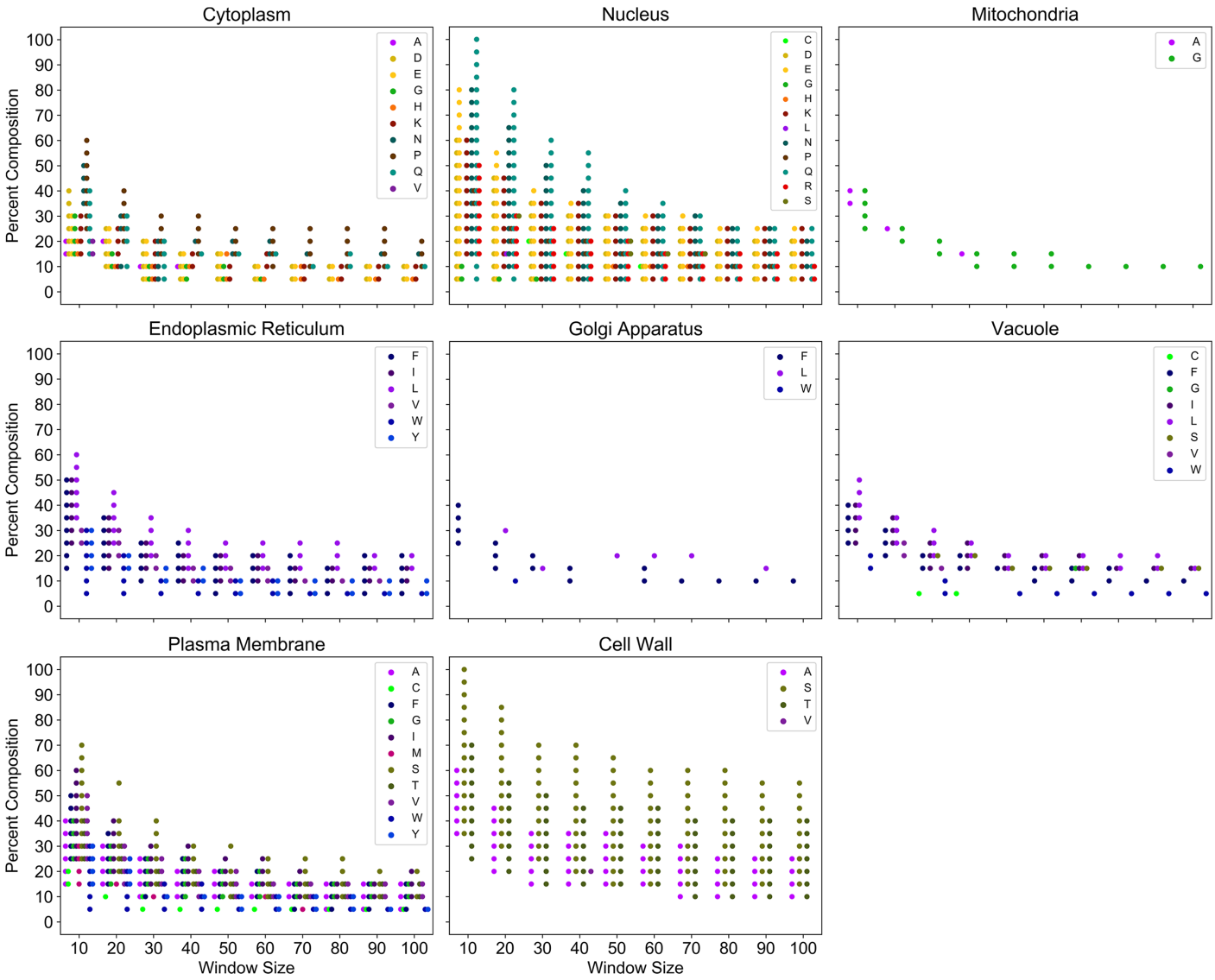

Fig 9. Compositional enrichment profiles associated with major subcellular compartments. All plotted points indicate protein sets for which association with the indicated subcellular compartment is statistically significant (Fisher's exact test, with Bonferonni-corrected $p < 0.05$). Warm colors (reds, oranges, and yellows) correspond to charged residues. Green colors indicate polar residues. Cool colors (purples and blues) correspond to hydrophobic and aromatic residues respectively.

https://doi.org/10.1371/journal.pcbi.1006256.g009

suggesting that the fundamental biophysical properties of these domains are related to the formation of membraneless organelles *in vivo*. However, while amino acid composition is acknowledged as a critical determinant of this behavior, the precise compositional requirements associated with membraneless organelles remain largely undefined.

Therefore, we also applied our bottom-up approach to infer the compositional enrichment profiles associated with protein components of stress granules and P-bodies (as defined in [66]). Stress granules and P-bodies have overlapping protein constituents and can exchange protein components [67,68], suggesting that they are closely related yet distinct organelles. Accordingly, we observe both shared and unique features in the composition profiles associated with stress granule and P-body proteins (Fig 10). As expected, both stress granules and P-bodies are strongly associated with proteins containing Q-rich or N-rich domains. For

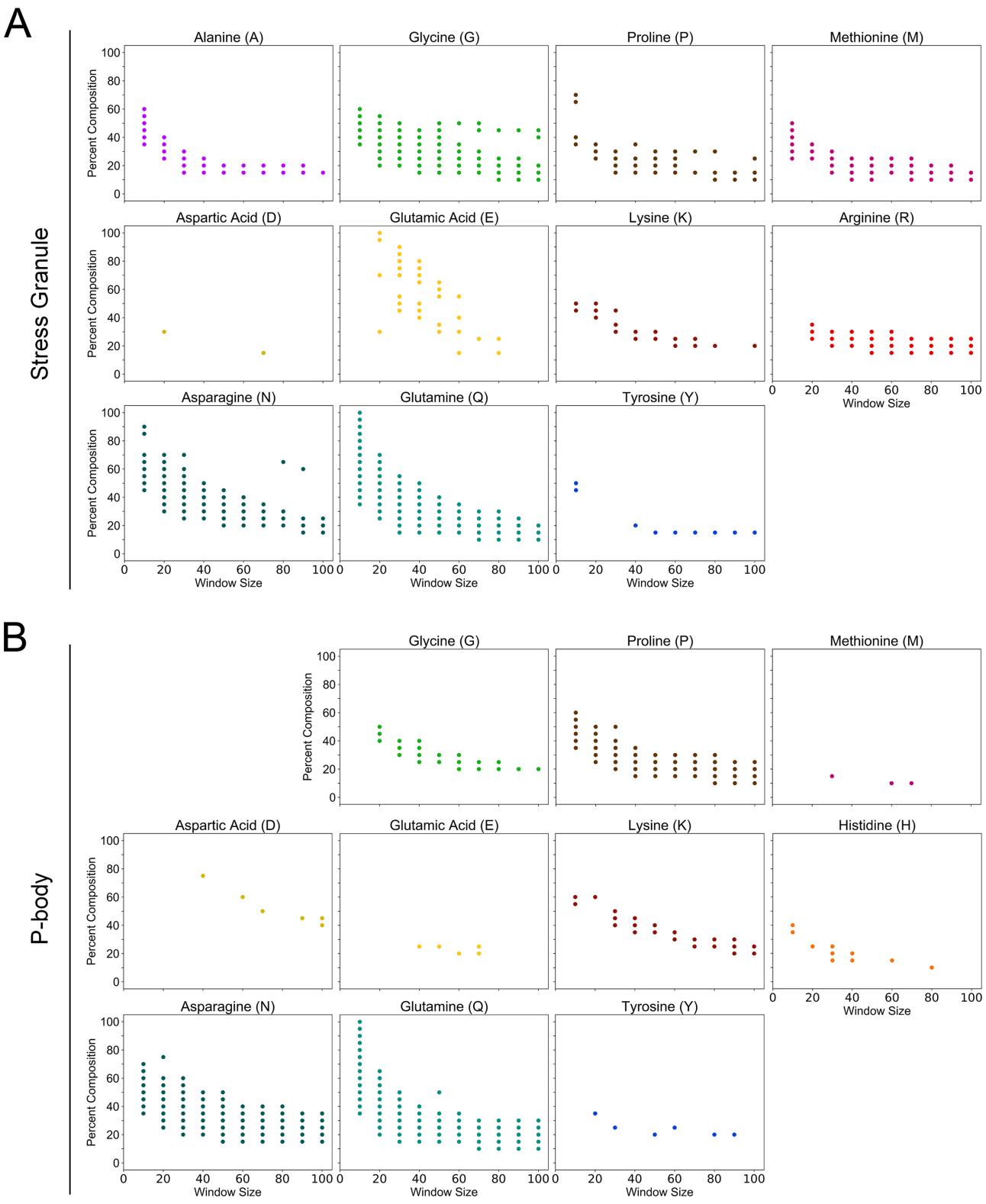

Fig 10. Composition profiles associated with membraneless organelles. All colored points indicate minimum percent composition thresholds for which components of stress granules (A) or P-bodies (B) are significantly enriched ($p < 0.05$). Only amino acids for which significant enrichment of stress granule or P-body proteins was observed in at least two composition bins are shown. For greater sensitivity, plots were generated using uncorrected *p*-values. Therefore, any individual point should be viewed with some skepticism: however, the presence of multiple consecutive significant points within each window size suggest that the observed trend is likely not an artifact of multiple hypothesis testing.

https://doi.org/10.1371/journal.pcbi.1006256.g010

example, minimum Q or N compositions significantly associated with stress granules range from ~15–100% at small window sizes ($\leq$30 amino acids) and ~10–30% at large window sizes ($\geq$80 amino acids), although these values vary slightly depending on window size and residue. Similarly, minimum Q or N compositions significantly associated with P-bodies range from ~15–100% at small window sizes and ~10–40% at larger window sizes.

In addition to the commonly appreciated link between stress granule/P-body components and Q/N-rich domains, we identify and define a variety of currently underappreciated compositional features common to stress granule and P-body components. Components of both stress granules and P-bodies are strongly associated with P-rich domains, weakly associated with K-rich domains, and very weakly (yet significantly) associated with Y-rich domains. Furthermore, while both stress granules and P-bodies are associated with proteins containing G-rich domains, stress granule components are associated with a much broader range of G enrichment, suggesting that G enrichment may be a more characteristic feature of stress granules than P-bodies. This is particularly striking in light of recent observations indicating that high glycine content helps maintain the liquid-like characteristics of phase-separated droplets and prevents droplet hardening *in vitro* [69].

Additionally, some compositional features are unique to either stress granules or P-bodies. For example, stress granule constituents are significantly associated with A-rich, M-rich, E-rich, and R-rich domains, whereas P-body constituents exhibit little or no preference for these compositional features (a key role for arginine in the phase separation of stress granule-associated proteins was also recently reported [69]). By contrast, P-body components are weakly associated with H-rich domains, whereas stress granule components are not enriched among proteins containing H-rich domains.

To our knowledge, this represents the first attempt to systematically define the range of amino acid compositions associated with membraneless organelles such as stress granules and P-bodies. These observations suggest that components of related, membraneless organelles have overlapping yet distinct compositional preferences. It is possible that shared compositional features facilitate the physical interactions between stress granules and P-bodies and allow for the exchange of components, while differences in compositional features facilitate their ability to function as independent organelles.

Discussion

Protein domains categorized as low-complexity, statistically-biased, or homopolymeric encompass broad, heterogeneous classes of sequences with diverse physical properties and cellular functions. These domains can play important roles in normal and pathological processes. However, challenges in categorizing proteins on the basis of sequence complexity or statistical bias have thus far precluded a complete, proteome-wide view of the effects of these domains on protein regulation and function. Here, we adopt an alternative, unbiased approach to examine proteome-wide relationships between local amino acid enrichment and the birth, abundance, functions, subcellular localization, and death of proteins. For nearly all amino acids, progressive local enrichment corresponds to clear transition thresholds with regard to translation efficiency, protein abundance, and protein half-life. Transition thresholds ubiquitously occurred at compositions preceding those required for classification as low-complexity or statistically-biased by traditional methods, indicating that our observed transition thresholds more closely reflect biologically-relevant composition criteria.

Protein sequences can range from perfectly diverse (i.e. a completely homogeneous mixture of amino acids with maximal spacing between identical amino acids) to lacking any diversity (i.e. homopolymeric sequences). While homopolymeric regions represent an extreme on this

spectrum and can influence protein metabolism [37], classically defined homopolymeric regions are not absolutely required for these effects (see S9–S11 Figs). This suggests that compositional enrichment may affect protein metabolism even upon some degree of primary sequence dispersion (i.e. greater linear spacing between identical amino acids). Defining the limits of this dispersion may shed additional light on the relationship between amino acid composition and protein metabolism.

An advantage of assessing compositional enrichment (as opposed to sequence complexity) is the ability to distinguish the effects of compositional enrichment for each amino acid type. The nature of the trends in translation efficiency, protein abundance, and protein half-life depend on the amino acid enriched in the protein sequences, indicating that local enrichment of different amino acids can have opposite effects. This highlights a key limitation when considering low-complexity, statistically-biased, or homopolymeric domains as a single class–grouping domains composed of radically different amino acids effectively skews any trends observed toward those of the most common type and, in some cases, can completely mask the effects of less common low-complexity, statistically-biased, or homopolymeric domains. Furthermore, even grouping these domains on the basis of common physicochemical properties can introduce the same complication. This is exemplified by the non-aromatic hydrophobic amino acids; while I-rich, L-rich, and M-rich domains are associated with poor translation efficiency, low abundance, and rapid degradation rate, A-rich and V-rich domains are associated with high translation efficiency, high abundance, and slow degradation rate. Additionally, the cellular functions associated with domains enriched in hydrophobic residues tend to differ; L-rich domains are predominantly associated with the ER or vacuole membrane, whereas I-rich domains are predominantly associated with carbohydrate transport at the plasma membrane. Similarly, N-rich domains are strongly associated with transcription-related processes, whereas Q-rich domains are more strongly associated with endocytosis and other processes in the cytoplasm. While there is some overlap between these two groups, this suggests that domains enriched in remarkably similar amino acids may yet be favored for specialized roles in the cell.

Finally, a bottom-up application of our composition-centric algorithm to membraneless organelles provides the first step in defining the distinct compositional profiles associated with each type of organelle. We find that even closely related and physically interacting organelles are associated with discernible differences in compositional enrichment, which may relate to differences in their properties, regulation, and function *in vivo*. It is important to note that, while the observed trends in compositional enrichment are significantly associated with stress granule proteins or P-body proteins as respective groups, these features may not be absolutely required for individual proteins to be incorporated into stress granules and/or P-bodies. It is possible, for example, that two proteins possessing non-overlapping subsets of the associated compositional features may still be recruited to stress granules, and that some stress granule proteins may be recruited for reasons entirely distinct from compositional enrichment (e.g. via RNA-binding domains). One might even imagine that differences in compositional features, while still allowing recruitment to stress granules and/or P-bodies, could favor differences in the dynamics of individual protein components (e.g. the kinetics of entry/exit, dwell time, the strength of the interactions, or the depth of penetration within the stress granule/P-body). Therefore, while the associated composition ranges observed here are collectively enriched among proteins associated with these membraneless organelles, each individual protein need not possess all of the compositional features simultaneously in order to function as a stress granule or P-body protein.

While a great deal of attention is rightfully devoted to understanding the effects of primary amino acid sequence on protein fates (including folding, regulation, and functions), amino

acid composition is increasingly believed to drive a variety of cellular and molecular processes. Here, we have developed an approach to examine relationships between local compositional enrichment and protein fates for each of the canonical amino acids, in the absence of *a priori* assumptions or pre-defined thresholds. Our results provide a coherent, proteome-wide view of the relationships between compositional enrichment and the fundamental aspects of protein life cycle, subcellular localization, and function in model eukaryotic organisms.

Methods

Composition-based proteome scanning algorithm

Protein sequences were parsed using FASTA sequence parsing module from the Biopython package [70]. For each amino acid in the set of 20 canonical amino acids, each protein in the translated ORF proteome (latest release from the Saccharomyces Genome Database website, last modified 13-Jan-2015) or the ORF coding sequences (organismID:UP000001940_6239, release date 23-May-2018 downloaded from the UniProt website) was scanned using a sliding window of defined size (ranging from 10 to 100 amino acids, in increments of 10). The percent composition of the amino acid of interest (AAoI) is calculated for each window, and the protein is sorted into bins based on the maximum percent composition achieved for the AAoI (ranging from 0 to 100 percent composition in 5 percent increments). Analyses were performed for all possible AAoI, window size, and percent composition combinations.

Normalized translation efficiency (nTE)

Translation efficiency for each gene was estimated using the normalized translation efficiency (nTE) scale [55], which is based on tRNA gene copy number, codon-anticodon wobble base-pairing efficiency, and transcriptome-wide codon usage. However, the original nTE algorithm plots all nTE values for each codon to generate a separate translation efficiency profile for each gene. In order to condense translation efficiency information to a single value for each gene (in a manner analogous to the tRNA adaptation index; [60]), the geometric mean of nTE values across the transcript was calculated as

$$nTE_{gene} = \left(\prod_{k=1}^{l_s} nTE_{i_{k_s}} \right)^{\frac{1}{l_s}} \quad (1)$$

where nTE_{iks} represents the translation efficiency value of the ith codon defined by the kth triplet in nucleotide sequence s, and l_s represents the length of the nucleotide sequence excluding stop codons. Therefore, nTE values reported in the current study represent whole-gene nTE values. nTE analyses were performed using an in-house Python script.

Defining Shannon entropy, statistical amino acid bias, and homopolymeric repeats

The Shannon entropy of each sequence was calculated as

$$SE = -\sum_{i=1}^{N=20} \frac{n_i}{L} \left(log_2 \frac{n_i}{L} \right) \quad (2)$$

where N represents the size of the residue alphabet ($N = 20$, for the canonical amino acids), n_i represents the number of occurrences of the ith residue within the given sequence window of length L. For comparison with established measures of sequence complexity, we defined low-complexity domains by using the default window size (12 amino acids) and Shannon entropy

threshold (SE $\leq$ 2.2bits) used in the first pass of the SEG algorithm to initially identify LCDs [1,10].

In the SEG algorithm, the complexity state vector used to calculate the Shannon entropy is blind to the amino acid composition (i.e. the n_i values in Eq 2 are not attributed their respective amino acids). Therefore, when indicated, in order to distinguish LCDs on the basis of the predominant amino acid, sequences for which the SE $\leq$ 2.2bits and $n_{AAoI} \geq n_{max}$ within the complexity state (indicating that the AAoI is a major contributor to the sequence's classification as an LCD) were assigned to the corresponding amino acid category (e.g. A-rich LCDs, C-rich LCDs, etc.). Single-LCD/CED proteins are proteins classified as LCDs or CEDs that do not appear on multiple amino acid-specific LCD/CED lists.

Statistical amino acid bias was calculated as described in [12]. Briefly, the lowest probability subsequence for each protein was determined by exhaustively scanning proteins with window sizes ranging from 25 to 2500 amino acids. For each window, the subsequence bias probability (P_{bias}) was defined as

$$P_{bias} = \left[\frac{w!}{n!(w-n)!}\right] \times (f_x)^n \times (1-f_x)^{w-n} \quad (3)$$

where w denotes the window size, n denotes the number of occurrences of the amino acid of interest within the subsequence, and f_x denotes the fraction of the amino acid of interest in the yeast translated proteome. The lowest probability subsequence for each protein is the subsequence with the lowest P_{bias}.

A suitable threshold to define statistically-biased proteins within the yeast protein was determined as previously described [12], except that more relaxed criteria were used in order to include additional proteins with less extreme biases. Briefly, the P_{bias} corresponding to the lowest probability subsequence (P_{min}) for each protein was plotted on a log-log plot against whole-protein sequence length. A line was fitted, then the y-intercept was decreased until only 15% of the proteome had P_{min} values below the line (previous analyses used a more stringent cutoff of 10% to define statistically-biased proteins [12]). Additionally, a length-independent threshold was designated as the P_{min} value at which 15% of the proteome had smaller absolute P_{min} values. This threshold was used when it was less than the P_{min} threshold given by the length-dependent method to avoid unreasonably relaxed bias criteria for small protein sequences. Amino acid bias was calculated using values from the translated orf proteome only, and implemented via an in-house Python script with pre-computed look-up tables for computational efficiency.

Proteins containing homopolymeric sequences were defined simply as any protein with a subsequence of five or more contiguous residues of the same amino acid, as previously described [37].

Protein abundance and protein half-life data

Yeast protein abundance values (in average number of molecules per cell per protein) were obtained from [48] (n = 5,391). Protein abundance values for *C. elegans* were obtained from [59]. Yeast protein half-life data were obtained from [49]. For simplicity of interpretation, only proteins with unambiguous, non-zero half-life or abundance values were included in the datasets. Proteins listed on separate lines with identical half-life or abundance values were retained, whereas protein half-life or abundance measurements assigned to more than one protein on the same line were excluded (these were often highly homologous genes, suggesting that the measurement could not be unambiguously assigned to one of the proteins). Furthermore, all proteins corresponding to "low-confidence" measurements in the half-life dataset were

excluded (see [49] for criteria). n = 3,525 for the filtered yeast half-life dataset, and n = 5,952 for the filtered *C. elegans* protein abundance dataset.

Statistics and plotting

For all AAoI/window size/percent composition bins, the distribution of nTE, abundance, or half-life values for proteins included in the given bin was compared to the distribution of the respective values of all proteins excluded from the given bin. Statistical significance was estimated using a two-sided Mann-Whitney *U* test (also referred to as the Wilcoxon rank-sum test; refer to Supplemental Experimental Procedures from [47] for a detailed description and rationale). Where indicated, *p*-values were adjusted within each window using the Bonferroni correction method for multiple hypothesis testing. All statistical tests were performed using modules available in the SciPy package with default settings, unless otherwise specified. All plots were generated using Matplotlib or Seaborn modules.

Gene ontology (GO) term enrichment analyses

GO term enrichment tests were performed using the GOATOOLS package (version 0.7.9) [71] for each set of proteins contained in a given amino acid/window size/percent composition bin. For each test, the set of background proteins was defined as all proteins from the translated ORF proteome of sequence length greater than or equal to the given window size. All reported *p*-values were adjusted using the Bonferroni correction during GO term association. To evaluate the compositional enrichment profiles associated with GO terms related to subcellular compartments, we applied a minimum-threshold-scanning approach to all partitioned proteomes. For each AAoI, window size, and percent composition bin, all proteins with maximum local compositions greater than or equal to the current percent composition under consideration are pooled and evaluated for possible enriched GO terms. This effectively evaluates possible GO term enrichment iteratively with increasing maximum local composition criteria. GO term results were subsequently evaluated for significant enrichment of a single GO term describing each subcellular compartment (or two related GO terms, "outer membrane" and "plasma membrane", in the case of the plasma membrane). *p*-values were further adjusted within each window size using the Bonferroni correction method.

Similar analyses were performed for the sets of experimentally-defined stress granule (n = 83) and P-body (n = 52) proteins [66]. Specifically, a minimum-threshold-scanning approach was applied to all partitioned proteomes. For each AAoI, window size, and percent composition bin, all proteins with maximum local compositions greater than or equal to the current percent composition under consideration are pooled. Significant enrichment of experimentally-defined stress granule or P-body proteins within each pool of proteins was evaluated using Fisher's exact test ($p < 0.05$).

Supporting information

S1 Table. Maximum local composition values for each amino acid and window size combination for all translated yeast proteins. For each protein in the translated ORF proteome, nTE, protein abundance, and protein half-life values are indicated. nTE was calculated according to the method described in [55]. Protein abundance and protein half-life values were reported in [48] and [54], respectively. Remaining columns contain the maximum local composition value (per 100) for each amino acid and window size combination for each protein. (CSV)

S2 Table. Maximum local composition values for each amino acid and window size combination for all *C. elegans* proteins. For each protein in the *C. elegans* proteome, protein abundance values (mean measured intensity across 3 biological replicates from [59]; see Methods section for inclusion criteria) are indicated. Remaining columns contain the maximum local composition value (per 100) for each amino acid and window size combination for each protein.
(CSV)

S3 Table. Residue-specific CEDs are associated with unique cellular structures and processes. All GO terms listed represent terms significantly associated with a set of residue-specific CEDs for at least one window size (Bonferroni-corrected $p \leq 0.05$).
(XLSX)

S1 Fig. Maximum local amino acid composition corresponds to residue-dependent differences in protein synthesis efficiency. Local enrichment for individual amino acids correspond to composition-dependent changes in experimentally-derived protein synthesis efficiency [56].
(TIF)

S2 Fig. Distribution of the *C. elegans* proteome based on maximum local amino acid composition. The number of proteins partitioned into each window size/percent composition bin for each of the 20 canonical amino acids are plotted as a function of maximum local composition for each window size. Scatter points are connected by line segments for visual clarity only.
(TIF)

S3 Fig. Associations between local compositional enrichment and nTE persist in the absence of proteins with statistically-biased domains. For each amino acid, nTE values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length $\geq$ the corresponding window size that were excluded from the bin. Red and blue points indicate bins for which the distribution of protein half-life values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicated bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.
(TIF)

S4 Fig. Associations between local compositional enrichment and protein abundance persist in the absence of proteins with statistically-biased domains. For each amino acid, protein abundance values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length $\geq$ the corresponding window size that were excluded from the bin. Red and blue points indicate bins for which the distribution of protein half-life values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicated bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.
(TIF)

S5 Fig. Associations between local compositional enrichment and protein half-life persist in the absence of proteins with statistically-biased domains. For each amino acid, protein half-life values corresponding to proteins partitioned into a given window size and percent

composition bin were compared to values for all proteins of length ≥ the corresponding window size that were excluded from the bin. Red and blue points indicate bins for which the distribution of protein half-life values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicated bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.
(TIF)

S6 Fig. Associations between local compositional enrichment and nTE persist in the absence of LCD-containing proteins. For each amino acid, nTE values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length ≥ the corresponding window size that were excluded from the bin. Red and blue points indicate bins for which the distribution of protein half-life values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicated bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.
(TIF)

S7 Fig. Associations between local compositional enrichment and protein abundance persist in the absence of LCD-containing proteins. For each amino acid, protein abundance values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length ≥ the corresponding window size that were excluded from the bin. Red and blue points indicate bins for which the distribution of protein half-life values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicated bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.
(TIF)

S8 Fig. Associations between local compositional enrichment and protein half-life persist in the absence of LCD-containing proteins. For each amino acid, protein half-life values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length ≥ the corresponding window size that were excluded from the bin. Red and blue points indicate bins for which the distribution of protein half-life values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicated bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.
(TIF)

S9 Fig. Associations between local compositional enrichment and nTE persist in the absence of proteins with homopolymeric repeats. For each amino acid, nTE values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length ≥ the corresponding window size that were excluded from the bin. Red and blue points indicate bins for which the distribution of protein half-life values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded

proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicated bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.
(TIF)

S10 Fig. Associations between local compositional enrichment and protein abundance persist in the absence of proteins with homopolymeric repeats. For each amino acid, protein abundance values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length $\geq$ the corresponding window size that were excluded from the bin. Red and blue points indicate bins for which the distribution of protein half-life values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicated bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.
(TIF)

S11 Fig. Associations between local compositional enrichment and protein half-life persist in the absence of proteins with homopolymeric repeats. For each amino acid, protein half-life values corresponding to proteins partitioned into a given window size and percent composition bin were compared to values for all proteins of length $\geq$ the corresponding window size that were excluded from the bin. Red and blue points indicate bins for which the distribution of protein half-life values differ significantly (Bonferroni-corrected $p \leq 0.05$) from those of excluded proteins: red points indicate bins with a lower median value relative to that of excluded proteins, whereas blue points indicated bins with a higher relative median value. Grey points indicate comparisons lacking statistical significance. Individual points are scaled within each subplot to reflect the sample sizes of proteins contained within each bin.
(TIF)

S12 Fig. Individual amino acid composition profiles for subcellular compartments predominantly associated with enrichment for polar and charged residues. Composition ranges for each amino acid significantly associated with the cytoplasm (A) and nucleus (B) are indicated. All plotted points indicate protein sets for which association with the indicated subcellular compartment is statistically significant (Bonferonni-corrected $p < 0.05$). Plots are shown only for amino acids with at least two composition bins significantly associated with the indicated subcellular compartment.
(TIF)

S13 Fig. Individual amino acid composition profiles for subcellular compartments predominantly associated with enrichment for hydrophobic and aromatic residues. Composition ranges for each amino acid significantly associated with the endoplasmic reticulum (A), Golgi apparatus (B), vacuole (C), and mitochondria (D) are indicated. All plotted points indicate protein sets for which association with the indicated subcellular compartment is statistically significant (Bonferonni-corrected $p < 0.05$). Plots are shown only for amino acids with at least two composition bins significantly associated with the indicated subcellular compartment.
(TIF)

S14 Fig. Individual amino acid composition profiles for subcellular compartments predominantly associated with enrichment for both polar and hydrophobic/aromatic

residues. Composition ranges for each amino acid significantly associated with the plasma membrane (A) and cell wall (B) are indicated. All plotted points indicate protein sets for which association with the indicated subcellular compartment is statistically significant (Bonferonni-corrected $p < 0.05$). Plots are shown only for amino acids with at least two composition bins significantly associated with the indicated subcellular compartment.
(TIF)

Author Contributions

Conceptualization: Sean M. Cascarina, Eric D. Ross.

Data curation: Sean M. Cascarina.

Formal analysis: Sean M. Cascarina.

Funding acquisition: Eric D. Ross.

Investigation: Sean M. Cascarina.

Methodology: Sean M. Cascarina.

Project administration: Sean M. Cascarina.

Resources: Eric D. Ross.

Software: Sean M. Cascarina.

Supervision: Sean M. Cascarina, Eric D. Ross.

Validation: Sean M. Cascarina.

Visualization: Sean M. Cascarina.

Writing – original draft: Sean M. Cascarina.

Writing – review & editing: Sean M. Cascarina, Eric D. Ross.

References

1. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem. 1993; 17: 149–163. https://doi.org/10.1016/0097-8485(93)85006-X
2. Sim KL, Creamer TP. Abundance and Distributions of Eukaryote Protein Simple Sequences. Mol Cell Proteomics. 2002; 1: 983–995. https://doi.org/10.1074/mcp.M200032-MCP200 PMID: 12543934
3. DePristo MA, Zilversmit MM, Hartl DL. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. Gene. 2006; 378: 19–30. https://doi.org/10.1016/j.gene.2006.03.023 PMID: 16806741
4. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. J Mol Biol. 1999; 293: 151–160. https://doi.org/10.1006/jmbi.1999.3136 PMID: 10512723
5. Shin SW, Kim SM. A new algorithm for detecting low-complexity regions in protein sequences. Bioinformatics. 2005; 21: 160–170. https://doi.org/10.1093/bioinformatics/bth497 PMID: 15333459
6. Claverie J-M, States DJ. Information Enhancement Methods Analysis * for. 1993; 17: 191–201.
7. Li X, Kahveci T. A novel algorithm for identifying low-complexity regions in a protein sequence. Bioinformatics. 2006; 22: 2980–2987. https://doi.org/10.1093/bioinformatics/btl495 PMID: 17018537
8. Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, et al. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Bioinformatics. 2000; 16: 915–922. https://doi.org/10.1093/bioinformatics/16.10.915 PMID: 11120681
9. Nandi T, Dash D, Ghai R, C BR, Kannan K, Brahmachari SK, et al. A novel complexity measure for comparative analysis of protein sequences from complete genomes. J Biomol Struct Dyn. 2003; 20: 657–668. doi:d=3013&c=4104&p=11631&do=detail [pii]\n https://doi.org/10.1080/07391102.2003.10506882 PMID: 12643768

10. Wootton JC, Federhen S. Analysis of Compositionally Biased Regions in Sequence Databases. Methods Enzymol. 1996; 266: 554–571. PMID: 8743706

11. Harrison PM, Gerstein M. A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes. Genome Biol. 2003; 4: R40. https://doi.org/10.1186/gb-2003-4-6-r40 PMID: 12801414

12. Harrison PM. Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila. BMC Bioinformatics. 2006; 7: 441. https://doi.org/10.1186/1471-2105-7-441 PMID: 17032452

13. Toll-Riera M, Radó-Trilla N, Martys F, Albà MM. Role of low-complexity sequences in the formation of novel protein coding sequences. Mol Biol Evol. 2011; 1–12. https://doi.org/10.1093/molbev/msr263 PMID: 22045997

14. Radó-Trilla N, Albà Mm. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evol Biol. 2012; 12: 155. https://doi.org/10.1186/1471-2148-12-155 PMID: 22920595

15. Kumari B, Kumar R, Kumar M. Low complexity and disordered regions of proteins have different structural and amino acid preferences. Mol Biosyst. 2015; 11: 585–594. https://doi.org/10.1039/c4mb00425f PMID: 25468592

16. Coletta A, Pinney JW, Solís D, Marsh J, Pettifer SR, Attwood TK. Low-complexity regions within protein sequences have position-dependent roles. BMC Syst Biol. 2010; 4: 43. https://doi.org/10.1186/1752-0509-4-43 PMID: 20385029

17. Radó-Trilla N, Arató K, Pegueroles C, Raya A, de la Luna S, Albà MM. Key Role of Amino Acid Repeat Expansions in the Functional Diversification of Duplicated Transcription Factors. Mol Biol Evol. 2015; 32: 2263–72. https://doi.org/10.1093/molbev/msv103 PMID: 25931513

18. Michelitsch MD, Weissman JS. A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. Proc Natl Acad Sci USA. 2000; 97: 11910–5. Available: http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=11050225 https://doi.org/10.1073/pnas.97.22.11910 PMID: 11050225

19. Cascarina SM, Ross ED. Yeast prions and human prion-like proteins: Sequence features and prediction methods. Cell Mol Life Sci. 2014; 71: 2047–2063. https://doi.org/10.1007/s00018-013-1543-6 PMID: 24390581

20. Alberti S, Halfmann R, King O, Kapila A, Lindquist S. A Systematic Survey Identifies Prions and Illuminates Sequence Features of Prionogenic Proteins. Cell. Elsevier Ltd; 2009; 137: 146–158. https://doi.org/10.1016/j.cell.2009.02.044 PMID: 19345193

21. King OD, Gitler AD, Shorter J. The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. Brain Res. Elsevier B.V.; 2012; 1462: 61–80. https://doi.org/10.1016/j.brainres.2012.01.016 PMID: 22445064

22. Galzitskaya O V. Repeats are one of the main characteristics of RNA-binding proteins with prion-like domains. Mol Biosyst. Royal Society of Chemistry; 2015; 11: 2210–2218. https://doi.org/10.1039/c5mb00273g PMID: 26022110

23. Harrison AF, Shorter J. RNA-binding proteins with prion-like domains in health and disease. Biochem J. 2017; 474: 1417–1438. https://doi.org/10.1042/BCJ20160499 PMID: 28389532

24. Das RK, Ruff KM, Pappu R V. Relating sequence encoded information to form and function of intrinsically disordered proteins. Curr Opin Struct Biol. 2015; 32: 102–112. https://doi.org/10.1016/j.sbi.2015.03.008 PMID: 25863585

25. Kroschwald S, Maharana S, Mateju D, Malinovska L, Nüske E, Poser I, et al. Promiscuous interactions and protein disaggregases determine the material state of stress-inducible RNP granules. Elife. 2015; 4. https://doi.org/10.7554/eLife.06807 PMID: 26238190

26. Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, et al. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. Cell. 2015; 162: 1066–1077. https://doi.org/10.1016/j.cell.2015.07.047 PMID: 26317470

27. Lin Y, Protter DSW, Rosen MK, Parker R. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. Mol Cell. 2015; 60: 208–219. https://doi.org/10.1016/j.molcel.2015.08.018 PMID: 26412307

28. Xiang S, Kato M, Wu LC, Lin Y, Ding M, Zhang Y, et al. The LC Domain of hnRNPA2 Adopts Similar Conformations in Hydrogel Polymers, Liquid-like Droplets, and Nuclei. Cell. Elsevier Inc.; 2015; 163: 829–839. https://doi.org/10.1016/j.cell.2015.10.040 PMID: 26544936

29. Kato M, Han TW, Xie S, Shi K, Du X, Wu LC, et al. Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels. Cell. 2012; 149: 753–767. https://doi.org/10.1016/j.cell.2012.04.017 PMID: 22579281

30. Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, et al. Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. Cell. 2015; 163: 123–133. https://doi.org/10.1016/j.cell.2015.09.015 PMID: 26406374

31. Weber JJ, Sowa AS, Binder T, Hübener J. From pathways to targets: Understanding the mechanisms behind polyglutamine disease. BioMed Research International. 2014. https://doi.org/10.1155/2014/701758 PMID: 25309920

32. Kim HJ, Kim NC, Wang Y-D, Scarborough EA, Moore J, Diaz Z, et al. Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. Nature. Nature Publishing Group; 2013; 495: 467–473. https://doi.org/10.1038/nature11922 PMID: 23455423

33. Mackenzie IR, Nicholson AM, Sarkar M, Messing J, Purice MD, Pottier C, et al. TIA1 Mutations in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Promote Phase Separation and Alter Stress Granule Dynamics. Neuron. 2017; 95: 808–816.e9. https://doi.org/10.1016/j.neuron.2017.07.025 PMID: 28817800

34. Lobanov MY, Klus P, Sokolovsky IV, Tartaglia GG, Galzitskaya OV. Non-random distribution of homorepeats: Links with biological functions and human diseases. Sci Rep. 2016; 6. https://doi.org/10.1038/srep26941 PMID: 27256590

35. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci U S A. 2002; 99: 333–338. https://doi.org/10.1073/pnas.012608599 PMID: 11782551

36. Albà MM, Guigó R. Comparative analysis of amino acid repeats in rodents and humans. Genome Res. 2004; 14: 549–554. https://doi.org/10.1101/gr.1925704 PMID: 15059995

37. Chavali S, Chavali PL, Chalancon G, de Groot NS, Gemayel R, Latysheva NS, et al. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. Nat Struct Mol Biol. 2017; https://doi.org/10.1038/nsmb.3441 PMID: 28805808

38. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, De La Banda MG, et al. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res. 2005; 15: 537–551. https://doi.org/10.1101/gr.3096505 PMID: 15805494

39. Kumar AS, Sowpati DT, Mishra RK. Single amino acid repeats in the proteome world: Structural, functional, and evolutionary insights. PLoS One. 2016; 11. https://doi.org/10.1371/journal.pone.0166854 PMID: 27893794

40. Simon M, Hancock JM. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol. 2009; 10. https://doi.org/10.1186/gb-2009-10-6-r59 PMID: 19486509

41. Lobanov MI, Bogatyreva NS, Galzitskaia O V. [Occurrence of motifs with six amino acid residues in three eukaryotic proteomes]. Mol Biol (Mosk). 2012; 46: 184–190.

42. Martin EW, Mittag T. The relationship of sequence and phase separation in protein low-complexity regions. Biochemistry. 2018; acs.biochem.8b00008. https://doi.org/10.1021/acs.biochem.8b00008 PMID: 29517898

43. Toll-Riera M, Radó-Trilla N, Martys F, Albá MM. Role of low-complexity sequences in the formation of novel protein coding sequences. Mol Biol Evol. 2012; 29: 883–886. https://doi.org/10.1093/molbev/msr263 PMID: 22045997

44. Galzitskaya O V., Lobanov MY. Phyloproteomic analysis of 11780 six-residue-long motifs occurrences. Biomed Res Int. 2015;2015. https://doi.org/10.1155/2015/208346 PMID: 26114101

45. Lobanov MY, Galzitskaya O V. Disordered patterns in clustered protein data bank and in Eukaryotic and bacterial proteomes. PLoS One. 2011; 6. https://doi.org/10.1371/journal.pone.0027142 PMID: 22073276

46. Lobanov MY, Galzitskaya O V. Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes. Mol Biosyst. 2012; 8: 327–337. https://doi.org/10.1039/c1mb05318c PMID: 22009164

47. van der Lee R, Lang B, Kruse K, Gsponer J, de Groot NS, Huynen MA, et al. Intrinsically disordered segments affect protein half-life in the cell and during evolution. Cell Rep. 2014; 8: 1832–1844. https://doi.org/10.1016/j.celrep.2014.07.055 PMID: 25220455

48. Ho B, Baryshnikova A, Brown GW. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. Cell Syst. 2018; 6: 192–205.e3. https://doi.org/10.1016/j.cels.2017.12.004 PMID: 29361465

49. Christiano R, Nagaraj N, Fröhlich F, Walther TC. Global Proteome Turnover Analyses of the Yeasts S. cerevisiae and S.pombe. Cell Rep. 2014; 9: 1959–1966. https://doi.org/10.1016/j.celrep.2014.10.065 PMID: 25466257

50. Tuller T, Kupiec M, Ruppin E. Determinants of protein abundance and translation efficiency in S. cerevisiae. PLoS Comput Biol. 2007; 3: 2510–2519. https://doi.org/10.1371/journal.pcbi.0030248 PMID: 18159940

51. Greenbaum D, Jansen R, Gerstein M. Analysis of mRNA expression and protein abundance data: An approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. Bioinformatics. 2002; 18: 585–596. https://doi.org/10.1093/bioinformatics/18.4.585 PMID: 12016056

52. Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK. Quantification of protein half-lives in the budding yeast proteome. Proc Natl Acad Sci U S A. 2006; 103: 13004–9. https://doi.org/10.1073/pnas.0605420103 PMID: 16916930

53. Martin-Perez M, Villén J. Determinants and Regulation of Protein Turnover in Yeast. Cell Syst. 2017; 5: 283–294.e5. https://doi.org/10.1016/j.cels.2017.08.008 PMID: 28918244

54. Christiano R, Nagaraj N, Fröhlich F, Walther TC. Global Proteome Turnover Analyses of the Yeasts S. cerevisiae and S.pombe. Cell Rep. 2014; 9: 1959–1966. https://doi.org/10.1016/j.celrep.2014.10.065 PMID: 25466257

55. Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol. 2012; 20: 237–243. https://doi.org/10.1038/nsmb.2466 PMID: 23262490

56. Lahtvee PJ, Sánchez BJ, Smialowska A, Kasvandik S, Elsemman IE, Gatto F, et al. Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. Cell Syst. 2017; 4: 495–504.e5. https://doi.org/10.1016/j.cels.2017.03.003 PMID: 28365149

57. Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci. 2009; 106: 3264–3269. https://doi.org/10.1073/pnas.0812841106 PMID: 19208812

58. Schreiber K, Csaba G, Haslbeck M, Zimmer R. Alternative splicing in next generation sequencing data of saccharomyces cerevisiae. PLoS One. 2015; 10. https://doi.org/10.1371/journal.pone.0140487 PMID: 26469855

59. Narayan V, Ly T, Pourkarimi E, Murillo AB, Gartner A, Lamond AI, et al. Deep Proteome Analysis Identifies Age-Related Processes in C. elegans. Cell Syst. 2016; 3: 144–159. https://doi.org/10.1016/j.cels.2016.06.011 PMID: 27453442

60. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: A test for translational selection. Nucleic Acids Res. 2004; 32: 5036–5044. https://doi.org/10.1093/nar/gkh834 PMID: 15448185

61. Pe'er I, Felder CE, Man O, Silman I, Sussman JL, Beckmann JS. Proteomic Signatures: Amino Acid and Oligopeptide Compositions Differentiate among Phyla. Proteins Struct Funct Genet. 2004; https://doi.org/10.1002/prot.10559 PMID: 14705021

62. Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, Plochowietz A, et al. Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. Mol Cell. 2015; 57: 936–947. https://doi.org/10.1016/j.molcel.2015.01.013 PMID: 25747659

63. Pak CW, Kosno M, Holehouse AS, Padrick SB, Mittal A, Ali R, et al. Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. Mol Cell. 2016; 63: 72–85. https://doi.org/10.1016/j.molcel.2016.05.042 PMID: 27392146

64. Altmeyer M, Neelsen KJ, Teloni F, Pozdnyakova I, Pellegrino S, Grøfte M, et al. Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). Nat Commun. 2015; 6: 8088. https://doi.org/10.1038/ncomms9088 PMID: 26286827

65. Sontag EM, Samant RS, Frydman J. Mechanisms and Functions of Spatial Protein Quality Control. Annu Rev Biochem. 2017; 86: 97–122. https://doi.org/10.1146/annurev-biochem-060815-014616 PMID: 28489421

66. Jain S, Wheeler JR, Walters RW, Agrawal A, Barsic A, Parker R. ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. Cell. 2016; 164: 487–498. https://doi.org/10.1016/j.cell.2015.12.038 PMID: 26777405

67. Buchan JR, Muhlrad D, Parker R. P bodies promote stress granule assembly in Saccharomyces cerevisiae. J Cell Biol. 2008; 183: 441–455. https://doi.org/10.1083/jcb.200807043 PMID: 18981231

68. Kedersha N, Stoecklin G, Ayodele M, Yacono P, Lykke-Andersen J, Fitzler MJ, et al. Stress granules and processing bodies are dynamically linked sites of mRNP remodeling. J Cell Biol. 2005; 169: 871–884. https://doi.org/10.1083/jcb.200502088 PMID: 15967811

69. Wang J, Choi J-M, Holehouse AS, Lee HO, Zhang X, Jahnel M, et al. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell. 2018; https://doi.org/10.1016/j.cell.2018.06.006 PMID: 29961577

70. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25: 1422–1423. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

71. Tang H, Klopfenstein D, Pedersen B, Flick P, Sato K, Ramirez F, et al. GOATOOLS: Tools for Gene Ontology. 2015; https://doi.org/10.5281/ZENODO.31628