scientific reports

Check for updates

OPEN

Homopeptide and homocodon levels across fungi are coupled to GC/AT-bias and intrinsic disorder, with unique behaviours for some amino acids

Yue Wang & Paul M. Harrison✉

Homopeptides (runs of one amino-acid type) are evolutionarily important since they are prone to expand/contract during DNA replication, recombination and repair. To gain insight into the genomic/proteomic traits driving their variation, we analyzed how homopeptides and homocodons (which are pure codon repeats) vary across 405 *Dikarya*, and probed their linkage to genome GC/AT bias and other factors. We find that amino-acid homopeptide frequencies vary diversely between clades, with the AT-rich *Saccharomycotina* trending distinctly. As organisms evolve, homocodon and homopeptide numbers are majorly coupled to GC/AT-bias, exhibiting a bi-furcated correlation with degree of AT- or GC-bias. Mid-GC/AT genomes tend to have markedly fewer simply because they are mid-GC/AT. Despite these trends, homopeptides tend to be GC-biased relative to other parts of coding sequences, even in AT-rich organisms, indicating they absorb AT bias less or are inherently more GC-rich. The most frequent and most variable homopeptide amino acids favour intrinsic disorder, and there are an opposing correlation and anti-correlation *versus* homopeptide levels for intrinsic disorder and structured-domain content respectively. Specific homopeptides show unique behaviours that we suggest are linked to inherent slippage probabilities during DNA replication and recombination, such as poly-glutamine, which is an evolutionarily very variable homopeptide with a codon repertoire unbiased for GC/AT, and poly-lysine whose homocodons are overwhelmingly made from the codon AAG.

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 some homopeptide types[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 lengths 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 prevalent homopeptides or homocodons are still largely unclear, and most might not be essential but rather create diversity in genomes which can be selected on[11]. Homopeptide lengths are often polymorphic between different individuals in a species, and even between different cell types or at different organismal ages[12,13]. Although phenotypic evolution is mostly modulated by *cis*-regulatory elements, homopeptide length polymorphisms are also linked to significant morphological differences, e.g., in dogs[14]. Homopeptide length variations are proposed as a 'tuning knob' that acts through expansion and contraction between generations, enabling 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 also insertions/deletions are frequently associated with homopeptides in their flanks[15,16]. Thus, homopeptides may enable rapid protein divergence, through creating more polymorphism.

Department of Biology, McGill University, Montreal, QC, Canada. ✉email: paul.harrison@mcgill.ca

Early studies found that eukaryotes have unique homopeptide distributions, i.e., their proteomes prefer/tolerate homopeptides at different lengths for different amino acids[17]. It was suggested that amino-acid preferences in low-complexity regions or homopeptides are largely driven by bias in genomic AT (adenine + thymidine) or GC (guanidine + cytidine), and are under selection pressures[16,18]. Also, previous analyses have shown that homopeptides are enriched in intrinsically disordered regions (IDRs)[19–22], as are tandem repeats generally[10,19,21].

Mularoni, et al. examined tandem repeat evolution across 12 vertebrate species, and by comparing to noncoding DNA repeats inferred that there is selection maintaining prevalent tandem repeats[23]. Schaper, et al. discovered that ~ 60% of tandem-repeat regions are deeply conserved as such across 61 eukaryotes[24]. A few other studies have focused on homopeptide evolution. In a comparison of 13 diverse eukaryotes, homopeptides were found to have no general GC- or AT -bias, and homocodons within homopeptides were longer than expected by chance[7]. Across five eukaryotes, homopeptides were enriched inside alternatively-spliced exons, which also had longer homocodons and lower codon diversity[18]. In a study of > 600 human genes, homopeptide tracts had relatively elevated mutation rates[22]. Mier, et al. discovered different positional trends for homopeptides made from different amino-acid types for a diverse sample of cellular organisms[25]. Distinct trends in conservation of compositional biases for different amino-acid types in annotated IDRs were observed in a survey of > 10,000 proteomes[26].

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[27]. Motivated by these findings, we hypothesized that the factors driving the evolution and variation 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*, comprised of the phyla *Basidiomycota* and *Ascomycota*. Previous studies have not analyzed how the factors underlying homopeptide/homocodon formation influence their variation between clades in a diverse organismal phylogeny in an integrated manner. In this study, we probe in detail how, over hundreds of millions of years of fungal evolution, both homopeptide and homocodon variation are coupled to or modulated by GC/AT bias and intrinsic disorder propensity, and discover some unique behaviours for specific amino acids and codons.

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:

1. To derive an overview of the variation in homopeptide frequencies, identifying any anomalous behaviour in specific clades;
2. 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;
3. 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);
4. 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 values are in the small range 1.8–2.4%. Only a few proteomes have homopeptide frequencies below this range, the rest varying from 2.4 to 4.8% (Fig. 1). Thus, 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* (comprising the phyla *Basidiomycota* and *Ascomycota*), and also examined other various attributes, including GC content and annotated IDR content (Fig. 2 and Suppl. Figure S1). Subphyla (and classes within the large subphylum *Pezizomycotina*) are analyzed in Fig. 2, with details of species names and prevalent amino-acid / codon types in Suppl. Figure S1. Fractions of homopeptides and IDRs are colour-coded by spectra in Suppl. Figure S1. The lowest homopeptide fractions are for *Saccharomycotina* and *Taphrinomycotina*, which are also low-GC and have the lowest annotated IDR fractions (Fig. 2). Variation of homopeptide fractions is obvious between different clades, but homopeptides and annotated IDRs can also accumulate in specific species over a short evolutionary time (Suppl. Figure S1, sections *a-b;* lighter colours for higher fractions).

Heat maps of the most abundant homopeptides and homocodons (i.e., perfect codon repeats) were derived (Suppl. Figure S1, sections *c* and *d*). The key to these heatmaps is supplied with the legend to Suppl. Fig. 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 "Methods" section. For these length distributions, lighter colours in heat map cells indicate more small homopeptides or homocodons, and darker colours a greater amount of long ones. One can see that generally there are more lighter cells for sections *c* and *d* (shorter homopeptides and homocodons) where the overall homopeptide fraction is lower (darker in section *a*) (Suppl. Figure S1). When we examine the relationship between the log(length) distribution slopes and corresponding homopeptide frequencies for each amino acid, we see that there are statistically significant correlations for most amino-acid types, although all but two have weak coefficients < 0.3 (Suppl. Figure S2). These results suggest that a tendency to shorter homopeptide sizes contributes in some way to there being fewer homopeptides in a proteome, or vice versa.

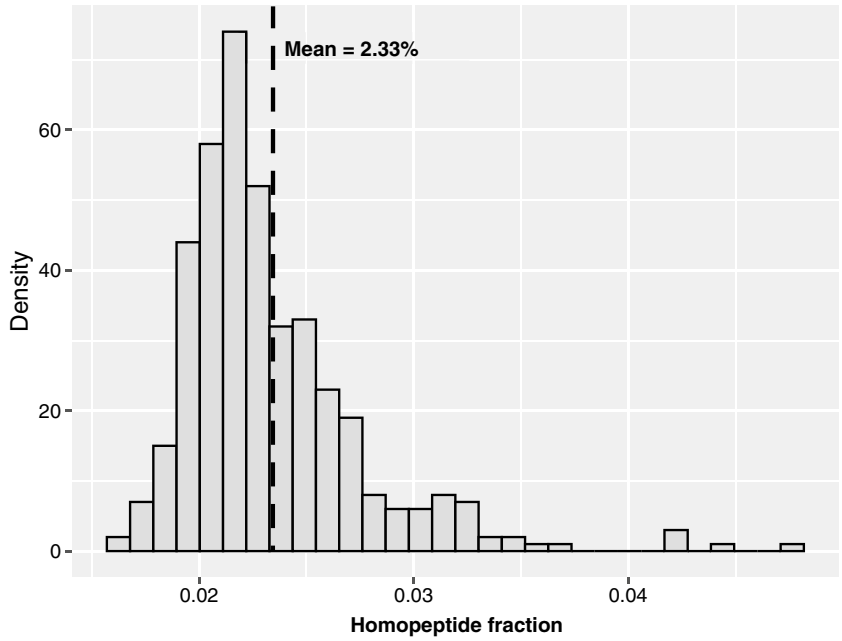

Figure 1. Distribution of overall homopeptide fraction in the proteomes. Mean = 0.023, standard deviation = 0.004, skewness = 2.004, kurtosis = 9.350. Each bin is 0.01 long and labelled with its lower bound.

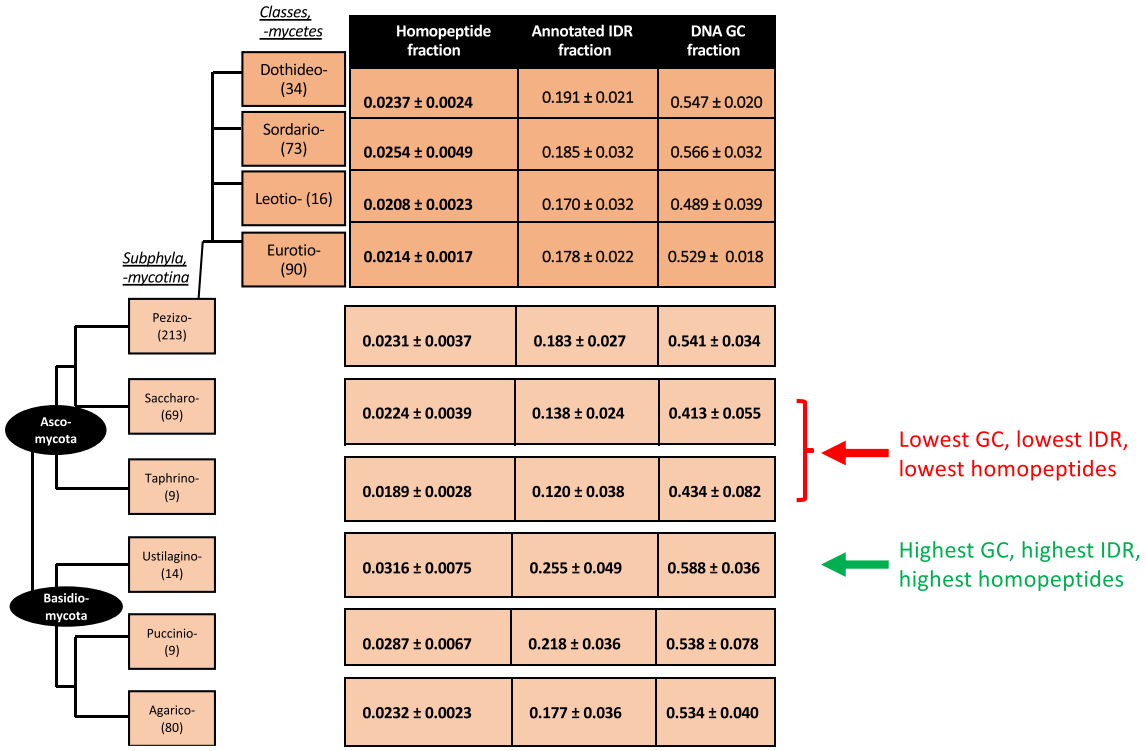

Figure 2. Schematic *Dikarya* phylogenetic tree with mean fractions of homopeptides, annotated IDRs and DNA GC content. The values for sub-phyla (clades suffixed '-mycotina'), and classes (suffixed '-mycetes') within large subphylum *Pezizomycotina* are shown.

The frequency ranking of homopeptides of different amino-acid types can also change within smaller clades and genera (Suppl. Figure S1). Such changes even appear between strains of one species. For example, among six strains of yeast *Saccharomyces cerevisiae*, most of the top ten homopeptides shift frequency ranking compared to other strains. Homopeptide lengths for aliphatic hydrophobic residues, i.e., poly-Leu, poly-Ile, poly-Val, are

Amino Acid	GC/AT bias in codon repertoire†		# of Codons	Homopeptide Frequency across *Dikarya*		Homopeptide Frequency in *Saccharomycotina*		Mean Homo-peptide Purity	Standard Deviation of Homopeptide Purity	Amino acid Disorder /Order Propensity (P_{diso})	Amino acid Hydro-phobicity †††
				Rank Mean	Rank Standard Deviation	Rank Mean††	Rank Standard Deviation				
A **	GC-biased ↑	**5/6 GC**	**4**	**2.8**	**2.3**	**6.7 ↓**	**2.5**	**0.710**	**0.041**	**–0.007**	**0.70**
P		**5/6 GC**	**4**	**4.7**	**2.9**	**10.4 ↓**	**1.9**	**0.674**	**0.041**	**0.241**	**0.32**
G		**5/6 GC**	**4**	**6.4**	**2.5**	**10.5 ↓**	**1.4**	**0.726**	**0.054**	**0.015**	**0.46**
R		**13/18 GC**	**6**	**7.8**	**2.4**	**12.1 ↓**	**2.1**	**0.635**	**0.069**	**–0.042**	**0.00**
W		2/3 GC	1	19.8	0.4	20.0	0.1	--	--	–0.304	0.40
S	AT/GC-even	1/2	6	1.4	0.5	1.1	0.3	0.607	0.035	0.192	0.41
E		1/2	2	5.3	1.7	3.3 ↑	0.9	0.767	0.032	0.146	0.11
D		1/2	2	7.3	1.5	5.4 ↑	1.2	0.772	0.031	0.015	0.11
T		1/2	4	8.6	1.5	8.1	2.3	0.709	0.040	0.028	0.42
Q ***		**1/2**	**2**	**10.0**	**2.6**	**6.2 ↑**	**2.8**	**0.768**	**0.033**	**0.173**	**0.11**
V		1/2	4	10.6	1.3	11.0	2.1	0.719	0.046	–0.144	0.97
H		1/2	2	15.6	1.0	16.6	0.7	0.779	0.032	–0.090	0.14
C		1/2	2	19.2	0.5	18.9	0.3	0.797	0.050	–0.223	0.78
L	AT-biased ↓	11/18 AT	6	4.0	1.4	3.0	1.4	0.625	0.045	–0.170	0.92
M		2/3 AT	1	18.0	0.3	18.1	0.3	--	--	–0.080	0.71
K		**5/6 AT**	**2**	**10.2**	**2.4**	**6.2 ↑**	**1.9**	**0.822**	**0.056**	**0.100**	**0.07**
N		**5/6 AT**	**2**	**13.3**	**2.9**	**8.7 ↑**	**4.2**	**0.820**	**0.044**	**–0.007**	**0.11**
F		5/6 AT	2	14.9	0.9	15.2	0.6	0.812	0.051	–0.207	0.81
Y		5/6 AT	2	16.8	0.5	16.2	0.2	0.805	0.043	–0.243	0.36
I		8/9 AT	3	13.2	1.2	12.6	1.7	0.774	0.060	–0.270	1.00

Table 1. Amino-acid homopeptide frequency ranks and purities, disorder propensities and hydrophobicities Amino acids are put in the order from the highest GC bias to the highest AT bias of their encoding codons, and if at the same bias level, they are in the order of the homopeptide frequency rank mean. The frequency rank mean is the mean of the ranking for the amino acid according to its frequency in homopeptides in a proteome; also listed are the rank standard deviations. These are calculated both across the whole *Dikarya* set and just within *Saccharomycotina*. **Coloured red are the top thirds of the list of amino acids sorted on: homopeptide frequency rank mean and standard deviation across *Dikarya* and within *Saccharomycotina*, and standard deviation of homopeptide purity. Similarly, the bottom thirds of these lists are coloured green. Also, the rows in the table for amino acids with homopeptide frequency rank deviation across *Dikarya* in the top one third of amino acids are in bold and shaded. *** Q (glutamine) is underlined since among amino acids with an AT/GC-even codon repertoire, the rank of poly-Q is the most variable across *Dikarya*. †GC/AT bias is the fraction of AT or GC in the encoding codon repertoire for an amino acid. ††Amino acids that rise or fall by > or < 1.0 ranking place on average in *Saccharomycotina*, compared to *Dikarya* generally, are labelled with up or down arrows respectively. †††Kyte-Doolittle hydrophobicity scale normalized to the interval 0.0 to 1.0.

generally short across all *Dikarya* (lighter cells in Suppl. Figure S1 heat maps), possibly due to 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 red in Table 1 ('Rank standard deviation' column). 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 influenced by GC bias, and high GC content is linked to homopeptide formation[28–32]. 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, section 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 analyses of prion-like proteins in *Saccharomycotina* that GC% influences the abundance of compositionally-biased protein regions encoded by GC- or AT-rich codons[27,33].

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 (Fig. 3a-b). To examine how different are the AT/GC levels within and outside homopeptides/homocodons, 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 (GC-poor) species, homocodons are more AT-rich than other proteome areas (Fig. 3a). For homopeptides, however, there is an underlying GC bias relative to outside of homopeptides even in AT-rich (GC-poor) organisms (i.e., mainly the *Saccharomycotina*) (Fig. 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. Pathogenic GC-rich homocodons such as CAG/GTC and CGG/GCC, 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[34–36]. The two repeats (CAG and CGG) are able to encode seven frequent homopeptide amino acids including Gln, Ser, Ala, etc., since 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[37].

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

Homocodon/homopeptide accumulation is strongly coupled to GC/AT bias, with a bi-furcated correlation arising between homocodon/homopeptide levels and GC/AT bias. We probed the relationship between homopeptide and homocodon levels and GC/AT bias, across proteomes (Fig. 3c–f). Interestingly, the correlation between homocodon fraction and AT/GC content splits into two directions from around 50% AT/GC (Fig. 3c–d, with linear regressions fitted to the data split into AT-biased and GC-biased groups). This indicates that homocodon abundance is positively correlated with the extremeness of AT/GC bias. Also, homocodon levels are lower for species that tend to mid-GC (~ 50% GC). Such a correlation is less strong for homopeptides but still significant (Fig. 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[38]. 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[39–41]. 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[42]. 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 or AT 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[42]. 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 homocodon/homopeptide levels. The opposite causation, i.e., that homocodon levels are driving GC/AT bias, is not likely since homocodons are such a small fraction of proteomes, 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 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 correlate with AT/GC bias for some codons, but not for others. It is known that the genomic GC level significantly affects codon usage bias[43–45], and this is also evident here in the 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-bias of coding regions. The predominant codon encoding poly-Glu in 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[46].

To further investigate the effects of AT/GC bias, we examined 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 (Fig. 4). Less negative values indicate smaller total relative amounts of short homocodons, and the overall density of the distributions in the different subphyla shows the prevalence of either alternative codon. Each dot in the plots is an occurrence in the top-20 lists of homocodons (arrayed in Suppl. Figure S1). Exceptionally, the predominant codon type for poly-Lys is always AAG, while its synonymous codon AAA only arises a few times in the top 20 frequency ranks even in AT-rich species (Fig. 4a; Suppl. Figure S1). This might be due to selection on poly-Lys at the protein level, and an inherent slippage difficulty for poly-AAA(K) during DNA replication. We focused on trends in the subphyla *Pezizomycotina*, *Saccharomycotina* and *Agaricomycotina*, since

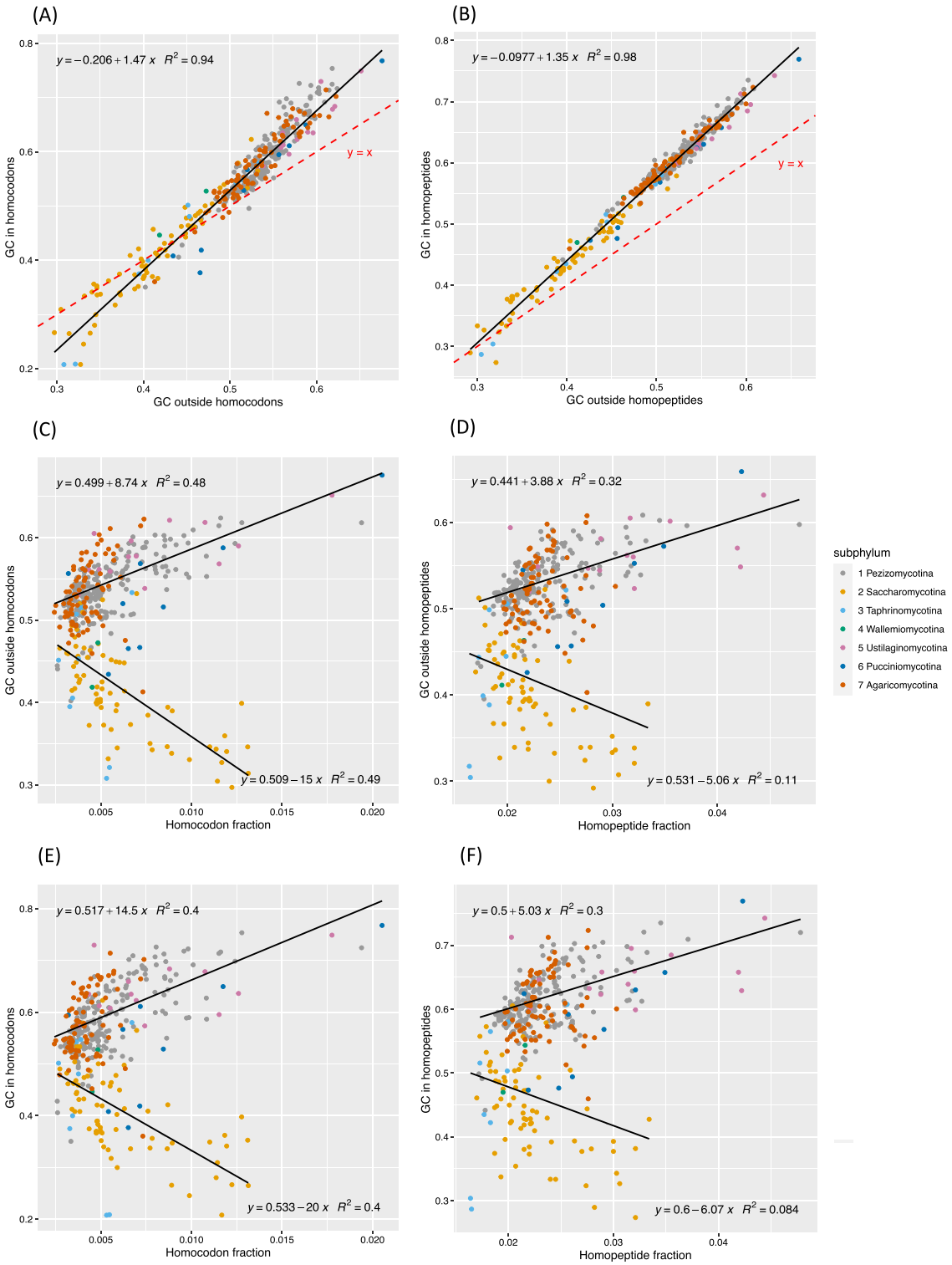

Figure 3. Relationship between homopeptide/homocodon level and GC/AT level. (**a**) GC/AT level in homocodons versus outside homocodons in coding regions. The red dashed line shows the default where GC/AT levels outside and inside homocodons are identical ($y=x$ line). (**b**) GC/AT level in homopeptides versus outside homopeptides. The $y=x$ line is shown (red dashed line). (**c**) GC/AT-level outside homocodons versus the fraction of homocodons, with separate linear regressions for GC-biased and AT-biased organisms. That is, they are separated into two groups one with GC fraction ≥ 0.5, and one with GC fraction < 0.5. (**d**) GC/AT-level outside homopeptides versus the fraction of homopeptides, with separate linear regressions for GC-biased and AT-biased organisms, as in part (**c**). (**e**) GC/AT-level in homocodons plotted versus the fraction of homocodons, with separate linear regressions for GC-biased and AT-biased organisms, as in part (c). (**f**) GC/AT-level in homopeptides plotted versus the fraction of homopeptides, with separate linear regressions for GC-biased and AT-biased organisms, as in part (**c**). All correlations in parts (**a**)–(**f**) are significant at $P<0.05$. A legend explaining the colour-coding for each subphylum is at the right of the figure.

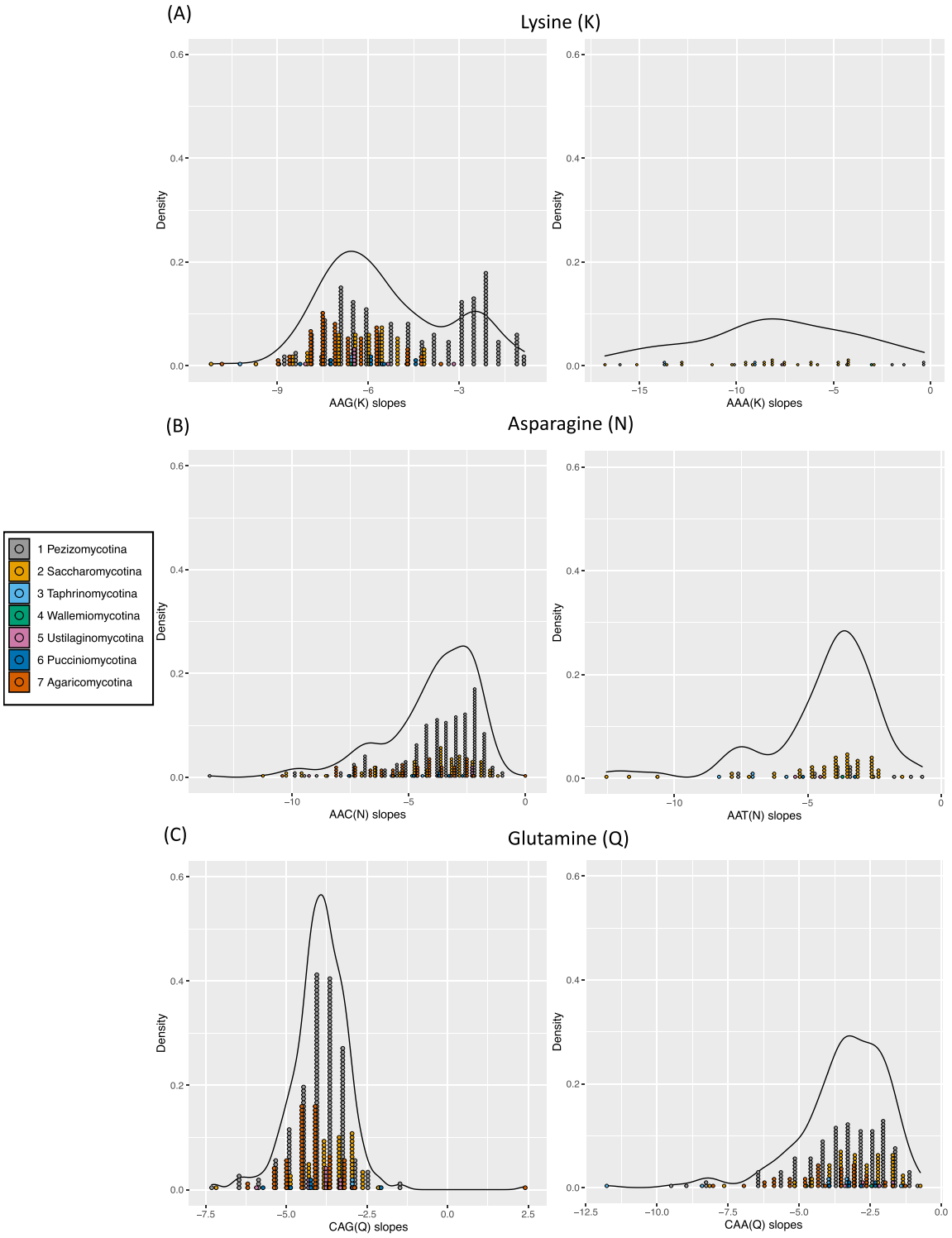

Figure 4. Histograms of length distribution slopes for two synonymous homocodons encoding poly-Lys, poly-Asn and poly-Gln from the top-20 lists of homocodon frequencies. Histograms of the log–log plot slopes for length distributions are plotted (one panel for each synonymous codon). They are binned in intervals of 0.5. For each pair of panels taken together, the total histogram area for each subphylum equals the number of occurrences in the top-20 lists for each homocodon in each subphylum. The lines indicate the overall distribution within each panel. More negative values indicate more, short homopeptides: (**a**) Comparison of Poly-AAG(K), left panel and Poly-AAA(K), right panel; (**b**) Comparison of Poly-AAC(N), left panel and poly-AAT(N), right panel; (**c**) Comparison of Poly-CAG(Q), left panel and poly-CAA(Q), right panel.

they are the largest subphyla. Strikingly, the slope distribution for poly-AAG(K) stands out as distinctly bimodal in *Pezizomycotina* (Fig. 4a). This indicates that many species in *Pezizomycotina* have poly-AAG(K) longer than the ordinary length of poly-AAG(K) in other clades. On the other hand, for some amino acids both synonymous homocodons are highly frequent. For example, poly-CAG(Q) and poly-CAA(Q) are both prevalent in *Pezizomycotina* (a GC-rich subphylum) and *Saccharomycotina* (AT-rich), and to a lesser extent *Agaricomycotina* (Fig. 4c). Also, poly-AAC(N) and poly-AAT(N) are both prevalent in *Saccharomycotina* (Fig. 4b; Suppl. Figure S1). This is despite the large-scale mutational trends during *Saccharomycotina* evolution, which have led to more amino acids encoded by more AT-biased codons, and wholesale generation of asparagine-rich regions especially[27]. Generally, these results show that some homocodons have codon preferences that do not follow the overall trends linked to GC/AT content. We surmise that this is due to the inherent slippage probability of specific codons during DNA replication and recombination.

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* expansions/contractions of homocodons (see "Methods" section). Homopeptide purity was calculated for each amino-acid type (Table 1, Table S1). These amino-acid homopeptide purities vary from clade to clade (Table S1). As explained in the "Methods" section, homopeptide purities will inherently be higher for amino acids with smaller codon repertoires, so we focussed on the standard deviations of purity for analysis. Only 1 of 6 Arg codons is AT-biased, thus although poly-Arg can contain codons with six-fold degeneracy, they can be relatively pure in AT-rich species, most notably *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 evidenced by their overall purity standard deviations, Table 1) have AT/GC-balanced codon repertoires, i.e., equal numbers of A + T and G + C. Thus, homopeptide purity variation is directly related to the GC/AT balance of the codon repertoires of each amino acid.

Intrinsic disorder is correlated with both homopeptide frequency and variability across *Dikarya*. Homopeptides are prone to accumulate in intrinsically disordered regions (IDRs)[10,19,20]. This phenomenon has however yet to be examined evolutionarily across a large phylogeny with many sub-clades with a spectrum of AT- and GC- bias. Thus here, we investigated how homopeptide variation and intrinsic disorder are associated across *Dikarya*.

A scale of intrinsic disorder propensity ($\mathbf{P_{diso}}$) was derived from independent data (*Methods*; scale listed in Table 1). We find that $\mathbf{P_{diso}}$ influences both amino-acid frequency (Fig. 5A) and variability (Fig. 5B) in homopeptides across *Dikarya*. Thus, amino acids with higher $\mathbf{P_{diso}}$ vary more from proteome to proteome as homopeptides. Significant correlations are not found for amino-acid hydrophobicities (listed in Table 1). Also, homopeptides are consistently more prevalent in annotated IDRs than in structured domains, and exhibit a far greater variance of frequencies (Fig. 5C). The much narrower variance of homopeptide fractions in structured domains indicates comparatively very tight constraint.

Furthermore, homopeptide fraction is significantly correlated with annotated IDR fraction across *Dikarya* and also within each subphylum (Fig. 6a,c), but anti-correlated with structured protein-domain content in proteomes (Fig. 6b). The large AT-rich sub-phylum *Saccharomycotina* has less correlation than GC-rich sub-phyla generally, maybe because of the favouring of GC-richness in IDRs (Fig. 6c, and see below for Fig. 6e). This result builds on previous observations on diverse cellular organisms that IDRs evolve along with homopeptide expansion[19,20]. Although homopeptides are also common in structured regions, total homopeptide lengths mostly vary in IDRs, and homopeptide abundance largely affects the size of IDRs but not of structured regions. Indeed, IDRs generally have higher insertion/deletion rates, and intrinsic disorder content is the major determinant of protein length[23,47–49]. Also, the general prevalence of the amino-acid types in homopeptides is mirrored by their prevalences in annotated IDRs (save for hydrophobic residues, particularly leucine and valine) (Suppl. Figure S4).

Previous research found that GC-richness is linked to increased proteomic intrinsic disorder[50,51]. Here, GC level and IDR fraction have significant positive correlation, but not to the same extent as for homopeptide levels *versus* IDR fractions; also, AT-biased genomes, especially *Saccharomycotina*, deviate more from the regression line (Fig. 6e). Indeed, 4 out of the 10 most common amino acids in homopeptides within 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 also positively correlates with IDR fraction, this is less than the correlation between homopeptides and IDRs (Fig. 6a,d), indicating that homocodons are less characteristic of IDRs. However, some sub-phyla are relatively more correlated indicating more homocodon content in their IDRs.

Two algorithms were used to annotate IDRs. IDRs rich in some amino acids might be underestimated, e.g., asparagine, considering its hydrophilicity and enrichment in *S. cerevisiae* prion domains, which have intrinsic disorder[52–55]. If so, IDR and homopeptide fractions (Fig. 6a) would be more correlated, and the correlation of IDR and GC level would be less (Fig. 6e).

Conclusions

Here we examined the diverse, well-sampled fungal sub-kingdom *Dikarya* for trends in the variation of homopeptides. The *Dikarya* fungi are particularly attractive for such analysis (as explained in full in "Methods" section), not least because they comprise large clades made from AT- and GC-biased species. We observed that amino-acid

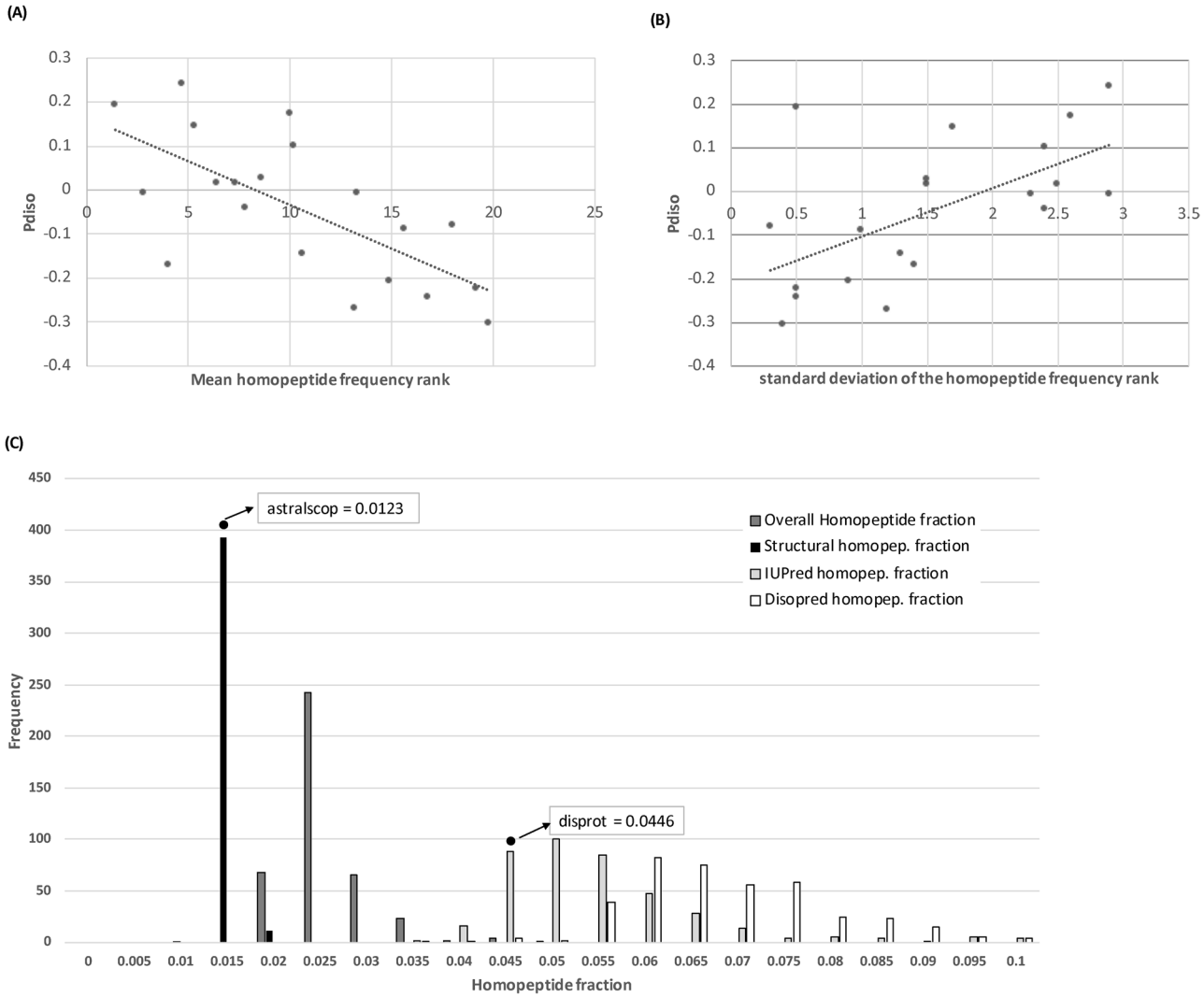

Figure 5. Intrinsic disorder propensity. The intrinsic disorder propensity ($\mathbf{P_{diso}}$) of the amino acids is plotted against (**A**) the mean frequency rank across proteomes of the amino acids in homopeptides (Pearson correlation coefficient R = −0.69, $P = 0.0008$), and (**B**) the standard 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 intrinsic disorder annotations, with the distribution of the overall homopeptide fractions in the proteomes for comparison. The fractions of homopeptides for disordered regions and for structured regions are calculated as fractions of the total number of residues in the disordered and structured subsets of residues respectively. Also indicated on the plot as points are the homopeptide fractions for the ASTRALSCOP40 and DISPROT databases.

homopeptide frequencies vary diversely between clades (even between closely related organisms), with the AT-rich *Saccharomycotina* trending distinctly. Dissection of this variation has yielded multiple insights, including:

1. *Homopeptides tend to be GC-rich even for AT-rich genomes, indicating they absorb AT bias less or are inherently more GC-rich.* This trend is less pronounced for homocodons. We surmised that these tendencies may be because GC level is easier to increase in homocodons/homopeptides than AT-level, owing to several factors including inherent slippage rates of individual trinucleotides such as CAG/GTC and CGG/GCC.
2. *Homocodon/homopeptide accumulation is strongly coupled to GC/AT bias, with a dual bi-furcated correlation between homocodon/homopeptide levels and GC or AT bias.* This indicates that mid-GC species tend to have fewer homocodons/homopeptides simply because they are mid-GC.
3. *Homocodon codon preferences are correlated with AT/GC bias for some codons, but not for others.* When homocodon codon preferences were examined for the amino acids encoded by two alternative codons, we found that while some amino-acid codon choices follow genomic AT/GC bias trends (e.g., Glu), others do not (e.g., Lys). Again, we surmise that this is due to different inherent slippage rates for different codons during DNA replication and recombination.
4. *The purity of homopeptides* (i.e., *the degree to which they are encoded by one specific codon) is modulated by GC/AT bias.* The amino acids that vary the least in homopeptide purity have codon repertoires that are balanced

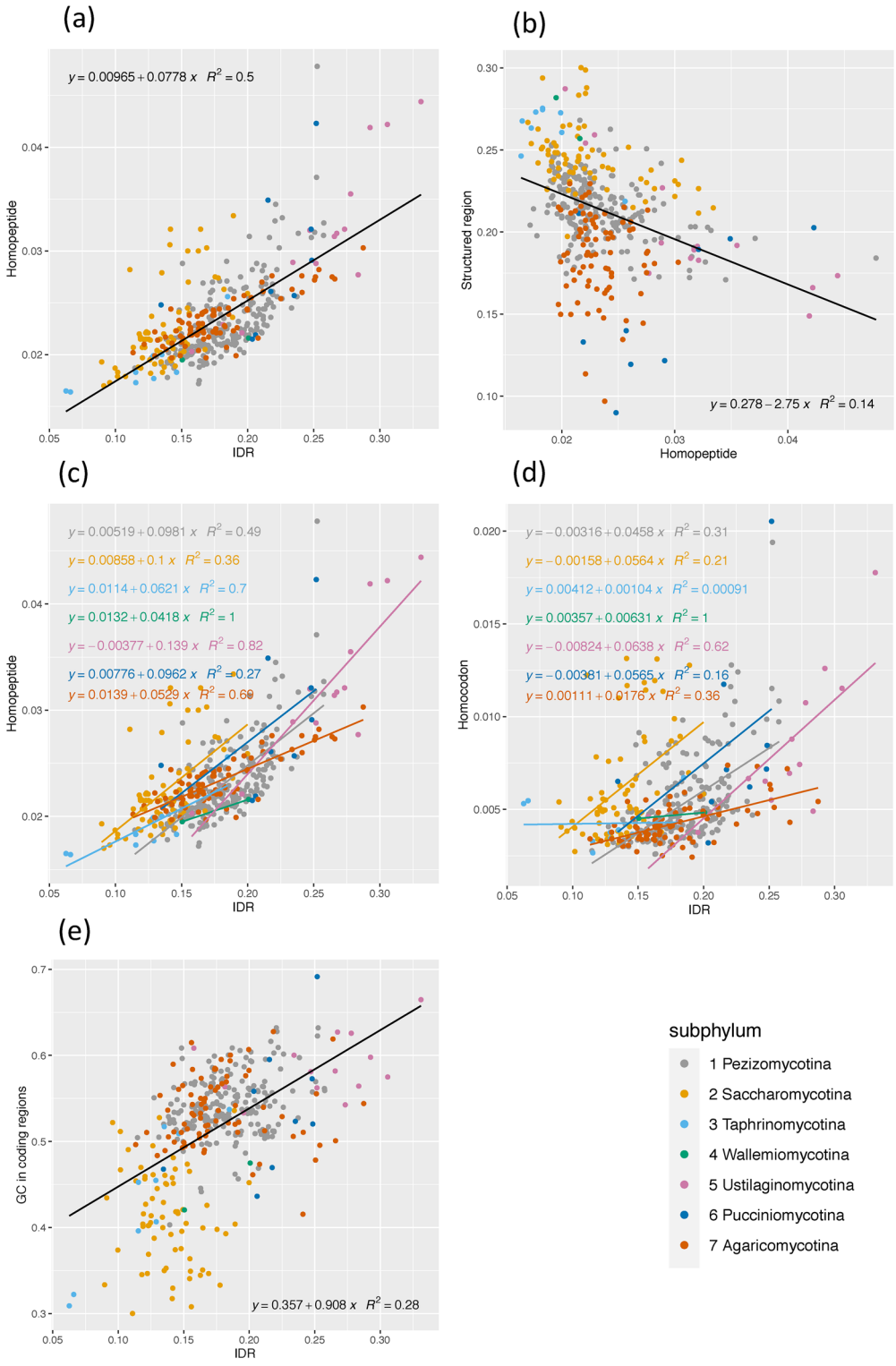

Figure 6. Relationship of homopeptide/homocodon fractions with intrinsic disorder, structured domains and GC content. Scatter plots are drawn of: (**a**) homopeptide fraction versus annotated IDR fraction, with an overall linear regression fitted (P value < 0.00001). (**b**) homopeptide fraction versus fraction of structured domains, with an overall linear regression (P value < 0.00001). (**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 (P value < 0.00001).

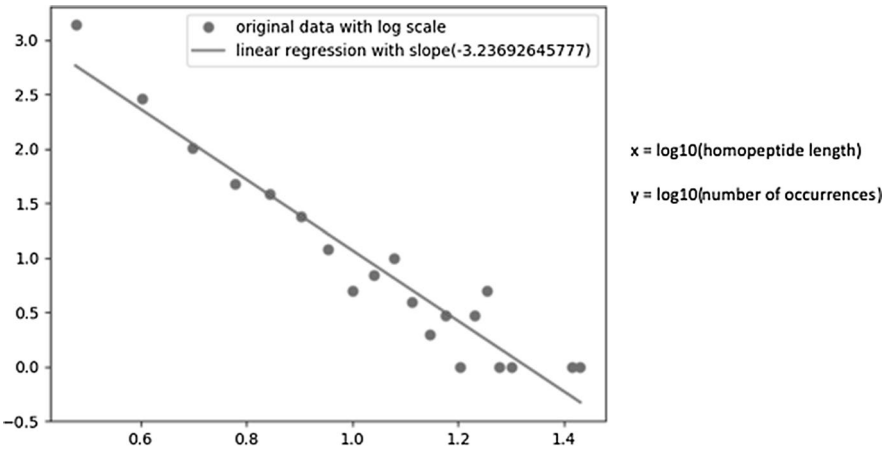

Figure 7. Example of a log–log plot used in the analysis of homopeptide or homocodon distributions. The length distributions are analyzed as log–log scale plots of the number of occurrences of a given homopeptide length versus homopeptide length. The distributions are characterized as linear regressions, yielding a calculated power-law relationship between homopeptide length and frequency for a given amino-acid type.

for A + T and G + C. Homopeptide codon usage is most volatile for poly-Arg which has only one AT-biased codon (out of six), presumably because response to an AT-biasing mutational trend or selection pressure is largely dependent on mutation to one codon, whereas five are available for the opposite trend/pressure.

5. *Intrinsic disorder is correlated with both homopeptide frequency and variability across* Dikarya, *but is less correlated for the AT-rich* Saccharomycotina. Also, we observe an opposing correlation and anti-correlation with homopeptide levels for intrinsic disorder and structured domains respectively; this (anti-)correlation pair may be capturing a signal from increased IDR insertion/deletion rates[23,47–49]. Some sub-phyla have homocodon levels relatively more correlated with IDR content, indicating more homocodon content in their IDRs.
6. *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 they are 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[56], combined with its codon CAG being one of the codons most prone to DNA slippage during replication[36]. For lysine (codons: AAG/AAA), the predominant codon overwhelmingly tends to AAG in homocodons; we hypothesize that this may also be due to inherent lack of slippage ability during DNA replication for the AAA codon. Also, arginine (codons: AGA/AGG/CGT/CGC/CGA/CGG) demonstrates high homopeptide purity in the AT-rich *Saccharomycotina* owing to it having only one AT-rich codon.

Methods

Proteome data. In total, 405 *Dikarya* reference proteomes (and corresponding coding regions) were downloaded from UniProt (www.uniprot.org) in July 2018[57]. *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[58]), and of fungi generally (the 1000 Fungal Genomes project[59]). Furthermore, our previous work on the evolution of prion and prion-like proteins which motivated the present study was focused on fungi[27]. They also contain large clades that are made from either AT- or GC-biased genomes[27].

***Dikarya* phylogenetic analysis.** *Dikarya* phylogenies were built from 18 s rRNA gene sequences, which are a prominent fungal phylogenetic marker[60]. The multiple sequence alignment (MSA) of the 18S rRNA gene was obtained from SILVA[61] 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[62], using aBayes branch support and defaults for nucleotide sequences. Trees and associated data were depicted with ggplot2[63] and ggtree[64].

Homopeptide and homocodon 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 10 most abundant amino acids in homopeptides (for example, Fig. 7). The slopes of linear regressions were used to indicate the general quantita-

tive distributions of the homopeptides, i.e., a steeper slope indicated a greater relative amount of short homopeptides in the proteome. The length distributions for the twenty most abundant homocodons were calculated in the same way as for homopeptides. Within each proteome, the types of amino acid were ranked according to their frequencies of homopeptides to give *frequency ranks*, i.e., rank 1 for the most frequent amino-acid homopeptide, rank 2 for the next, etc. Mean frequency ranks (and standard deviations of frequency rank) were calculated for each amino-acid type across *Dikarya* and *Saccharomycotina* to show the variation in the frequencies of homopeptides made from these amino acids (Table 1). Similar rankings were made for homocodon codons.

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 with equal codon numbers. However, those encoded by codons with six-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).

Intrinsic disorder. Intrinsically disordered regions (IDRs) in proteomes were annotated by the default DisoPred3 and IUPred2A programs[65,66]. Many IDR annotators are only available as webservers, so cannot be used here. IUPred and DisoPred are available standalone and were ranked in the top three in at least one assessment[67]. Combined use of multiple such programs improves annotation[68]. Only IDRs ≥ 30 residues long were considered, since typically an IDR of ≥ 30 residues is classified as a 'long', about a third of eukaryotic proteins have such long IDRs, and programs trained on long IDRs are less accurate for shorter IDRs[68]. We used the union set of IUPred and DisoPred results after comparing the differences in their annotation, since we did not want to be restricted by any tendency of a program to under-annotate IDRs with specific compositional traits. In total, only 5.6% of DisoPred results are not predicted by IUPred with a proximity threshold of 10 amino acids; 20.15% of IUPred prediction are not predicted by DisoPred.

A scale of 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[52] (version 7.0, reduced for redundancy as previously described[69]), and from the ASTRALSCOP40 protein domain database[70] (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 $\mathbf{P_{diso}}$) 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[70] onto proteomes using BLASTP (e-value threshold = 0.0001)[71]. 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.

Data availability

The data analyzed are publicly available from the Uniprot[57], SILVA[61], DISPROT[69], and ASTRALSCOP databases[70]. Some generated data is available in Table 1 and in the Supplementary Information. Other generated data is available from the authors upon request.

Received: 30 November 2020; Accepted: 22 April 2021
Published online: 11 May 2021

References

1. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940. https://doi.org/10.1038/nature05977 (2007).
2. La Spada, A. R. & Taylor, J. P. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.* **11**, 247. https://doi.org/10.1038/nrg2748 (2010).
3. Amiel, J., Trochet, D., Clément-Ziza, M., Munnich, A. & Lyonnet, S. Polyalanine expansions in human. *Hum. Mol. Genet.* **13**, R235–R243. https://doi.org/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. https://doi.org/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. https://doi.org/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. *Nat. Struct. Mol. Biol.* **24**, 765. https://doi.org/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. https://doi.org/10.1101/gr.3096505 (2005).
8. Björklund, Å. K., Ekman, D. & Elofsson, A. Expansion of protein domain repeats. *PLoS Comput. Biol.* **2**, e114 (2006).
9. Hancock, J. M. & Simon, M. Simple sequence repeats in proteins and their significance for network evolution. *Gene* **345**, 113–118. 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. *FEBS J.* **277**, 2673–2682. https://doi.org/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. https://doi.org/10.1002/bies.080122 (2009).
13. Hannan, A. J. Tandem repeat polymorphisms. in *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. *Proc. Natl. Acad. Sci.* **101**, 18058–18063. https://doi.org/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 Biol.* **9**, e1000622 (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. https://doi.org/10.1093/gbe/evu042 (2014).
17. Sim, K. L. & Creamer, T. P. Abundance and distributions of eukaryote protein simple sequences. *Mol. Cell. Proteom.* **1**, 983–995. https://doi.org/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. https://doi.org/10.1093/gbe/evr045 (2011).
19. Tompa, P. Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays* **25**, 847–855. https://doi.org/10.1002/bies.10324 (2003).
20. Simon, M. & Hancock, J. M. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol.* **10**, R59. https://doi.org/10.1186/gb-2009-10-6-r59 (2009).
21. Delucchi, M., Schaper, E., Sachenkova, O., Elofsson, A. & Anisimova, M. A new census of protein tandem repeats and their relationship with intrinsic disorder. *Genes (Basel)* https://doi.org/10.3390/genes11040407 (2020).
22. Gojobori, J. & Ueda, S. Elevated evolutionary rate in genes with homopolymeric amino acid repeats constituting nondisordered structure. *Mol. Biol. Evol.* **28**, 543–550. https://doi.org/10.1093/molbev/msq225 (2011).
23. 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. https://doi.org/10.1016/j.ygeno.2006.11.011 (2007).
24. Schaper, E., Gascuel, O. & Anisimova, M. Deep conservation of human protein tandem repeats within the eukaryotes. *Mol. Biol. Evol.* **31**, 1132–1148. https://doi.org/10.1093/molbev/msu062 (2014).
25. Mier, P., Alanis-Lobato, G. & Andrade-Navarro, M. A. Context characterization of amino acid homorepeats using evolution, position, and order. *Proteins* **85**, 709–719. https://doi.org/10.1002/prot.25250 (2017).
26. Kastano, K. *et al.* Evolutionary study of disorder in protein sequences. *Biomolecules* https://doi.org/10.3390/biom10101413 (2020).
27. An, L., Fitzpatrick, D. & Harrison, P. M. Emergence and evolution of yeast prion and prion-like proteins. *BMC Evol. Biol.* **16**, 24. https://doi.org/10.1186/s12862-016-0594-3 (2016).
28. 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. *Hum. Mol. Genet.* **8**, 1061–1067. https://doi.org/10.1093/hmg/8.6.1061 (1999).
29. 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. https://doi.org/10.1016/j.gene.2006.03.023 (2006).
30. 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. https://doi.org/10.1371/journal.pone.0006231 (2009).
31. Alba, M. M. & Guigo, R. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.* **14**, 549–554. https://doi.org/10.1101/gr.1925704 (2004).
32. Zhou, Y., Liu, J., Han, L., Li, Z. G. & Zhang, Z. Comprehensive analysis of tandem amino acid repeats from ten angiosperm genomes. *BMC Genom.* **12**, 632. https://doi.org/10.1186/1471-2164-12-632 (2011).
33. Harrison, P. M. Variable absorption of mutational trends by prion-forming domains during Saccharomycetes evolution. *PeerJ* **8**, e9669. https://doi.org/10.7717/peerj.9669 (2020).
34. Liu, G. & Leffak, M. Instability of (CTG)n•(CAG)n trinucleotide repeats and DNA synthesis. *Cell Biosci.* **2**, 7. https://doi.org/10.1186/2045-3701-2-7 (2012).
35. 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. *J. Biol. Chem.* **275**, 18382–18390 (2000).
36. Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J. & Deka, R. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci.* **94**, 1041–1046. https://doi.org/10.1073/pnas.94.3.1041 (1997).
37. Jiang, H. *et al.* High recombination rates and hotspots in a *Plasmodium falciparum* genetic cross. *Genome Biol.* **12**, R33. https://doi.org/10.1186/gb-2011-12-4-r33 (2011).
38. Hildebrand, F., Meyer, A. & Eyre-Walker, A. Evidence of Selection upon Genomic GC-Content in Bacteria. *PLoS Genet.* **6**, e1001107. https://doi.org/10.1371/journal.pgen.1001107 (2010).
39. Fitzpatrick, D. A. Horizontal gene transfer in fungi. *FEMS Microbiol. Lett.* **329**, 1–8. https://doi.org/10.1111/j.1574-6968.2011.02465.x (2012).
40. Gladieux, P. *et al.* Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Mol. Ecol.* **23**, 753–773. https://doi.org/10.1111/mec.12631 (2014).
41. 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 Biol. Evol.* **9**, 2560–2579. https://doi.org/10.1093/gbe/evw201 (2016).
42. Yona, A. H. *et al.* tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* **2**, e01339–e01339. https://doi.org/10.7554/eLife.01339 (2013).
43. Behura, S. K. & Severson, D. W. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biol. Rev.* **88**, 49–61. https://doi.org/10.1111/j.1469-185X.2012.00242.x (2013).
44. Hershberg, R. & Petrov, D. A. Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLoS Genet.* **6**, e1001115. https://doi.org/10.1371/journal.pgen.1001115 (2010).
45. Li, J., Zhou, J., Wu, Y., Yang, S. & Tian, D. GC-Content of Synonymous codons profoundly influences amino acid usage. *G3 (Bethesda, Md)* **5**, 2027–2036. https://doi.org/10.1534/g3.115.019877 (2015).
46. Huntley, M. A. & Clark, A. G. Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. *Mol. Biol. Evol.* **24**, 2598–2609. https://doi.org/10.1093/molbev/msm129 (2007).
47. 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. https://doi.org/10.1093/molbev/mst157 (2013).
48. Brown, C. J., Johnson, A. K., Dunker, A. K. & Daughdrill, G. W. Evolution and disorder. *Curr. Opin. Struct. Biol.* **21**, 441–446. https://doi.org/10.1016/j.sbi.2011.02.005 (2011).
49. Schuler, A. & Bornberg-Bauer, E. Evolution of protein domain repeats in metazoa. *Mol. Biol. Evol.* **33**, 3170–3182. https://doi.org/10.1093/molbev/msw194 (2016).

50. Basile, W., Sachenkova, O., Light, S. & Elofsson, A. High GC content causes orphan proteins to be intrinsically disordered. *PLoS Comput. Biol.* **13**, e1005375 (2017).
51. Peng, Z., Uversky, V. N. & Kurgan, L. Genes encoding intrinsic disorder in Eukaryota have high GC content. *Intrinsically Disord. Proteins* **4**, e1262225. https://doi.org/10.1080/21690707.2016.1262225 (2016).
52. Hatos, A. *et al.* DisProt: intrinsic protein disorder annotation in 2020. *Nucl. Acids Res.* **48**, D269–D276. https://doi.org/10.1093/nar/gkz975 (2020).
53. 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. https://doi.org/10.1371/journal.pone.0100615 (2014).
54. Su, W. C. & Harrison, P. M. Deep conservation of prion-like composition in the eukaryotic prion-former Pub1/Tia1 family and its relatives. *PeerJ* **8**, e9023. https://doi.org/10.7717/peerj.9023 (2020).
55. Harrison, P. M. fLPS: fast discovery of compositional biases for the protein universe. *BMC Bioinf.* **18**, 476. https://doi.org/10.1186/s12859-017-1906-3 (2017).
56. Campen, A. *et al.* TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.* **15**, 956–963. https://doi.org/10.2174/092986608785849164 (2008).
57. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res* **31**, 365–370 (2003).
58. Calhoun, S., Mondo, S. J. & Grigoriev, I. V. Yeasts and how they came to be. *Nat. Rev. Microbiol.* **17**, 649. https://doi.org/10.1038/s41579-019-0274-6 (2019).
59. Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucl. Acids Res* **42**, D699-704. https://doi.org/10.1093/nar/gkt1183 (2014).
60. 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. Genom.* **36**, 33–39. https://doi.org/10.1016/j.margen.2017.05.009 (2017).
61. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res.* **41**, D590–D596. https://doi.org/10.1093/nar/gks1219 (2012).
62. 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. https://doi.org/10.1093/sysbio/syq010 (2010).
63. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
64. 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 Ecol. Evol.* **8**, 28–36. https://doi.org/10.1111/2041-210x.12628 (2017).
65. 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. https://doi.org/10.1016/j.jmb.2004.02.002 (2004).
66. 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 (Oxford, England)* **21**, 3433–3434. https://doi.org/10.1093/bioinformatics/bti541 (2005).
67. Meng, F., Uversky, V. N. & Kurgan, L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell. Mol. Life Sci.* **74**, 3069–3090. https://doi.org/10.1007/s00018-017-2555-4 (2017).
68. 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. https://doi.org/10.3390/ijms160819040 (2015).
69. Harrison, P. M. Compositionally biased dark matter in the protein universe. *Proteomics* **18**, e1800069. https://doi.org/10.1002/pmic.201800069 (2018).
70. 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. *Nucl. Acids Res.* **42**, D304-309. https://doi.org/10.1093/nar/gkt1240 (2014).
71. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402 (1997).

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.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-89650-1.

Correspondence and requests for materials should be addressed to P.M.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.