Homopeptide and Homocodon Levels are Coupled to GC/AT Bias Levels, Intrinsic Disorder Propensity and other Factors Across Diverse Fungi

Yue Wang
McGill University
Paul Harrison (paul.harrison@mcgill.ca)
McGill University

Research Article

Keywords: Homopeptides , proteome evolution , DNA , homopeptide frequencies

Posted Date: December 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-118390/v1

Abstract

Homopeptides (consecutive runs of one amino-acid type) are suggested to play important roles in proteome evolution, since they are prone to expand/contract during DNA replication, recombination and repair. It is currently not clear how homopeptide frequencies vary as organisms evolve, and which genomic/proteomic traits drive variation. Thus, to gain insight, we analyzed how homopeptides and homocodons (which are pure codon repeats) vary across 405 Dikarya, and probed how this variation is linked to GC/AT bias amongst other factors. We observe that amino-acid homopeptide frequencies vary diversely between clades (even close relatives), with the AT-rich Saccharomycotina trending distinctly. As organisms evolve, homocodon and homopeptide numbers are majorly coupled to GC/AT-bias, with medium GC/AT genomes having markedly fewer. Despite this, homopeptides tend to be more GC-rich than other proteome areas, even in AT-rich organisms, indicating they absorb AT bias less or are inherently more GC-rich. Furthermore, the purity of homopeptides (i.e., the degree one codon type predominates in them) varies least for amino acids with GC/AT-balanced codon repertoires, with most variation for arginine since it has only one AT-rich codon (out of six). The most frequent and most variable homopeptide amino acids have greater intrinsic disorder propensity, and annotated intrinsic disorder fractions are strongly correlated with homopeptide levels (unlike structured domain fractions, which are anti-correlated). Poly-glutamine uniquely behaves as an evolutionarily very variable homopeptide with a codon repertoire unbiased for GC/AT. In summary, homopeptide/homocodon levels are coupled to or influenced by several factors, including GC/AT bias and amino-acid intrinsic disorder propensity.

Introduction

Homopeptides and homocodons (which are perfect codon repeats) are well known for their roles in inherited human diseases, such as poly-CAG/poly-Gln in Huntington's disease, and poly-Ala linked to congenital developmental disorders [1]. The pathogenic mechanisms of these diseases are various. While many diseases might be essentially caused by the aggregation propensity of certain types of homopeptides [2,3], the soluble forms of proteins with longer mutant repeats could also be problematic by competing with functional homopeptides in normal proteins for molecular interactions [4]. Homopeptides and homocodons not exceeding certain length thresholds are prevalent and can be beneficial for eukaryotes [5]. About 15% of proteins in any eukaryotic proteome contain at least one stretch of ≥5 identical residues [6]. These homopeptide-containing proteins function diversely, especially in DNA/RNA binding, signaling and regulation [7-9]. Homopeptides levels generally exceed those of other amino-acid repeat types [10].

Nevertheless, the functions of widely prevalent homopeptides or homocodons are still largely unclear, and most of them might not have essential roles but rather create diversities quickly in genomes which can be selected on during evolution [11]. Homopeptide length polymorphisms are frequently found in different individuals of a species, and even between different cell types or at different ages of an individual [12,13].

Although phenotypic evolution is mostly modulated by cis-regulatory elements, homopeptide length polymorphisms are also found linked to significant morphological differences, *e.g.,* in dogs [14]. Opposed to the binary effect of single nucleotide polymorphisms, homopeptides length variations are proposed as a “digital modulator” or “tuning knob” that acts through expansion and contraction between generations, leading to greater phenotypic variability in a population [11]. Besides the high mutation rate of homopeptides themselves, DNA substitution rate is also strongly correlated with the distance to homopeptides and insertion/deletion mutations are frequently associated with homopeptides in their flanks [15,16]. Thus, homopeptides may drive rapid divergence of proteins that contain them, through creating more polymorphism.

Early studies found that eukaryotes have unique homopeptide distributions, *i.e.,* their proteomes prefer/tolerate homopeptides at different lengths for different amino acids [17]. For example, poly-Ser, -Ala, -Glu, and -Gly ≥10 residues long are preferred in the *Drosophila melanogaster* proteome, while poly-Asn and -Asp at all lengths and poly-Ser ≥20 residues are preferred in *Saccharomyces cerevisiae*, which is less tolerant of poly-Gly, -Arg, -Pro [17]. It was suggested that amino-acid preferences in low-complexity regions or homopeptides are largely driven by genome AT/GC bias, and are under selection pressures [16,18]. Also, homopeptides are prone to accumulate in intrinsically disordered regions (IDRs) [10,19,20].

Previously, it was observed that a large-scale emergence of prion-like regions during *Saccharomycetes* yeast evolution was caused by mutational trends that produced more poly-asparagine tracts [21]. Motivated by these findings, we hypothesized that the factors driving the evolution of homopeptides and homocodons in general would also be discernible through analysis of their trends across a large diverse fungal clade, *i.e.,* the subkingdom *Dikarya*. We discovered that, over hundreds of millions of years of fungal evolution, homopeptide and homocodon accumulation is coupled to or modulated by GC/AT bias, intrinsic disorder propensity and other factors, such as the inherent design of the genetic code.

Results And Discussion

The evolutionary behaviour of homopeptides and homocodons (perfect codon repeats) is surveyed across the fungal Dikarya sub-kingdom. In this survey, we had the following objectives:

(i) To derive an overview of the variation in homopeptide frequencies, identifying any anomalous behaviour in specific clades;

(ii) To examine how homopeptide frequencies are influenced by or coupled to genomic AT/GC bias, which is the most basic compositional parameter typically studied in such analyses;

(iii) To examine how codon preferences in homocodons and homopeptides are affected by such AT/GC bias, in doing so deriving a measure of homopeptide purity (i.e., the predominance of one specific codon in homopeptides);

(iv) To examine how proteomic homopeptide frequencies are influenced by intrinsic disorder and structured domain content in proteins.

Homopeptide levels vary extensively across diverse fungi

The distribution of homopeptide frequencies (1.64–4.78%) in the 405 proteomes of Dikarya shows a heavy-tailed right-skewed distribution. Nearly 70% of the distribution is in the small range 1.84–2.44%. Only a few proteomes have homopeptide frequencies below this range, the rest varying from 2.44 % to 4.78% (Figure 1). This shows that while most proteomes have similar homopeptide fractions, there is a bias towards homopeptide accumulation for values away from this peak.

We examined the trends in homopeptide frequencies across 405 Dikarya, and also examined other various attributes, including GC content (Figure 2 and Suppl. Figure S1). Heat maps of the most abundant homopeptides and homocodons (i.e., perfect codon repeats) were derived (Suppl. Figure S1). Subphyla (and classes within the large subphylum Pezizomycotina) are analyzed in Figure 2, with details of species names and prevalent amino-acid / codon types in Suppl. Figure S1. The lowest homopeptide fractions are for Saccharomycotina and Taphrinomycotina, which are also low-GC and have the lowest fractions of annotated IDRs (Figure 2). Obvious variations of homopeptide fractions tend to appear between different clades, but homopeptides can also accumulate in individual species within a short evolutionary time (Suppl. Figure S1). To show which homopeptides and homocodons predominate, they are ranked in decreasing order of overall frequency (i.e., total fraction of amino acids or codons of that type) in each proteome. Homopeptide and homocodon length distributions are characterised using slopes from log-log plots as described in Materials and Methods. For these length distributions, darker colours in heat map cells indicate more, long homopeptides or homocodons of a specific type. One can see that generally there are more lighter cells for bands c and d (shorter homopeptides and homocodons) where the overall homopeptide fraction is lower (darker in band a) (Suppl. Figure S1). Indeed, when we examine the relationship between the log(length) distribution slopes and corresponding homopeptide frequencies for each amino acid, we see that most exhibit correlations, some highly significantly, most notably glycine (Suppl. Figure S2). Thus, most amino acids tend to longer homopeptides when more homopeptides are in a proteome.

The frequency ranking of homopeptides of different types of amino acids can also change within smaller clades and genera (Suppl. Figure S1). Such changes even appear between different strains of the same species. For example, among the six strains of budding yeast Saccharomyces cerevisiae, most of the top ten homopeptides shift frequency ranking compared to the other strains. The lengths of homopeptides of aliphatic hydrophobic residues, i.e., poly-Leu, poly-Ile, poly-Val, are generally short in all Dikarya species (lighter cells in Suppl. Figure S1 heat maps), which may be due to the selection against protein aggregation [17], and constraints of side-chain packing in protein domain hydrophobic cores.

The amino acids that vary the most in homopeptide amount are discerned from examining the standard deviations for their ranking for homopeptide frequencies (Table 1). The top one third of the homopeptides

that change the most across Dikarya are especially highlighted in Table 1. All but one of these are from amino acids whose codon repertoire is biased for GC or AT (Table 1). However, poly-Gln specifically stands out as encoded by a codon repertoire that has no overall GC/AT-bias, but it still greatly changes in the frequency ranks across Dikarya (Table 1).

Saccharomycotina have distinct behaviour for homopeptide and homocodon evolution

Previous work on limited data sets indicated that the prevalent types of homopeptides are strongly influenced by GC bias [22-24]. Here, we investigated the effect of GC/AT levels on homopeptide and homocodon evolution on a large scale across Dikarya, and for Saccharomycotina in particular. Saccharomycotina are mostly AT-rich while species in other subphyla are mostly GC-rich, which causes homopeptide composition in Saccharomycotina to be distinct (Suppl. Figure S1, band c). The four homopeptide types which drop most in the frequency ranks in Saccharomycotina are all for GC-rich amino acids (Table 1), while the two types that rise the most in rank are poly-Asn and poly-Lys, which have AT-rich codons (Suppl. Figure S1; Table 1). This result concurs with the discovery in an analysis of prion-like proteins in Saccharomycotina that GC% influences the abundance of compositionally-biased protein regions encoded by GC- or AT-rich codons [21].

Given that homopeptides behave differently in the AT-rich Saccharomycotina relative to other subphyla, we investigated more closely how homopeptide and GC/AT trends are related.

Homopeptides tend to be GC-rich even for AT-rich genomes

It is obviously expected that the AT/GC level in coding regions outside of homopeptides/homocodons and within them are positively correlated to each other (Figure 3a-b). To examine how different the AT/GC levels within and outside homopeptides/homocodons are, we examined how the linear regressions deviate from the y=x line for both homopeptides and homocodons. GC level tends to be higher within homocodons in both AT- and GC-rich organisms, but for a large fraction of AT-rich species, homocodons are more AT-rich than other proteome areas (Figure 3a). For homopeptides, however, there is an underlying GC bias relative to outside of homopeptides even in AT-rich organisms (i.e., mainly the Saccharomycotina) (Figure 3b). This is also evident in Table 1, where only one of the top ten overall most frequent amino acids in homopeptides has an AT-biased codon repertoire, but five of them have a GC-biased codon repertoire. This may be because GC level is easier to increase in homocodons/homopeptides than AT level. Disease-causing homocodons such as CAG/GTC and CGG/GCC, which are mostly GC-rich, are found to be particularly prone to expand in models and in experiments, with a higher inherent slippage rate which is determined by propensity to form stable mismatched secondary structures [25-27]. The two repeats (CAG and CGG) are able to encode seven frequent homopeptide amino acids including Gln, Ser, Ala, etc., since the reading frame should not affect

the inherent slippage rate. Also, GC-rich low-complexity regions (including homopeptides) are recombination hotspots which may lead to increased homopeptide content [28].

Given these trends, we investigated the relationship between homocodon/homopeptide levels and GC- or AT-bias across Dikarya.

Homocodon accumulation is strongly influenced by or coupled to GC/AT bias

We probed the relationship between homopeptide and homocodon levels and GC/AT bias, across proteomes (Figure 3c-f). Interestingly, the correlation between homocodon fraction and AT/GC content splits into two directions from around 50% AT/GC (Figure 3c-d). This indicates that homocodon abundance is positively correlated with the extremeness of AT/GC bias. Also, homocodon levels are lower for proteomes that tend to medium GC/AT. Such a correlation is less strong for homopeptides but still significant (Figure 3e-f). We would expect there to be no major bars on homocodon formation simply because a genome has medium GC/AT levels. Thus, general selection pressures or mutational biases governing GC/AT bias are majorly coupled to homocodon formation and also strongly influence the appearance of homopeptides.

The factors leading to the variation of genomic GC level during evolution are complicated, including both mutational bias and natural selection [29]. When the global GC content switches due to events such as horizontal gene transfer and biased gene conversion, the concentrations of tRNA with different anticodons could quickly readjust to fit the new GC level, which would further drive the shift in codon usage bias gradually from current abundant codons to new optimal codons [30-32]. The decrease of concentrations of the previously optimal tRNAs could induce selective pressure or point mutations in previous optimal homocodons, since homocodons demanding previous tRNAs would slow down translation [33]. Also, the increase of the new optimal tRNA could influence expansion of corresponding homocodons. On the other hand, homopeptide expansion is an efficient way to increase local GC bias, and point mutation rates are also higher in homopeptides, since they are generally located in regions under less constraint, which both lead to faster GC level change, to be further selected on by the changed tRNA concentrations [33]. AT/GC-biased regions also naturally accumulate homocodons more easily due to a higher possibility of the same codons co-occurring within a biased region.

The results here imply that general selection pressures or mutational biases governing GC or AT bias influence levels of homocodon and homopeptide formation. The opposite causation, i.e., that homocodon levels are driving GC/AT bias, is not likely since homocodons are such a small percentage of the proteome, although there may be a degree of feedback as newly formed homocodons accumulate mutations. Despite this link, homopeptides tend to be more GC-rich than other areas of proteomes, even in AT-rich organisms, indicating that they absorb AT bias trends less than other areas of the proteome, or have an inherent tendency to higher GC content, as discussed above.

Homocodon codon preferences are correlated with AT/GC bias for many codons, but there are exceptions

It is known that the genomic GC level significantly affects codon usage bias [34-36], and this is also evident here in the varying rankings of homocodon frequencies across Dikarya (Suppl. Figure S1). To probe this phenomenon, we analyzed the variation in codon preference for the five most common amino acids that are encoded by two alternative codons (E, GAA/GAG; D, GAT/GAC; K, AAG/AAA; N, AAC/AAT; Q, CAG/CAA). Not surprisingly, given the overall trends linked to AT/GC bias discussed above, the codon types in homocodons also change according to the GC/AT-level in coding regions. The predominant codon encoding poly-Glu in the clades of GC-rich species is GAG, but it switches to GAA in the AT-rich Saccharomycotina (Suppl. Figure S3). Likewise, the predominant codon encoding poly-Asp switches from GAC to GAT in Saccharomycotina (Suppl. Figure S3). Such switching has also been observed for Drosophila species [37].

We further investigated the log-log plot slopes that indicate the length distributions of homocodons for three different residue types that are encoded by two alternative codons, namely K, N and Q (Figure 4). Less negative values indicate longer homocodons, and the overall density of the distributions in the different subphyla shows the prevalence of either alternative codon. Exceptionally, the predominant codon type for poly-Lys is always AAG, while its synonymous codon AAA only shows up a few times in the top 20 frequency ranks even in AT-rich species (Figure 4a; Suppl. Figure S1). This might be due to selection on poly-Lys at the protein level, and the inherent slippage difficulty of poly-AAA(K) during DNA replication. On the other hand, for some amino acids both synonymous homocodons are highly frequent. Poly-CAG(Q) and poly-CAA(Q) are both prevalent in Pezizomycotina (a GC-rich subphylum) and Saccharomycotina (AT-rich) (Figure 4c). Poly-AAC(N) and poly-AAT(N) are both prevalent in Saccharomycotina (Figure 4b; Suppl. Figure S1). The most striking distribution is of poly-AAG(K), which is bimodal in Pezizomycotina (Figure 4a). This indicates that many species in Pezizomycotina have poly-AAG(K) longer than the ordinary length of poly-AAG(K) in other clades. Thus, some homocodons have codon preferences that appear not to follow the overall trends relating to GC/AT content.

Purity of homopeptides is modulated by GC/AT bias

Next, we set out to examine the bias of homopeptides for specific codons. To do this, we calculated homopeptide purity. This is defined as the proportion of the most dominant codons in homopeptides, which is influenced by the relative importance of synonymous point mutations versus the expansions/contractions of homocodons (see Materials and Methods). We calculated the purity of homopeptides for each amino-acid type (Table 1 and Table S1). The homopeptide purity of individual amino acids varies from clade to clade (Table S1). As explained in Materials and Methods, homopeptide purities will inherently be higher for amino acids with smaller codon repertoires, so we focussed on the standard deviations of purity for analysis. However, poly-Arg in AT-rich species has only 1 of 6 Arg codons is AT-biased, thus although poly-Arg can be encoded by codons with 6-fold degeneracy, they can be

relatively pure in AT-rich species, most notably in Saccharomycotina (highlighted red in Table S1; the arginine purity value for Saccharomycotina is an outlier). Because of this arginine-specific behaviour, its homopeptide purity varies the most across Dikarya (i.e., it has the highest standard deviation of purity, Table 1). In contrast, amino acids that vary the least in homopeptide purity (as evident from their overall purity standard deviations, Table 1) have AT/GC-balanced codon repertoires. Thus, homopeptide purity variation is directly related to the GC/AT balance of the codon repertoires of each amino acid.

Intrinsic disorder is an influencing factor on homopeptide frequency

Homopeptides are prone to accumulate in intrinsically disordered regions (IDRs) [20][10,19]. This phenomenon has however yet to be examined evolutionarily on a large scale. Thus, we investigated how homopeptide accumulation and intrinsic disorder are related across Dikarya.

A scale of intrinsic disorder propensity (P_{diso}) was derived from independent data, as described in Materials and Methods (the scale is listed in Table 1). We find that P_{diso} is an influencing factor in the frequency of amino acids in homopeptides across Dikarya, since the mean frequency rank of amino acids in homopeptides is correlated with it (Figure 5A). Also, the standard deviation of the frequency rank is also correlated (to less extent) with P_{diso} (Figure 5B). This indicates that amino acids with higher P_{diso} vary more from proteome to proteome as homopeptides. Significant correlations are not found for amino-acid hydrophobicity values (listed in Table 1). In addition, homopeptides generally are more prevalent in annotated IDRs than in structured domains, and exhibit a greater variance of frequencies (Figure 5C). The much narrower variance of homopeptide fractions for structured domains indicates that they are more constrained for homopeptide formation, as organisms evolve.

Furthermore, homopeptide fraction is significantly correlated with predicted IDR fraction across Dikarya and also within each subphylum (Figure 6a, c), but anti-correlated with the total amount of structured protein domains in a proteome (Figure 6b). This result builds on a previous observation that IDRs evolve along with the expansion of homopeptides [19,20]. The trends in Figure 6a-b are understandable since, although homopeptides are also common in structured regions, length variations of homopeptides mostly occur in IDRs, thus the abundance of homopeptides largely affects the size of IDRs but not of structured regions [38,39]. Indeed, the general prevalence of the individual amino-acid types in homopeptides is mirrored by their prevalences in annotated intrinsic disorder, with the exception of hydrophobic residues, particularly leucine and valine (Suppl. Figure S4).

Previous research found that GC-richness is linked to increased protein disorder in a proteome [40]. Here, GC level and IDR fraction have positive correlation, but not with the high significance of homopeptide levels versus IDR fractions; also, there is greater deviation from the regression line for AT-biased genomes, particularly Saccharomycotina (Figure 6e). Indeed, four of the ten most common amino acids that form homopeptides in annotated intrinsic disorder have GC-biased codon repertoires (P, A, G, R), five have

AT/GC-even repertoires (S, E, D, Q, T), and only one AT-rich (K) (Figure S4D-E). Although homocodon fraction is also positively correlated with IDR fraction, this is less significant with more deviation compared with the correlation between homopeptides and IDRs (Figure 6a, d), indicating that homocodons are less characteristic of IDRs. Previously it was shown that the expansion of 'flexible' IDRs (where disorder is conserved but amino-acid sequences are quickly evolving) is largely due to homocodon slippage, but less so for sequence-conserved IDRs [41].

IDRs here are annotated with two algorithms. IDR annotations for regions rich in some amino acids such as asparagine might be underestimated, considering its hydrophilicity and enrichment in S. cerevisiae prion-forming domains, which are intrinsically disordered [42,43]. If so, the correlation of IDR and homopeptides in Figure 6a would be more significant and the correlation of IDR and GC level in Figure 6e would be less.

Conclusions

Homopeptide and homocodon accumulation is influenced by the extremeness of GC/AT bias in coding regions, and by the intrinsic disorder propensity of specific amino acid types. Medium-GC/AT organisms have lower levels of homocodons/homopeptides, compared to those with extremer GC/AT bias. Even so, homopeptides inherently tend to be more GC-rich than other proteome areas. This might be due to GC-rich low-complexity regions (including homopeptides) being recombination hotspots, and also having increased slippage rates during DNA replication, as discussed above. Also, the most common residues in annotated intrinsic disorder tend have GC-rich or GC-even codon repertoires. *Saccharomycotina* have behaviour distinct from other fungal subphyla, since they are an AT-rich clade, while all other large clades are GC-rich (Figure 2). *Saccharomycotina* may have lower annotated IDR content because of deficiencies in the data sets on which the algorithms for IDR annotation are trained, although on average they do have lower fractions of homopeptides, possibly because homopeptides tend to be inherently GC-rich, as noted above (Figure 2).

Despite the overall trends involving GC/AT bias and intrinsic disorder, some amino acids have unique behaviours. For example, polyglutamine levels are highly variable across *Dikarya*, yet it is encoded by a GC/AT-balanced codon repertoire (CAG/CAA). We suggest that this variability is linked to glutamine preferring to exist in IDRs, which are under less structural constraints [44], combined with its codon CAG being one of the codons most prone to DNA slippage during replication [27]. Other residue-specific behaviours are revealed for: lysine (codons: AAG/AAA), whose predominant codon overwhelmingly tends to AAG in homopeptides; and arginine (codons: AGA/AGG/CGT/CGC/CGA/CGG), which demonstrates high homopeptide purity in the AT-rich *Saccharomycotina* owing to it having only one AT-rich codon.

Materials And Methods

Proteome data

In total, 405 *Dikarya* reference proteomes (and corresponding coding regions) were downloaded from UniProt (www.uniprot.org) in July 2018 [45]. *Dikarya* provide a good set for analyzing the principles and trends of proteome evolution, since they are comprised of the two main currently well-sampled fungal phyla (*Ascomycota* and *Basidiomycota*), that contain hundreds of fungi of interest as pathogens, and useful for food, biotechnology and laboratory research. Also, there are currently major genome-sequencing initiatives underway to improve further the sampling of the phylogenetic tree of *Saccharomycotina* (the Y1000+ project [46]), and of fungi generally (the 1000 Fungal Genomes project [47]). Furthermore, our previous work on the evolution of prion and prion-like proteins which motivated the present study was focused on fungi [21].

Dikarya phylogenetic analysis

Dikarya phylogeny was built from 18s rRNA gene sequences, which are a prominent fungal phylogenetic marker [48]. The multiple sequence alignment (MSA) of the 18S rRNA gene was obtained from SILVA [49] in March 2018, and reduced to the 405 *Dikarya* reference species. Based on the MSA, phylogenetic trees were made with the maximum likelihood phylogeny program PhyML 3.0 [50], using aBayes branch support and defaults for nucleotide sequences. Trees and associated data were depicted with ggplot2 [51] and ggtree [52].

Homopeptide andhomocodon frequencies

Homopeptides or homocodons were defined as runs of consecutive single amino acids or codons respectively. In this study, the minimum length of homopeptides and homocodons is three, and only homocodons in coding regions were considered. The positions and lengths of homopeptides were found and calculated for each proteome. The length distributions of homopeptides were further calculated in log scale and made into log-log scatter plots for each of the most abundant 10 amino acids in homopeptides (for example, Figure 7). The slopes of linear regressions were used to indicate the general quantitative distributions of the homopeptides, *i.e.,* a less steep slope indicated a greater number of long homopeptides of the amino acid in the proteome. The length distributions for the twenty most abundant homocodons were calculated in the same way as for homopeptides. Mean frequency ranks and standard deviations of frequency rank were calculated to show the variation degree of frequency rankings of homopeptides of all amino acid types among *Dikarya* (Table 1).

Homopeptide purity

A homopeptide could be composed of different codons encoding the same amino acid. To measure the extent to which homopeptides are encoded by a predominant codon, we calculated the ‘purity’ of homopeptides for each type of amino acid X using the equation below:

$$purity_{aa} = \frac{\sum n}{N}$$

with the counts given by:
n = number of the predominant (most frequent) codons in one X-homopeptide
N = number of codons in all X-homopeptides

The purity of each amino acid is further scaled through dividing by the maximum purity across the 405 proteomes for amino acids which equal codon numbers. However, those encoded by codons with 6-fold degeneracy will be generally less pure than those encoded by codons with less degeneracy. Thus, only the overall variance of purity is comparable between different amino acid types (in Table 1).

GC/AT Content

GC/AT content is calculated for coding regions. GC/AT level within and outside of homopeptides/homocodons are calculated separately for some analyses.

Intrinsic disorder

Positions of intrinsically disordered regions (IDRs) in each proteome were annotated by the default DisoPred3 and IUPred2A programs [53,54]. Many IDR annotators are only available as webservers, which cannot be used for the large-scale data here. IUPred and DisoPred are available standalone and have been ranked in the top three in at least one assessment [55]. Combined use of multiple such programs improves annotation [56]. Only IDRs ³ 30 residues long were considered, since typically an IDR of ³30 residues is classified as a long IDR, roughly a third of eukaryotic proteins on average have such long IDRs, the programs trained on long IDRs are less accurate for shorter IDRs [56]. We used the union set of IUPred and DisoPred results after comparing the differences in their IDR annotations, since we did not want to be restricted by any tendency of either program to under-annotate IDRs with specific compositional traits. On average, only 5.6% of DisoPred results are not predicted by IUPred with a proximity threshold of 10 amino acids. On average, 20.15% of IUPred prediction are not predicted by DisoPred.

A scale indicating the propensity of amino-acid types to favour disorder or structure was calculated. The fractions of each amino-acid type were derived for an IDR set from the DISPROT database [42] (version 7.0, reduced for redundancy as previously described [57]), and from the ASTRALSCOP40 protein domain database [58] (version 2.06). For the latter, the sequences derived from the Protein Data Bank file atom records were used, to minimize inclusion of intrinsic disorder. The fractions for each amino acid in the DISPROT set were then divided by the corresponding fractions in ASTRALSCOP. The logarithm of this

ratio was calculated to make a propensity (termed P_{diso}) that is positive for amino acids favouring disorder and negative for those favouring structure. Table 1 lists the scale.

Structured domain annotations

Annotations of structured domains were made by mapping the ASTRALSCOP95 data set [58] onto proteomes using BLASTP (e-value threshold =0.0001) [59]. Blast matches were sorted on increasing order of e-value, and progressively de-selected from the list if they overlap a match of smaller e-value.

Declarations

Acknowledgements

This work was supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada.

Author Contributions

Y.W. analysed data, prepared figures and tables, and wrote the paper. P.H. conceived the project, analysed data, prepared figures and tables, and wrote the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Data Availability

The data analyzed are publicly available from the Uniprot [45], SILVA [49] , DISPROT [57], and ASTRALSCOP databases [58]. Some generated data is available in Table 1 and in the Supplementary Information. Other generated data is available from the authors upon request.

References

1. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature***447**, 932-940, doi:10.1038/nature05977 (2007).
2. La Spada, A. R. & Taylor, J. P. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nature Reviews Genetics***11**, 247, doi:10.1038/nrg2748 (2010).
3. Amiel, J., Trochet, D., Clément-Ziza, M., Munnich, A. & Lyonnet, S. Polyalanine expansions in human. *Human Molecular Genetics***13**, R235-R243, doi:10.1093/hmg/ddh251 (2004).

4. Arrasate, M., Mitra, S., Schweitzer, E. S., Segal, M. R. & Finkbeiner, S. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature***431**, 805-810, doi:10.1038/nature02998 (2004).
5. Gemayel, R., Vinces, M. D., Legendre, M. & Verstrepen, K. J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet***44**, 445-477, doi:10.1146/annurev-genet-072610-155046 (2010).
6. Chavali, S. *et al.* Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. *Nature Structural & Molecular Biology***24**, 765-+, doi:10.1038/nsmb.3441 (2017).
7. Faux, N. G. *et al.* Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res***15**, 537-551, doi:10.1101/gr.3096505 (2005).
8. Björklund, Å. K., Ekman, D. & Elofsson, A. Expansion of protein domain repeats. *PLoS computational biology***2** (2006).
9. Hancock, J. M. & Simon, M. Simple sequence repeats in proteins and their significance for network evolution. *Gene***345**, 113-118, doi:https://doi.org/10.1016/j.gene.2004.11.023 (2005).
10. Jorda, J., Xue, B., Uversky, V. N. & Kajava, A. V. Protein tandem repeats – the more perfect, the less structured. *The FEBS Journal***277**, 2673-2682, doi:10.1111/j.1742-4658.2010.07684.x (2010).
11. Nithianantharajah, J. & Hannan, A. J. Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. *Bioessays***29**, 525-535 (2007).
12. Brouwer, J. R., Willemsen, R. & Oostra, B. A. Microsatellite repeat instability and neurological disease. *BioEssays***31**, 71-83, doi:10.1002/bies.080122 (2009).
13. Hannan, A. J. Tandem Repeat Polymorphisms. *Tandem Repeat Polymorphisms: Genetic Plasticity, Neural Diversity and Disease*, 1 (2013).
14. Fondon, J. W. & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences***101**, 18058-18063, doi:10.1073/pnas.0408118101 (2004).
15. McDonald, M. J., Wang, W.-C., Huang, H.-D. & Leu, J.-Y. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS biology***9** (2011).
16. Lenz, C., Haerty, W. & Golding, G. B. Increased substitution rates surrounding low-complexity regions within primate proteins. *Genome Biol Evol***6**, 655-665, doi:10.1093/gbe/evu042 (2014).
17. Sim, K. L. & Creamer, T. P. Abundance and Distributions of Eukaryote Protein Simple Sequences. *Molecular & Cellular Proteomics***1**, 983-995, doi:10.1074/mcp.M200032-MCP200 (2002).
18. Haerty, W. & Golding, G. B. Increased polymorphism near low-complexity sequences across the genomes of Plasmodium falciparum isolates. *Genome Biol Evol***3**, 539-550, doi:10.1093/gbe/evr045 (2011).
19. Simon, M. & Hancock, J. M. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biology***10**, R59, doi:10.1186/gb-2009-10-6-r59 (2009).

20. Tompa, P. Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays***25**, 847-855, doi:10.1002/bies.10324 (2003).
21. An, L., Fitzpatrick, D. & Harrison, P. M. Emergence and evolution of yeast prion and prion-like proteins. *BMC Evol Biol***16**, 24, doi:10.1186/s12862-016-0594-3 (2016).
22. Brock, G. J. R., Anderson, N. H. & Monckton, D. G. Cis-Acting Modifiers of Expanded CAG/CTG Triplet Repeat Expandability: Associations with Flanking GC Content and Proximity to CpG Islands. *Human Molecular Genetics***8**, 1061-1067, doi:10.1093/hmg/8.6.1061 (1999).
23. DePristo, M. A., Zilversmit, M. M. & Hartl, D. L. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene***378**, 19-30, doi:10.1016/j.gene.2006.03.023 (2006).
24. Dalby, A. R. A Comparative Proteomic Analysis of the Simple Amino Acid Repeat Distributions in Plasmodia Reveals Lineage Specific Amino Acid Selection. *PLOS ONE***4**, e6231, doi:10.1371/journal.pone.0006231 (2009).
25. Liu, G. & Leffak, M. Instability of (CTG)n•(CAG)n trinucleotide repeats and DNA synthesis. *Cell & Bioscience***2**, 7, doi:10.1186/2045-3701-2-7 (2012).
26. Hartenstine, M. J., Goodman, M. F. & Petruska, J. Base stacking and even/odd behavior of hairpin loops in DNA triplet repeat slippage and expansion with DNA polymerase. *Journal of Biological Chemistry***275**, 18382-18390 (2000).
27. Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J. & Deka, R. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences***94**, 1041-1046, doi:10.1073/pnas.94.3.1041 (1997).
28. Jiang, H. *et al.* High recombination rates and hotspots in a Plasmodium falciparum genetic cross. *Genome Biol***12**, R33, doi:10.1186/gb-2011-12-4-r33 (2011).
29. Hildebrand, F., Meyer, A. & Eyre-Walker, A. Evidence of Selection upon Genomic GC-Content in Bacteria. *PLOS Genetics***6**, e1001107, doi:10.1371/journal.pgen.1001107 (2010).
30. Fitzpatrick, D. A. Horizontal gene transfer in fungi. *FEMS Microbiology Letters***329**, 1-8, doi:10.1111/j.1574-6968.2011.02465.x (2012).
31. Gladieux, P. *et al.* Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Mol Ecol***23**, 753-773, doi:10.1111/mec.12631 (2014).
32. Sun, Y., Tamarit, D. & Andersson, S. G. E. Switches in Genomic GC Content Drive Shifts of Optimal Codons under Sustained Selection on Synonymous Sites. *Genome Biology and Evolution***9**, 2560-2579, doi:10.1093/gbe/evw201 (2016).
33. Yona, A. H. *et al.* tRNA genes rapidly change in evolution to meet novel translational demands. *eLife***2**, e01339-e01339, doi:10.7554/eLife.01339 (2013).
34. Behura, S. K. & Severson, D. W. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biological Reviews***88**, 49-61, doi:10.1111/j.1469-185X.2012.00242.x (2013).

35. Hershberg, R. & Petrov, D. A. Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLOS Genetics* **6**, e1001115, doi:10.1371/journal.pgen.1001115 (2010).
36. Li, J., Zhou, J., Wu, Y., Yang, S. & Tian, D. GC-Content of Synonymous Codons Profoundly Influences Amino Acid Usage. *G3 (Bethesda)* **5**, 2027-2036, doi:10.1534/g3.115.019877 (2015).
37. Huntley, M. A. & Clark, A. G. Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. *Molecular Biology and Evolution* **24**, 2598-2609, doi:10.1093/molbev/msm129 (2007).
38. Light, S., Sagit, R., Sachenkova, O., Ekman, D. & Elofsson, A. Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol* **30**, 2645-2653, doi:10.1093/molbev/mst157 (2013).
39. Mularoni, L., Veitia, R. A. & Albà, M. M. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* **89**, 316-325, doi:https://doi.org/10.1016/j.ygeno.2006.11.011 (2007).
40. Basile, W., Sachenkova, O., Light, S. & Elofsson, A. High GC content causes orphan proteins to be intrinsically disordered. *PLoS computational biology* **13**, e1005375 (2017).
41. Rorick, M. M. & Wagner, G. P. The origin of conserved protein domains and amino acid repeats via adaptive competition for control over amino acid residues. *J Mol Evol* **70**, 29-43, doi:10.1007/s00239-009-9305-7 (2010).
42. Hatos, A. *et al.* DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res* **48**, D269-D276, doi:10.1093/nar/gkz975 (2020).
43. Harbi, D. & Harrison, P. M. Interaction networks of prion, prionogenic and prion-like proteins in budding yeast, and their role in gene regulation. *PloS one* **9**, e100615, doi:10.1371/journal.pone.0100615 (2014).
44. Campen, A. *et al.* TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein and peptide letters* **15**, 956-963, doi:10.2174/092986608785849164 (2008).
45. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-370 (2003).
46. Calhoun, S., Mondo, S. J. & Grigoriev, I. V. Yeasts and how they came to be. *Nat Rev Microbiol* **17**, 649, doi:10.1038/s41579-019-0274-6 (2019).
47. Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res* **42**, D699-704, doi:10.1093/nar/gkt1183 (2014).
48. Yarza, P., Yilmaz, P., Panzer, K., Glockner, F. O. & Reich, M. A phylogenetic framework for the kingdom Fungi based on 18S rRNA gene sequences. *Mar Genomics* **36**, 33-39, doi:10.1016/j.margen.2017.05.009 (2017).
49. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590-D596, doi:10.1093/nar/gks1219 (2012).

50. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol***59**, 307-321, doi:10.1093/sysbio/syq010 (2010).
51. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer, 2016).
52. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution***8**, 28-36, doi:10.1111/2041-210x.12628 (2017).
53. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol***337**, 635-645, doi:10.1016/j.jmb.2004.02.002 (2004).
54. Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics***21**, 3433-3434, doi:10.1093/bioinformatics/bti541 (2005).
55. Meng, F., Uversky, V. N. & Kurgan, L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cellular and Molecular Life Sciences***74**, 3069-3090, doi:10.1007/s00018-017-2555-4 (2017).
56. Atkins, J. D., Boateng, S. Y., Sorensen, T. & McGuffin, L. J. Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies. *Int J Mol Sci***16**, 19040-19054, doi:10.3390/ijms160819040 (2015).
57. Harrison, P. M. Compositionally Biased Dark Matter in the Protein Universe. *Proteomics***18**, e1800069, doi:10.1002/pmic.201800069 (2018).
58. Fox, N. K., Brenner, S. E. & Chandonia, J. M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res***42**, D304-309, doi:10.1093/nar/gkt1240 (2014).
59. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res***25**, 3389-3402 (1997).

Table

Due to technical limitations, table 1 is only available as a download in the Supplemental Files section.

Figures

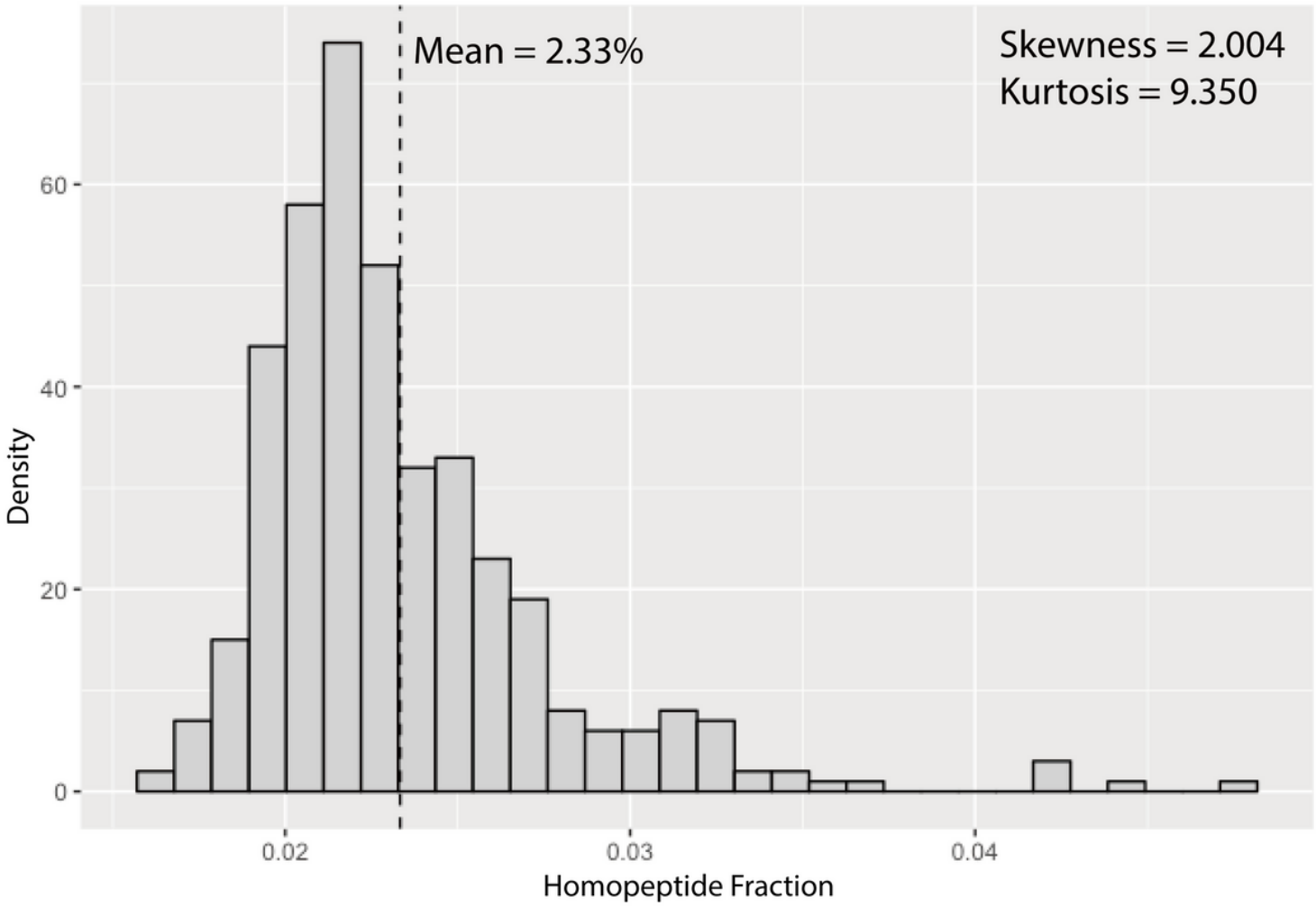

Figure 1

Distribution of overall homopeptide fraction in the proteomes. Mean = 0.023, standard deviation = 0.004, skewness = 2.004, kurtosis = 9.350.

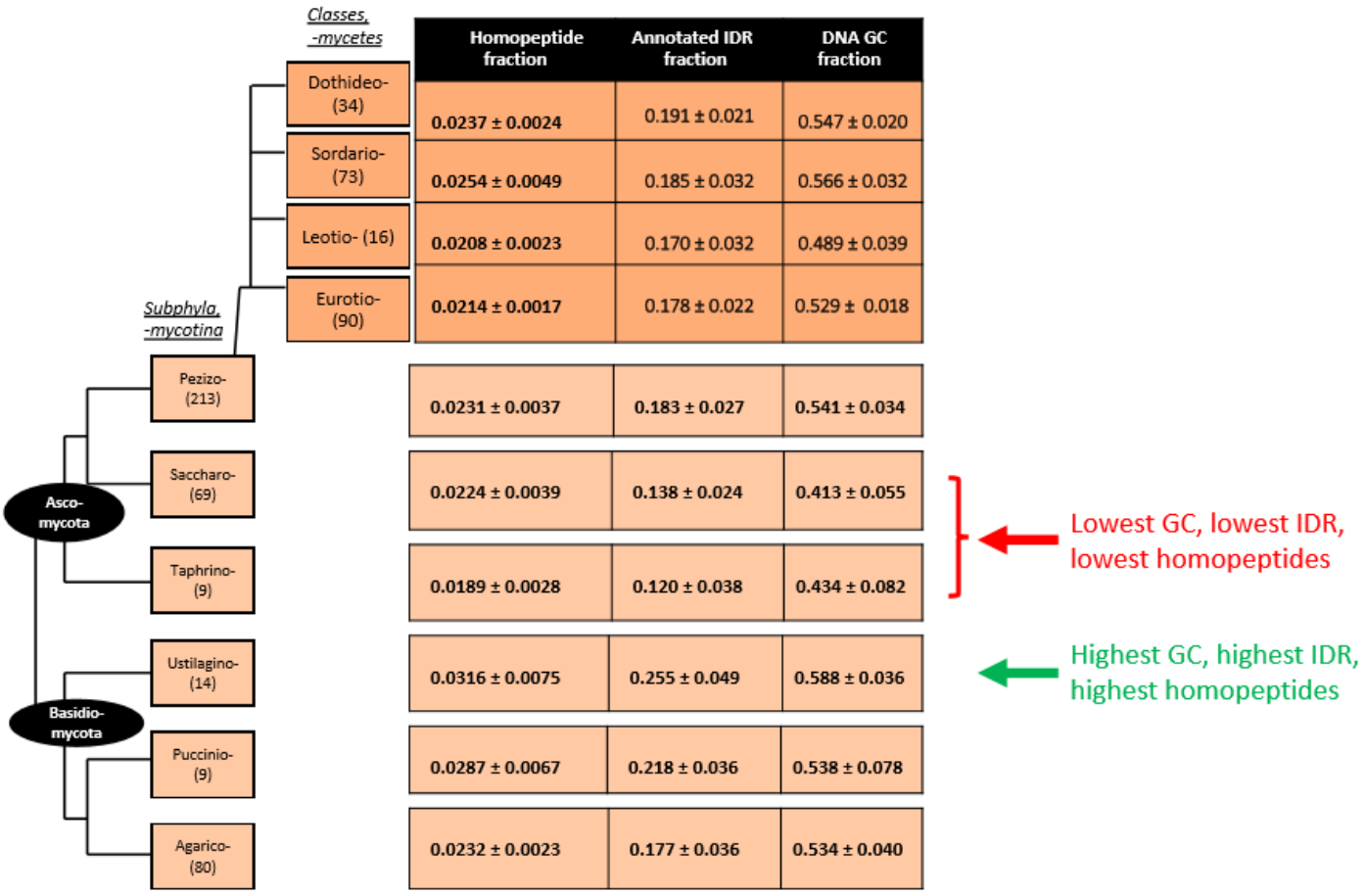

Figure 2

Schematic Dikarya phylogenetic tree with mean fractions of homopeptides, annotated IDRs and DNA GC content. The values for sub-phyla, and classes within large subphylum Pezizomycotina are shown.

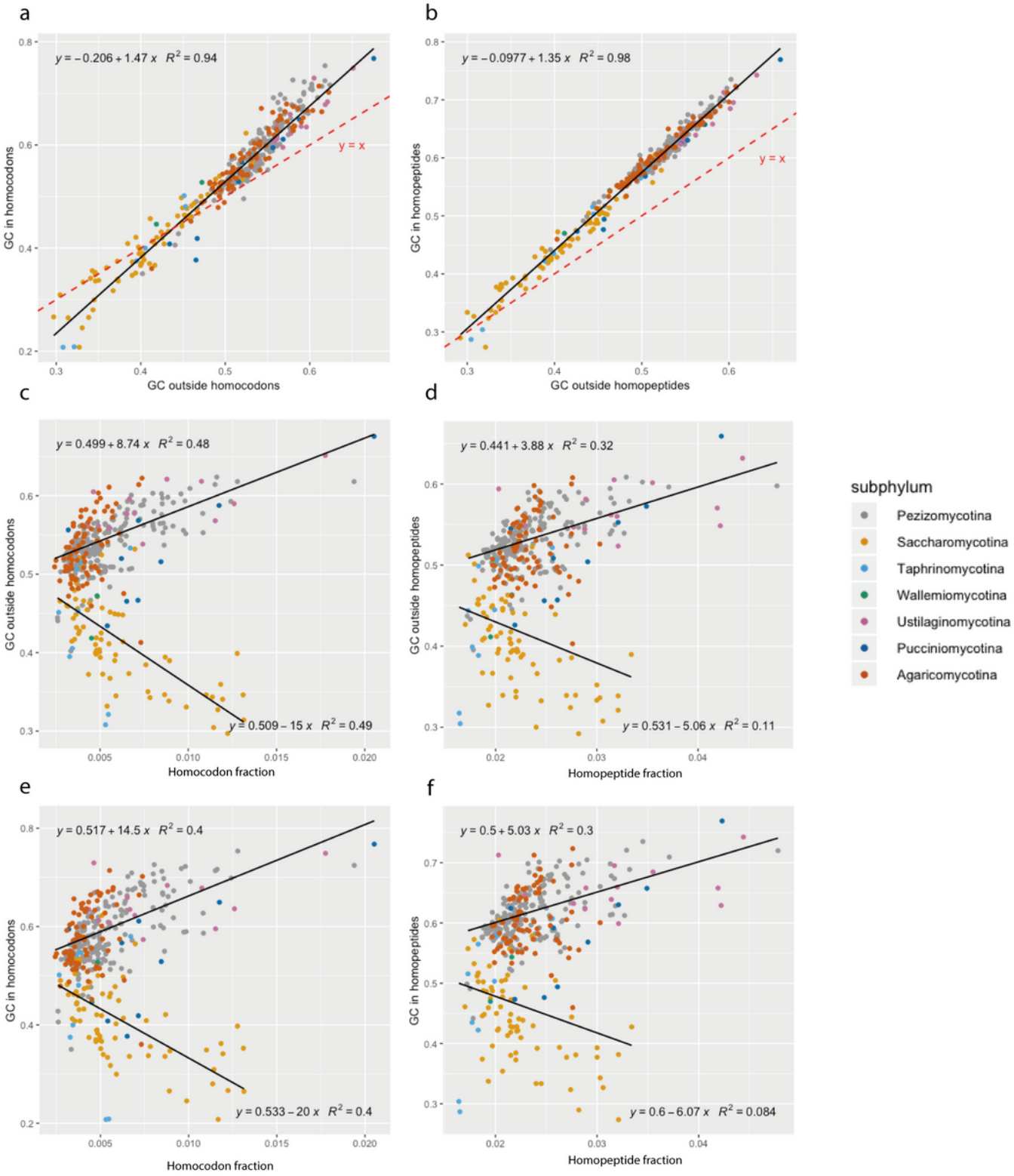

Figure 3

Relationship between homopeptide / homocodon level and GC/AT level. (a) GC/AT level in homocodons versus outside homocodons in coding regions. The red line shows the default where GC/AT levels outside and inside homocodons are identical (y=x line). (b) GC/AT level in homopeptides versus outside homopeptides. The y=x line is shown. (c) GC/AT-level outside homocodons versus the fraction of homocodons, with separate linear regressions for GC-biased and AT-biased organisms. (d) GC/AT-level

outside homopeptides versus the fraction of homopeptides, with separate linear regressions for GC-biased and AT-biased organisms. (e) GC/AT-level in homocodons plotted versus the fraction of homocodons, with separate linear regressions for GC-biased and AT-biased organisms. (f) GC/AT-level in homopeptides plotted versus the fraction of homopeptides, with separate linear regressions for GC-biased and AT-biased organisms. All correlations in parts (a)-(f) are significant at P<0.05.

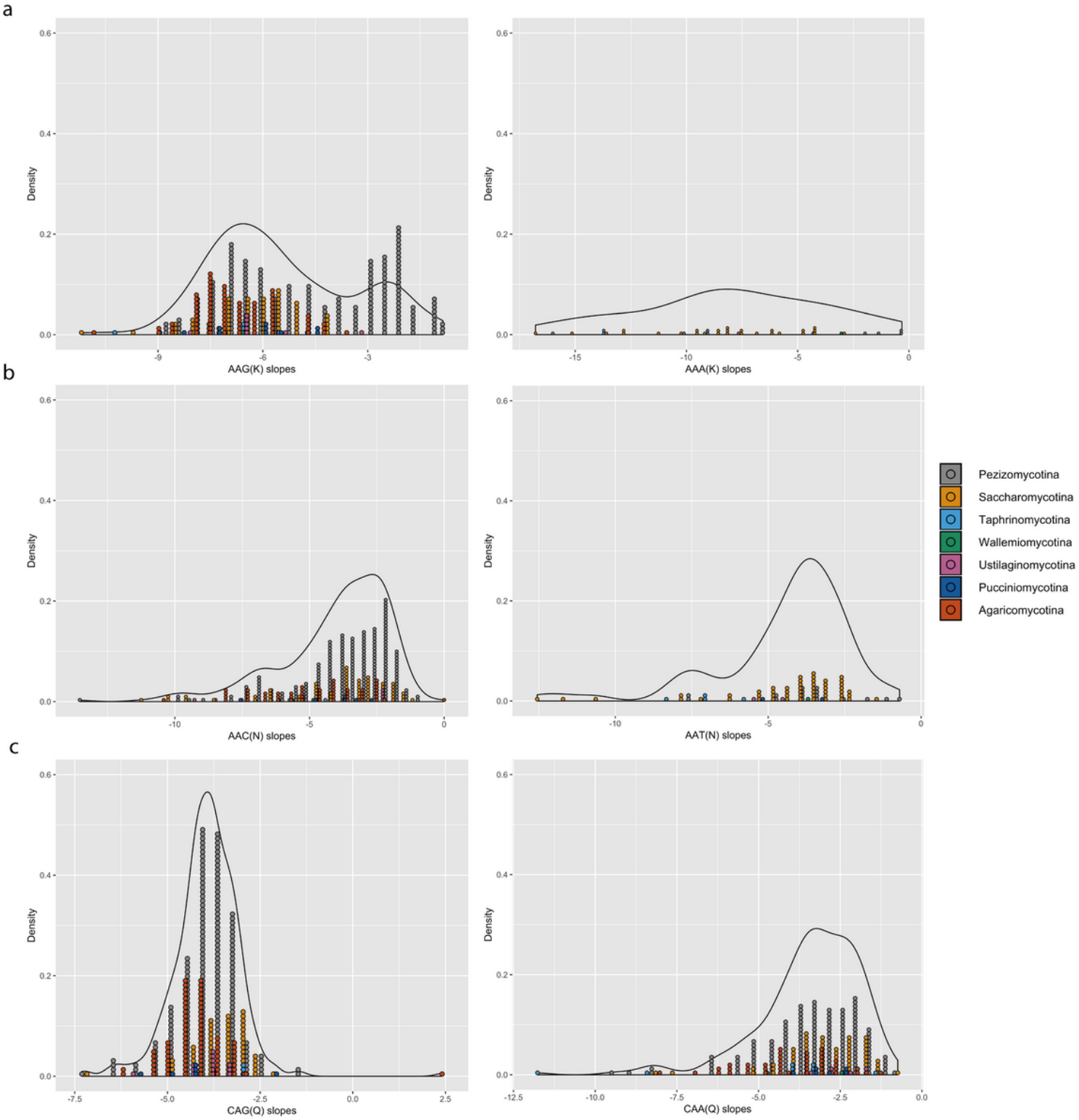

Figure 4

Length distributions of two synonymous homocodons encoding poly-Lys, poly-Asn and poly-Gln from the top-20 lists of homocodon frequencies. Histograms of the log-log plot slopes for lengths distributions are plotted. They are binned in intervals of 0.5. The total areas of the histograms for each panel indicate the total amount of each codon. The lines indicate the overall distribution within each panel. Less negative values indicate longer homopeptides: (a) Comparison of Poly-AAG(K) and Poly-AAA(K); (b) Comparison of Poly-AAC(N) and poly-AAT(N); (c) Comparison of Poly-CAG(Q) and poly-CAA(Q).

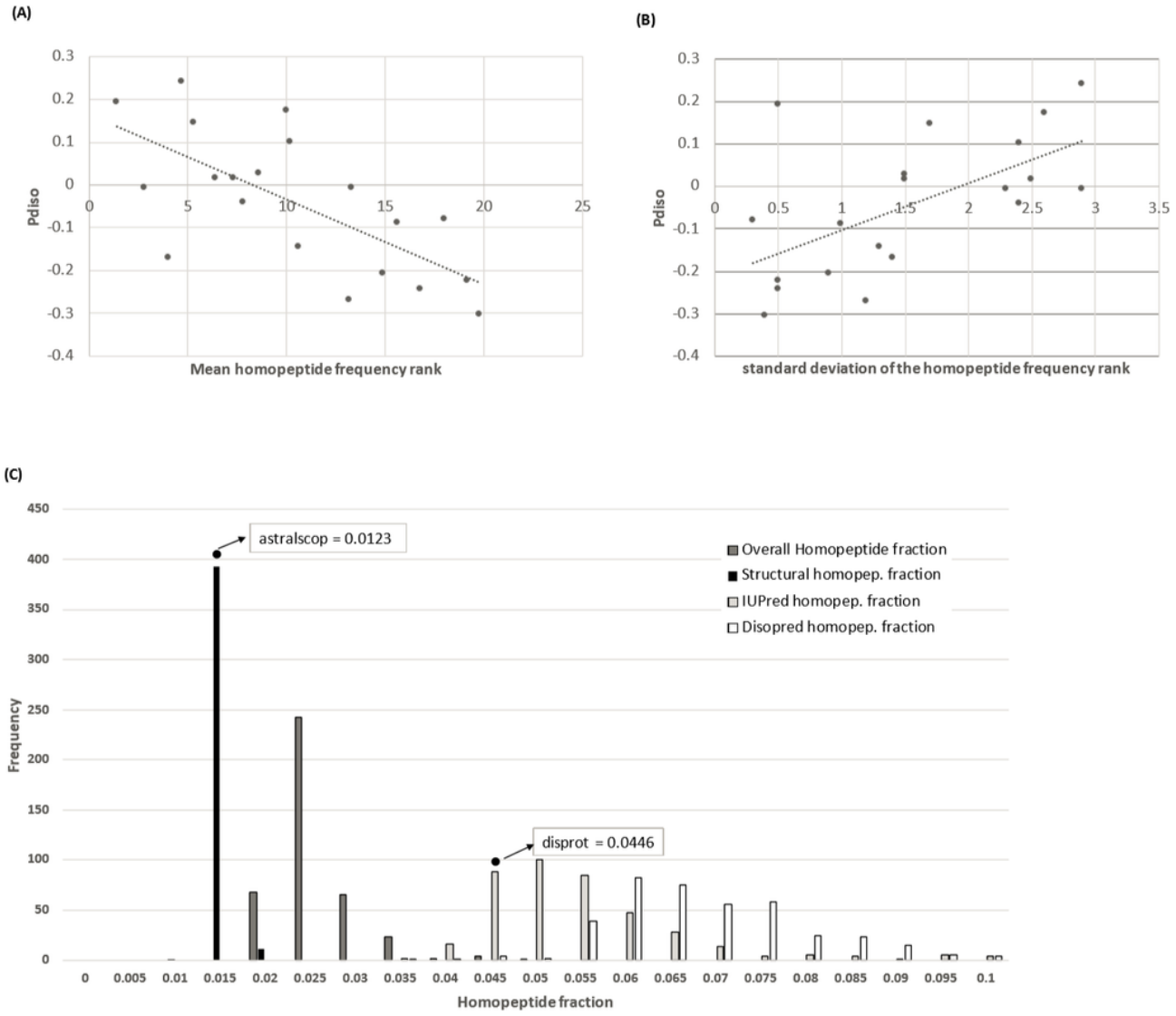

Figure 5

Intrinsic disorder propensity. The intrinsic disorder propensity (Pdiso) of the amino acids is plotted against (A) the mean frequency rank across proteomes of the amino acids in homopeptides (Pearson correlation coefficient R=0.69 (P=0.0008), and (B) the standard deviation of the frequency rank of the amino acids in homopeptides (R=0.60, P=0.005). In part (C), histograms are depicted of the homopeptide fractions of structured regions (annotations made using SCOP domains), and of the IUPred and DISOPRED annotations, with the distribution of the overall homopeptide fractions in the proteomes for

comparison. Also indicated on the plot as points are the homopeptide fractions for the ASTRALSCOP40 and DISPROT databases.

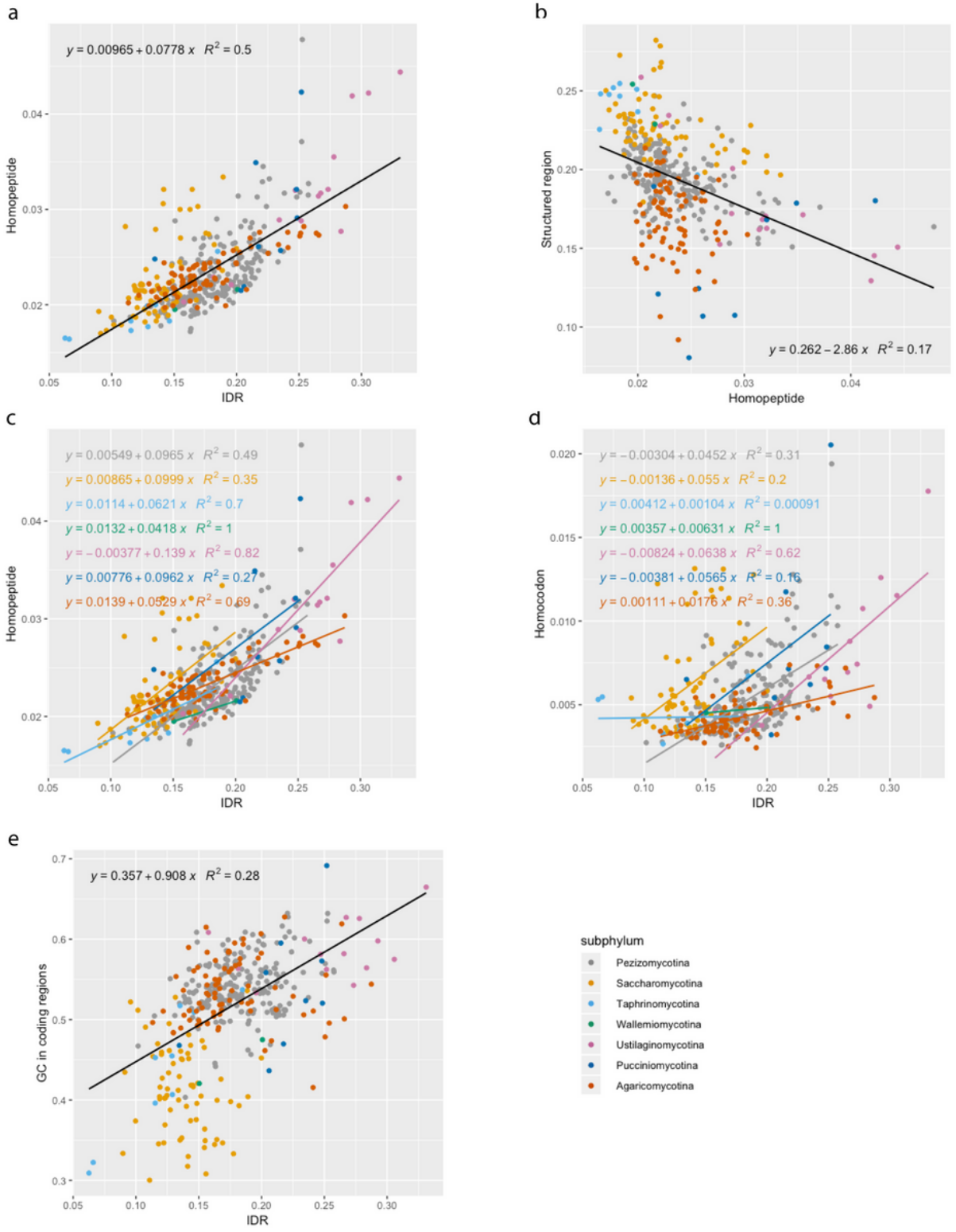

Figure 6

Relationship of homopeptide/homocodon fractions with intrinsic disorder, structured domains and GC content. Scatter plots are drawn of: (a) homopeptide fraction versus annotated IDR fraction, with an overall linear regression fitted. (b) homopeptide fraction versus fraction of structured domains, with an

overall linear regression. (c) homopeptide fraction versus annotated IDR fraction, with linear regressions fitted for each subphylum. P-values for correlations are <0.05, except for Wallemiomycotina. (d) homocodon fraction versus annotated IDR fraction, with regressions for each subphylum (correlation P-values are <0.05, except for Wallemio-, Taphrino- and Pucciniomycotina). (e) GC fraction in coding regions versus annotated proteome IDR fraction.

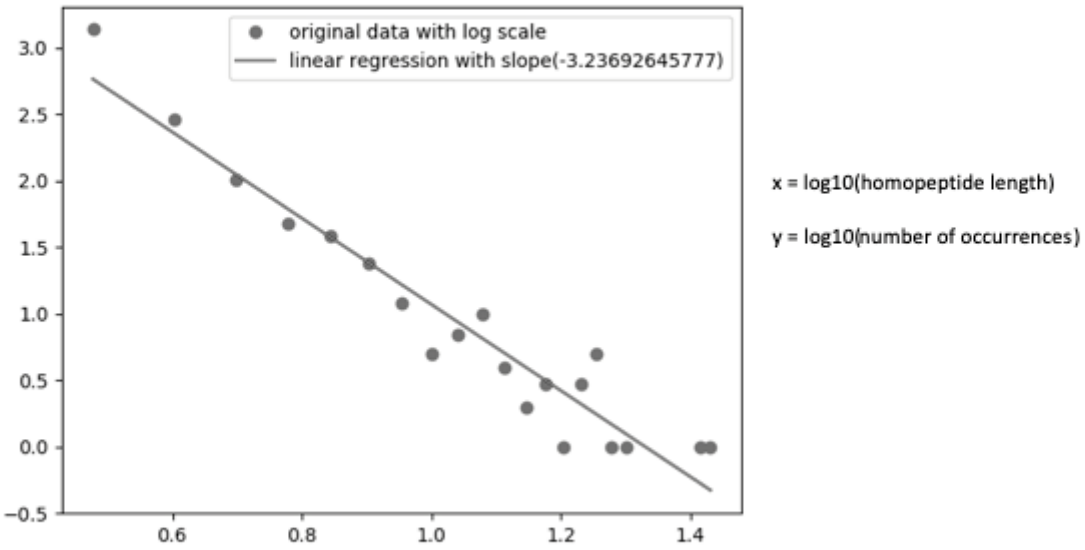

Figure 7

Example of a log-log plot used in the analysis of homopeptide or homocodon distributions. The length distributions are analyzed as log-log scale plots of the number of occurrences of a given homopeptide length versus homopeptide length. The distributions are characterized as linear regressions, yielding a calculated power-law relationship between homopeptide length and frequency for a given amino-acid type.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplFileSciRep.pdf
- Table1.docx