Low Complexity Regions in Proteins and DNA are Poorly Correlated

Johanna M. Enright, Zachery W. Dickson ,* and G. Brian Golding*

Department of Biology, McMaster University, Hamilton, ON, Canada

***Corresponding authors:** E-mails: dicksoz@mcmaster.ca; golding@mcmaster.ca.

Associate editor: Banu Ozkan

Abstract

Low complexity sequences (LCRs) are well known within coding as well as non-coding sequences. A low complexity region within a protein must be encoded by the underlying DNA sequence. Here, we examine the relationship between the entropy of the protein sequence and that of the DNA sequence which encodes it. We show that they are poorly correlated whether starting with a low complexity region within the protein and comparing it to the corresponding sequence in the DNA or by finding a low complexity region within coding DNA and comparing it to the corresponding sequence in the protein. We show this is the case within the proteomes of five model organisms: *Homo sapiens*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana*. We also report a significant bias against mononucleic codons in LCR encoding sequences. By comparison with simulated proteomes, we show that highly repetitive LCRs may be explained by neutral, slippage-based evolution, but compositionally biased LCRs with cryptic repeats are not. We demonstrate that other biological biases and forces must be acting to create and maintain these LCRs. Uncovering these forces will improve our understanding of protein LCR evolution.

Key words: **low complexity, entropy, amino acid sequence, DNA sequence.**

Introduction

Low complexity regions (LCRs) are segments of a protein or DNA sequence which are biased in composition (Wootton and Federhen 1993). LCRs can present as periodic repeats, ambiguous cryptic repeats, or can contain no apparent pattern at all, but simply deviate from a randomized composition (Tautz et al. 1986; Wootton 1994a). LCRs contain low information and have a low entropy (Wootton and Federhen 1993). Entropy, as measured by Shannon's Entropy equation (Shannon 1948), is a measure of compositional complexity which uses the proportion of residue(s) in a subsequence to measure the compositional state of that subsequence (Wootton 1994b). A lower variety of residues would result in a lower entropy for that subsequence. Thus, minimal entropy would contain a subsequence consisting of only a single residue, whereas maximal entropy would contain all possible residues in an alphabet in equal proportions. In proteins, LCRs are typically composed of hydrophilic and small amino acid residues (Faux et al. 2005).

Interest in protein LCRs has grown in recent decades as involvement of LCRs in protein function and disease has been further illuminated. Due to a lack of motif conservation and a tendency to form non-globular protein domains, LCRs were once considered to be merely tolerated within their protein, offering no functional, biologic contribution (Huntley and Golding 2000, 2002). It is now believed that LCRs may offer a range of functions to various proteins, many which are linked to this non-globular, intrinsically disordered nature. Intrinsically disordered regions can allow for longer, more accessible protein domains, protein flexibility, and plasticity in molecular binding partners (Dosztányi et al. 2006; Ekman et al. 2006). As such, LCRs are often found in proteins involved in signaling pathways and can act as scaffolds in the formation of large protein complexes (Dyson and Wright 2005; Coletta et al. 2010). They are also enriched in transcription factors (Millard et al. 2020), developmental proteins (Huntley and Clark 2007) and can offer accessible regions for posttranslational modifications (Jeronimo et al. 2016; Monahan et al. 2017).

As protein LCRs are ultimately the result of changes to the underlying DNA, their evolution is likely similar to that of intergenic, non-coding DNA microsatellites (DePristo et al. 2006). Microsatellites are believed to evolve rapidly by expansion or contraction via two main mechanisms: the first and predominant mechanism being polymerase slippage, in which the DNA template and coding strand shift relative to one another and re-anneal with another repeat unit causing either insertion or deletion of a repeat unit (Levinson and Gutman 1987; Viguera et al. 2001). The second mechanism is unequal recombination which occurs via the misalignment of homologous repeat sequences during meiosis and results in the gain of repeats in the sequence of one chromosome and loss of repeats in other (Richard and Paques 2000). Other factors,

Article

including mismatch repair mechanisms (Levinson and Gutman 1987), repeat unit length (Schug et al. 1998), ability of the DNA to form structures (Moore et al. 1999; Dere et al. 2004; Murat et al. 2020), and repeat unit composition (Gragg et al. 2002) play a role in the rate of microsatellite slippage. Microsatellites above a certain threshold repeat length will undergo slippage and expand or contract, with longer microsatellites being more unstable and more likely to undergo slippage (Lai and Sun 2003). In coding regions, slippage of repeats whose units are multiples of three are more likely to be permitted compared with other repeat unit lengths because an insertion or deletion will not cause a frameshift mutation in the downstream coding sequence (Metzgar et al. 2000). Thus, they will be less likely to result in a deleterious mutation that will be selected against (Metzgar et al. 2000).

Codon homogeneity is also an important factor in LCR evolution. Because LCRs are believed to arise primarily via polymerase slippage, microsatellites of homogeneous codon runs are thought to be more unstable, evolve faster, and to be less conserved than sequences encoded by a heterogeneous mixture of synonymous codons (Albà et al. 1999). Over time, accumulation of synonymous mutations in LCR coding regions can help to conserve the LCR by breaking up repeats and reducing the chance of slippage (Albà et al. 1999).

With important physiological roles and high mutation rates, it follows that LCRs have the potential for pathogenesis. In humans, one of the most notable examples is Huntington's disease (Everett and Wood 2004). Slippage of long tracts of CAG trinucleotide repeats result in an expanded polyglutamine tract. The mutant Huntingtin protein develops a toxic gain of function effect within the cell (Everett and Wood 2004). Other neurodegenerative diseases resulting from trinucleotide repeat expansion include spinocerebellar ataxia and muscular dystrophy, encoding polyglutamine, and polyalanine tracts, respectively (Brown and Brown 2004; Everett and Wood 2004). As well, LCRs have been shown to contribute to antigenic variation and immune system evasion of human pathogens (Verstrepen et al. 2005; Velasco et al. 2013; Kebede et al. 2019).

Various studies have suggested that low entropy in nucleotide content correlates with LCRs in proteins. Li et al. (2015) showed how GC content constrains the types of amino acids which can be encoded, resulting in a bias towards amino acids encoded by codons with a high GC proportion and a bias against those with a lower GC proportion. The malaria parasite, *Plasmodium falciparum*, contains a high genomic AT content, which is strongly associated with the presence of protein LCRs, leading to preference for certain codons and amino acid types over others (DePristo et al. 2006). Xue and Forsdyke (2003) suggest that LCRs at the protein level are a result of AG content bias at the nucleotide level, and thus pressures at the DNA level can explain the presence of LCRs at the protein level. This was further supported by analyzing the nucleotide composition at first, second, and third codon bases, where AG content was higher in the first two, suggesting the importance of AG content to encode particular amino acids (Xue and Forsdyke 2003).

The structure of codons can significantly impact the entropy of the sequence they make up. Codons can be classified by the number of unique nucleotides in the codon, a property that we will refer to as nucleodicity. The DNA level entropy varies significantly among these codons: mononucleic codons such as AAA have an entropy of zero, dinucleic codons such as AGA have an entropy of 0.918, and trinucleic codons such as AGC have an entropy 1.58. However, this property can only affect entropy at the DNA level as the information is lost when the codon is translated to a single amino acid. LCRs with codons of different nucleodicities may evolve differently as the repeats have variable abilities to form secondary structure during transcription and translation (Barik 2017). This property is thought to be an influencing factor in the likelihood of polymerase slippage (Murat et al. 2020).

Polymerase slippage, such as that seen in microsatellite expansion, is suggestive of a neutral model of evolution whereby the unstable LCR is merely tolerated within the protein so long as it does not impart deleterious effects (Radó-Trilla and Albà 2012). The LCR can then be preferentially retained if it confers a selective advantage. This is in contrast to a strict selective model of LCR evolution which maintains that LCRs within a protein are a result of selective pressures constraining the types and ordering of amino acids so as to create an amino acid motif which confers a particular function (Haerty and Golding 2010).

There have been multiple studies supporting both neutral evolution as well as selective evolution, with LCRs being created due to forces acting on the protein/amino acid level. Evidence for selective neutrality includes a large variation in LCR tract size both intra and interspecifically (Haerty and Golding 2010). Such high length polymorphic variability is also associated with a homogeneous codon tract and high slippage rates (Mularoni et al. 2007). Whereas conservation of LCR motifs and selective evolution may entail a heterogeneous codon tract of synonymous codons. That is, assuming the codons were a result of pressure from the protein level and not due to the degeneration of trinucleotide repeats (Albà et al. 1999; Huntley and Golding 2006). Neutral proteins could contain a high variation in repeat tract size as a result of unstable replicative slippage and also could undergo non-synonymous mutations which would be permitted due to the lack of purifying selection (Mularoni et al. 2007). However, increased repeat length has been observed to correspond with low non-synonymous mutation rate, suggesting the conservation of long LCRs (Mularoni et al. 2007). Studies showing synonymous mutations closer to LCRs have indicated that these regions may be evolutionarily conserved and hold functional significance (Lenz et al. 2014).

In this study, we have identified LCRs in proteins and assessed the correlation between their entropy and their corresponding DNA sequence entropy. We also identified

LCRs in DNA sequences and compared their entropy to that of their corresponding amino acid sequence. If the origin and evolution of LCRs were primarily a result of mutation acting at the DNA level via polymerase slippage and LCR expansion being allowed due to low selective constraints, we would expect to see a high correlation between protein entropy and its corresponding coding sequence entropy in LCRs. As DNA entropy decreased, codon types would be constrained thereby constraining and lowering the protein entropy as well as increasing chances of polymerase slippage for further LCR generation. If selection was the predominant mechanism by which LCRs were formed, we would expect to see a lower correlation between DNA and protein sequence entropy of corresponding sequences. This is because selection, unlike slippage, would not necessarily favor a homogeneous run of codons, but could instead allow a more random collection of synonymous codons for a particular amino acid residue. Ultimately, this would allow for a wider range of possible DNA entropies given a particular protein LCR.

Materials and Methods

All custom scripts and commands used in this analysis can be found on `GitHub` at https://github.com/JohannaEnright/LCREntropyProject/.

Sequence Data

Two correlation studies of sequence entropy were conducted. The first identified LCRs in proteins from the entire proteome of five model organisms *S. cerevisiae*, *H. sapiens*, *A. thaliana*, *C. elegans*, and *D. melanogaster*. For each protein, we identified LCRs and compared the entropy in the corresponding coding DNA sequence. The second study did the inverse; identifying LCRs in the coding DNA sequences and compared their entropy to the entropy of the amino sequence that they encode. The genomes of the five organisms were downloaded from NCBI. Access dates and accession numbers are listed in supplementary table S1, Supplementary Material online.

Annotated sequences representing a haploid assembly for each organism were downloaded in genbank and fasta format. A custom python script was written to identify LCRs within coding sequences and locate their corresponding amino acid sequences and vice versa (Van Rossum and Drake 2009). LCRs were identified using the `Seg` algorithm (Wootton and Federhen 1993). Adjustments were made to `Seg` to account for alphabet size depending on the sequence type. Ambiguous characters were accounted for when identifying regions of low complexity by adding a fractional count to each residue represented by the ambiguous character. When searching for LCRs within proteins, `Seg` parameters were set to a window length (W) of 15, a trigger complexity (K1) of 1.9, and an extension complexity (K2) of 2.2 (see supplementary table S2, Supplementary Material online for alternate parameters examined). To identify LCRs in coding sequences, parameters were set to 45 for W (three times the length used for amino acids) 1.3 for K1, and 1.5 for K2 (see supplementary table S3, Supplementary Material online for alternate parameters examined).

A non-redundant set of coding sequences was selected by retaining only the longest isoform for each gene. This was done to reduce redundancy introduced by splice variants, duplicate genes in the pseudo-autosomal regions of the X and Y chromosomes, and duplicate genes present in alternate assemblies. If isoforms were the same length, the one which mapped to a chromosome was chosen over that from an alternate assembly. As well, an X chromosome isoform was chosen over a Y chromosome isoform. In addition, any coding sequences which encoded only a portion of a final protein product, such as immunoglobulin gene segments, or sequences which were not exactly three times the length of the amino acid sequence, were excluded as a direct mapping between amino acid sequence and coding sequence could not be made. For later simulations, the codon frequency, protein length, and proportion of proteins containing LCRs were calculated using this set.

For LCR analysis, only the longest LCR from each isoform was taken. We have observed in human data that less than 10% of LCR-containing proteins have multiple LCRs, and the composition of LCRs tends to be similar within the same protein. Taking the longest allows for a simpler analysis with one signal per protein. If multiple LCRs from a single sequence were the same length, the one with the lowest entropy was chosen. Once all LCRs were obtained, the entropy of the corresponding amino acid or DNA sequence was calculated using Shannon's Entropy equation (Shannon 1948):

$$H = \sum_{i=1}^{n} p_i \log_2 p_i, \quad (1)$$

where p_i refers to the proportion of each unique letters in a sequence and n refers to the total number of unique letters (eq. 1).

While calculating entropy, ambiguous characters were handled in the same manner as above. DNA LCRs were trimmed at the ends to ensure a direct correspondence between a codon and its amino acid. These end adjustments were taken into account before determining the longest LCR from a coding sequence.

Scatter plots were created for each organism and set of parameters. A linear regression and correlation coefficient calculation was performed for each plot in R (R Core Team 2022).

For the purpose of later simulations, the codons in the LCRs were classified by nucleodicity, the number of unique nucleotides in a codon. The nucleodicity of the LCR as whole was set to match the most frequent nucleodicity class amongst its codons. For example, an LCR made up of 6 AAA, 2 ATC, and 1 GCA would be assigned a nucleodicity of 1. In the case of a tie for the most frequent nucleodicity, an LCR would be partially assigned all tied nucleodicities. For example, an LCR with equal counts of

mono-, di-, and tri- nucleic codons would be counted as one-third for each class. Using these potentially partial counts the number of observed LCRs of each nucleodicity was counted. The number of expected LCRs for each nucleodicity class was calculated based on the codon usage for each organism by multiplying the total frequency for each class by the total number of observed LCRs. As an example, with completely unbiased codon usage, 6.6% ($\frac{4}{61}$) of LCRs would be expected be mononucleic, 57% ($\frac{35}{61}$) dinucleic, and 36% ($\frac{22}{61}$) trinucleic. The actual proportions vary between species based their codon usage.The significance of differences between observed and expected numbers of LCRs in each class was evaluated using a χ^2 test. A preference coefficient was calculated for each nucleodicity class to represent the observed preference for a codon class relative to the expected value. The coefficients are normalized relative to the most preferred codon class and are calculated as

$$P_{i,j} = \frac{O_{i,j}/E_{i,j}}{\max_k (O_{i,k}/E_{i,k})}, \quad (2)$$

where P is the preference coefficient, O is the observed number of LCRs, E is the expected number of LCRs, organisms are indexed by i organism, and the number of unique codons in a codon class is indexed by j and k.

LCR Simulations

It was critical to have null expectations to which to compare the biologically observed values for entropy and correlation. To that end, several simulations were implemented with the python language (Van Rossum and Drake 2009). Simulations were performed separately for each organism studied and for several models of evolution. In each simulation, a set of 100,000 equal length coding sequences were generated according to the relevant evolutionary model as well as the organism's codon usage. Each coding sequence had as many codons as the organism's average protein length (n) and a stop codon, for a total of $n + 1$ codons. The evolutionary models considered are intended to simulate varying levels of replication slippage, codon nucleodicity class, and substitution. Overall, five models were used: Null, Slip, Slip + CC, Slip + Syn, and Slip + CC + Syn.

The first model, Null, is the simplest and is intended as a naïve model. Each coding sequence was constructed by randomly sampling from the 61 amino acid encoding codons. Each codon had an equal probability of being sampled. This model does not generate a significant numbers of LCRs.

In the Slip model, codons are randomly selected according to each specific organism's genomic codon bias. However, once the same codon has been sampled at least twice the weighting for that codon is increased. As a result, runs of identical codons which encode LCRs are more likely. This elevated weighting is maintained until a different codon is selected at which point it returns to using the original genomic codon bias. The amount by which the probability is increased, the "slope," is dynamically set such that the overall proportion of LCR containing proteins in the generated proteome matches that observed in the organism. The increase in weight is applied each time the codon is consecutively sampled. As a result, the probability of slippage increases linearly with the length of the LCR. Slippage on the basis of codons was used instead of nucleotide-based slippage as there is strong selection against frameshift mutations in coding sequences (Metzgar et al. 2000).

There may be biological preferences for or against runs of identical codons in each nucleodicity class. Hence, the Slip+CC model generates sequences according to the Slip model and has a later additional step which attempts to mimic the species-specific use of nucleodicity class. After a protein was constructed, according to the Slip model, `Seg` is used to identify any LCRs. The LCR is classified by its nucleodicity and is retained with a probability equal to the organism-specific preference coefficient (See eq. 2). If a protein is not retained, a new protein is generated. This process continues until a proteome of 100,000 proteins with the same proportion of LCRs as observed biologically is constructed. In addition, this will generate a proteome which has LCRs in each nucleodicity class with the same proportions as is biologically observed.

As a final step, subsequent synonymous substitutions maintaining the amino acid sequence were simulated for the Slip model (Slip+Syn), and the Slip+CC model (Slip +CC+Syn). This was implemented by randomly selecting a codon within the previously simulated sequence. Any codon in the artificial protein could be selected, regardless of inclusion in an LCR. Then the first or third position nucleotide was randomly selected and randomly changed to any of the three other nucleotides with equal weight. This change was only accepted if the resulting codon was synonymous with the original. This process was repeated until 1,000 accepted synonymous mutations were made. Each attempt was completely independent of any previous iterations, therefore potential mutation sites were sampled with replacement. Differences in probability between transitions and transversions were not explicitly accounted for, however the nature of the genetic code forces more synonymous transitions than synonymous transversions (Koonin and Novozhilov 2009) since transitions are more likely to be synonymous than transversions. The process of adding 1,000 synonymous mutations was repeated for each of the proteins in the simulated proteome.

For all simulations, the same python script and `Seg` parameters described in the previous section were used to identify protein LCRs, calculate their entropy, and calculate the entropy of their corresponding coding sequences. The same was done for LCRs within coding sequences. See supplementary tables S4–S8, Supplementary Material online, for alternate parameters examined in the Null, Slip, and Slip+Syn simulations. Each pair of entropy values were plotted and a linear regression and correlation coefficient were calculated (R Core Team 2022).

Confidence Intervals for Correlation Coefficient

To determine if the entropy correlations were significantly different, 95% confidence intervals ($\alpha = 0.05$) for the correlation coefficient were calculated. A Fisher transformation (eq. 3) was first performed on the r values to improve normality with increasing sample size (David Shen 2006). The lower and upper confidence limits were then calculated (eq. 4) and these limits were transformed back (eq. 5) (David Shen 2006).

Calculations were performed using the following equations:

$$f_r = 0.5 \ln\left(\frac{1+r}{1-r}\right) \tag{3}$$

$$\begin{aligned} \zeta_l &= f_r - z_{(1-\alpha/2)}\sqrt{\frac{1}{n-3}} \\ \zeta_u &= f_r + z_{(1-\alpha/2)}\sqrt{\frac{1}{n-3}} \end{aligned} \tag{4}$$

$$\begin{aligned} r_l &= \tanh(\zeta_l) \\ r_u &= \tanh(\zeta_u). \end{aligned} \tag{5}$$

Identifying LCRs in Codons

LCRs at the codon level were identified and compared against the entropy of their encoded protein sequences. `Seg` was modified to be able to use an alphabet with 61 letters. `Seg` parameters used to identify codon LCRs were 15 for W, 2.5 for K1, and 2.9 for K2. All additional steps were performed as in the previous sections.

Identifying Periodic Repeats in LCRs

Mono-, di-, and tri- periodic repeats were identified at the protein level from the previously determined protein and DNA LCRs using a custom python script. Minimal repeat lengths for mono-, di-, and tri- repeats were 6, 5, and 4, respectively. LCRs which contained one or more of the three repeat types were classified as periodic LCRs, whereas LCRs which did not contain any of the three repeat types were classified as cryptic LCRs. Minimal repeat length parameters were varied to ensure consistent trends in sequence correlation for periodic repeats. Results for periodic LCRs and cryptic LCRs of alternate repeat lengths can be found at supplementary tables S9 and S10, Supplementary Material online, respectively.

Results

Entropy of LCRs in Protein and DNA Correlate Poorly with Corresponding Sequence Entropies

Protein and DNA sequence entropy comparisons were performed on the genome and proteome of five model organisms *Saccharomyces cerevisiae*, *Homo sapiens*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. In general, we observed a low correlation between corresponding sequence entropies when LCRs were identified in both coding regions and proteins. This lack of correlation was observed for all organisms, all of which had correlation coefficients at or below $r = 0.579$ for both LCR types. The low correlation suggests that DNA LCRs can encode a variety of amino acid sequence compositional complexities and that protein LCRs can be encoded by a mixture of nucleotide complexities and by a heterogeneous mixture of synonymous codons. To avoid being redundant, only the results from *S. cerevisiae* and *H. sapiens* will be described in detail. Corresponding results and figures for *A. thaliana*, *C. elegans*, and *D. melanogaster* can be found in the supplementary material at S1, S2, and S3, Supplementary Material online.

Despite the consistently low correlation between sequence entropies, there were significant differences in correlation coefficient values between LCR sequence types in some, but not all organisms. For example in *S. cerevisiae*, sequence entropy comparisons in protein LCRs yield a correlation coefficient of $r = 0.488$ (95% CI: 0.435–0.538; fig. 1A). The correlation between DNA LCRs and their corresponding protein sequences in *S. cerevisiae* was lower than for protein LCRs, although not significantly ($r = 0.428$, 95% CI: 0.281–0.555; fig. 2A). In *H. sapiens*, the correlation coefficient for sequences entropies between protein LCRs and DNA was $r = 0.374$ (95% CI: 0.347–0.400; fig. 1B). However, the correlation coefficient between DNA LCRs and their corresponding amino acid sequences was significantly higher at $r = 0.579$ (95% CI: 0.541–0.614; fig. 2B). A summary of results for all five organisms can be found in supplementary table S11, Supplementary Material online.

Next, we examined the general trends in the entropy distributions. The majority of protein and DNA LCRs have entropies near or within the low and high cut `Seg` parameter values and quickly taper off as they near more extreme entropies. Protein LCRs are encoded predominantly by higher entropy coding sequences with very few being encoded by low entropy coding sequences (fig. 1). Comparatively, DNA LCRs typically encode relatively midrange entropy protein sequences and are more evenly distributed within the possible protein sequence entropy range (fig. 2). This indicates that low entropy DNA sequences encode comparatively lower entropy protein sequences whereas, low entropy protein sequences can still be encoded by relatively high entropy DNA sequences. At the extremes of the distribution, a vertical line at a protein entropy of 0 was observed in protein and DNA LCRs of both species (figs. 1 and 2). This line corresponds to homopeptide repeats of a single amino acid residue which was evidently encoded by codons with various nucleotide compositions as well as a potential mix of heterogeneous, synonymous codons, hence the wide range of corresponding DNA entropies.

Other examples of extreme deviations include having high protein entropy but low DNA entropy or vice versa. Examples of both are shown in figure 3, and provide insight

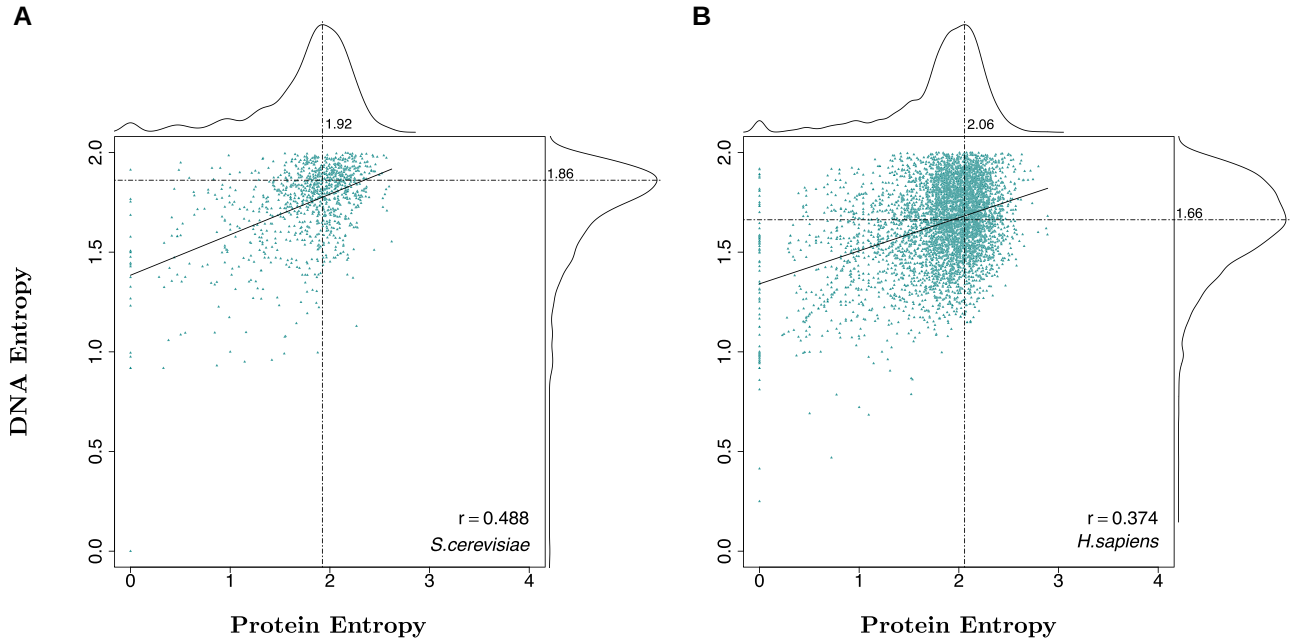

Fig. 1. Entropy comparisons of protein LCRs and corresponding sequences in the *S. cerevisiae* and *H. sapiens* genome. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. (*A*) 1,034 LCRs were identified from 6,016 protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of the corresponding coding sequences ($r = 0.488$). (*B*) 5,005 LCRs were identified from 133,689 protein sequences in *H. sapiens* and their entropies were plotted against the entropies of the corresponding coding sequences ($r = 0.374$).

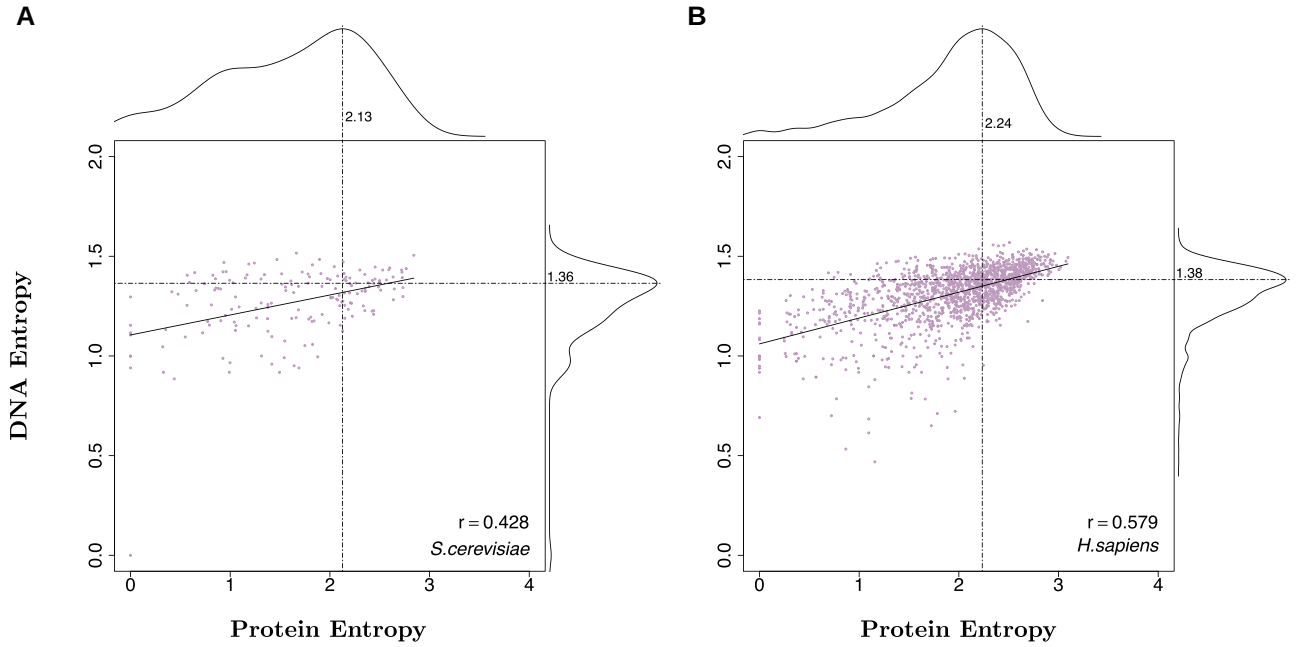

Fig. 2. Entropy comparisons of LCRs and corresponding sequences in the *S. cerevisiae* and *H. sapiens* genome. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. (*A*) 171 LCRs were identified from 6,016 coding sequences and their entropies were plotted against the entropies of the corresponding protein sequences ($r = 0.428$). (*B*) 1,571 DNA LCRs were identified from 133,689 coding sequences and their entropies were plotted against the entropies of the corresponding protein sequences ($r = 0.579$).

into the low sequence correlations observed. Protein LCRs with high protein entropy and unexpectedly low DNA entropy consist of amino acid residues whose codons share the same nucleotides and contain two or fewer different nucleotides (fig. 3). For example, the protein LCR with relatively higher entropy composed of R, E, and K are encoded by the codons AGA, AGG (R), GAG, GAA (E), and AAG, AAA (K), all of which share the nucleotides, A and/or

Protein LCRs

> NP_001369.1 NC_000002 REEKKRKEEERKKKE	$H = 1.52$	> NP_001369.1 NC_000002 AGAGAGGAAAAGAAGAGAAAAGAA GAAGAAAGGAAAAAAAAAGAA	$H = 0.867$
> NP_056007.1 NC_000020 DDDDDDDD	$H = 0$	> NP_056007.1 NC_000020 GACGACGACGATGATGATGATGAC	$H = 1.92$

DNA LCRs

> XP_024307650.1 NC_000020 GGTTTTTTGTTTTTTGTTTTTTTG GTTTTTGTTTGTTTGTTTTTT	$H = 0.722$	> XP_024307650.1 NC_000020 GFLFFVFLVFVCLFF	$H = 1.96$
> XP_011508832.1 NC_000002 CCACCGCCGCCGCCGCCCCCTCC TCCACCTCCTCCTCCCCCACCGC CCCCTCCGCCTCCTCCTCTC	$H = 1.28$	> XP_011508832.1 NC_000002 PPPPPPPPPPPPPPPPPPPPL	$H = 0.267$

Fig. 3. Example entropy comparisons from the opposing extremes among *H. sapiens* sequences, obtained from LCRs in protein sequences (Protein LCRs) and LCRs in DNA sequences (DNA LCRs). ***Protein LCRs)*** On the top, a relatively higher entropy protein LCR is encoded by a comparatively low entropy DNA sequence. On the bottom, an extremely low entropy protein LCR is encoded by a high entropy DNA sequence. ***DNA LCRs)*** On the top, a low entropy DNA LCR codes for a relatively high entropy protein sequence. On the bottom, a relatively higher entropy DNA LCR codes for a relatively low entropy protein sequence.

G. In contrast, low entropy protein regions with high entropy DNA sequences can be the result of a DNA sequence composed of distinct, but synonymous codons which are composed of three different nucleotides. In this case, the aspartic acid homopolymer encoded GAC and GAT. The DNA LCRs with low entropies and unexpectedly high protein entropies again tend to be encoded by codons which all share the same nucleotides in different rearrangements, but encode different amino acids. Additionally, few distinct, synonymous codons are used for each amino acid. In this case, the sequences are composed of: G (GGT), F (TTT), L (TTG), V (GTT), and C (TGT). In DNA LCRs with relatively high DNA entropy and comparatively low protein entropy, different residues are encoded by synonymous codons which often do not share the same nucleotides. Hence, the degree of codon homogeneity, the codon nucleodicity, as well as the potential for shared nucleotides between codons, all affect the degree of correlation between entropies of protein and DNA sequences in LCRs.

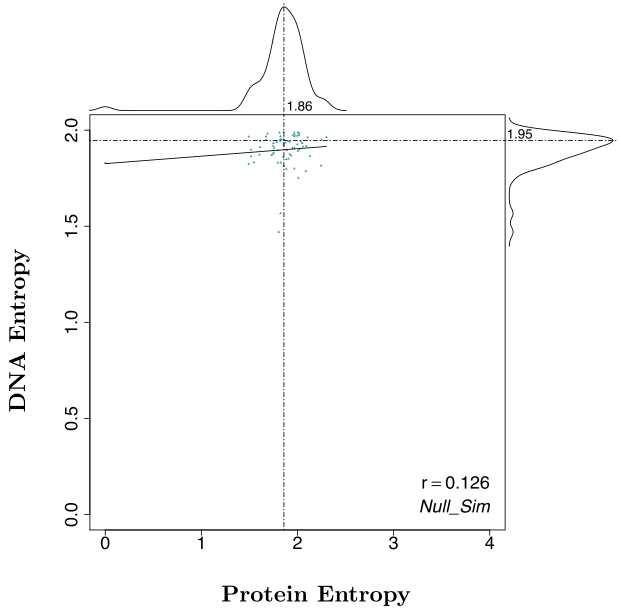

Fig. 4. Entropy of LCRs from the null simulated proteomes. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. 70 LCRs were identified from a sample of 100,000 protein sequences.

Using Slippage and Substitution Models to Compare and Explain Observed Entropy Correlations in Biological Sequences

To further examine the significance of the sequence entropy correlations, proteomes were simulated according to five different slippage and substitution models. The first proteome, generated according to the Null model, contained only 70 proteins with LCRs. The correlation between protein and coding sequence entropy for these LCRs was low at $r = 0.126$ (95% CI: −0.139–0.374; fig. 4). The correlation coefficient for the random simulation was lower than that in both *H. sapiens* and *S. cerevisiae*, although this difference was only significant in *S. cerevisiae*. There was only one DNA LCR which was identified from the 100,000 Null model DNA sequences. Due to the lack of data points, a linear regression could not be performed. The small number of LCRs identified from the Null model sequences in both proteins and DNA, as well as the low sequence entropy correlation in protein LCRs suggests, not surprisingly, that LCRs are not the sole result of

randomness in nature and are the result of some biological driving force. This is consistent with the literature which shows that LCRs are caused by replication error events like polymerase slippage which are exacerbated by an increase in repeat length (Levinson and Gutman 1987; Viguera et al. 2001; Lai and Sun 2003).

The second proteome, generated according to the Slip model, consisted of sequences with a propensity to form LCRs in a repeat length-dependent manner in an attempt to mimic LCR formation by DNA polymerase slippage. In general, the Slip model resulted in higher LCR sequence entropy correlations than in the biological LCRs. When looking at protein LCRs and their corresponding DNA sequences in the *S. cerevisiae* specific simulation, the correlation coefficient was significantly higher than for the *S. cerevisiae* biological sequences at $r = 0.566$ (95% CI: 0.555–0.577) (fig. 5A). DNA LCRs and their corresponding protein sequences in *S. cerevisiae* had a correlation coefficient of $r = 0.591$ (95% CI: 0.573–0.608) (fig. 5B) which was also significantly higher than the biological *S. cerevisiae* DNA LCRs. In the Slip simulation specific to *H. sapiens*, when looking at protein LCRs and their corresponding DNA sequences, the

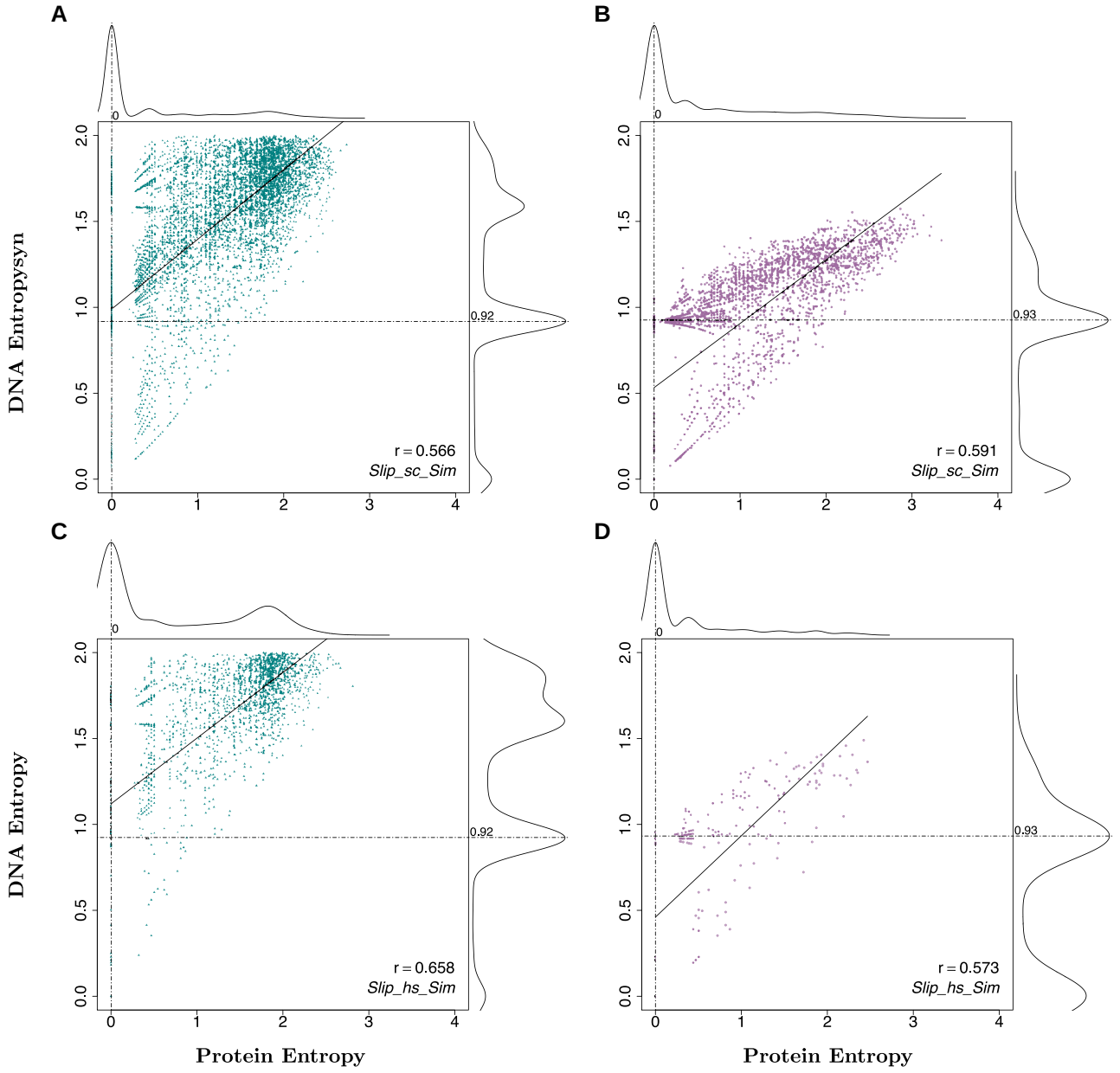

FIG. 5. Entropy of LCRs from the Slip simulated proteomes specific to *S. cerevisiae* and *H. sapiens*. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. LCRs were identified from a sample of 100,000 sequences. (*A*) 17,239 LCRs were identified from protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.566$). (*B*) 6,615 LCRs were identified from coding sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.591$). (*D*) 3,765 LCRs were identified from protein sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.658$). (*D*) 488 LCRs were identified from coding sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.573$).

correlation coefficient was significantly higher than for the biological protein LCRs in *H. sapiens* at $r = 0.658$ (95% CI: 0.637–0.678; fig. 5C). The correlation coefficient for the DNA LCRs and their corresponding protein sequences in the *H. sapiens* Slip model was slightly lower than the *H. sapiens* biological sequences although not significantly at $r = 0.573$ (95% CI: 0.503–0.636; fig. 5D). Overall, the higher correlations in the Slip model suggest that if LCRs were formed strictly in a neutral manner by DNA polymerase slippage, we would expect to see higher correlations in the biological sequences than what were actually observed.

To simulate conservation of the amino acid sequences, 1,000 synonymous mutations were added to the coding sequences from the Slip model, generating a third proteome, the Slip+Syn model. In general, implementing synonymous mutations into the coding sequences decreased the correlation compared with the Slip model. Correlations when going from the Slip model to the Slip+Syn model in the *S. cerevisiae* specific simulation were significantly lower for protein LCRs and their corresponding sequences ($r = 0.459$; 95% CI: 0.446–0.472; fig. 6A) as well as the reverse ($r = 0.517$; 95% CI: 0.489–0.544; fig. 6B). In the *H. sapiens* specific Slip+Syn

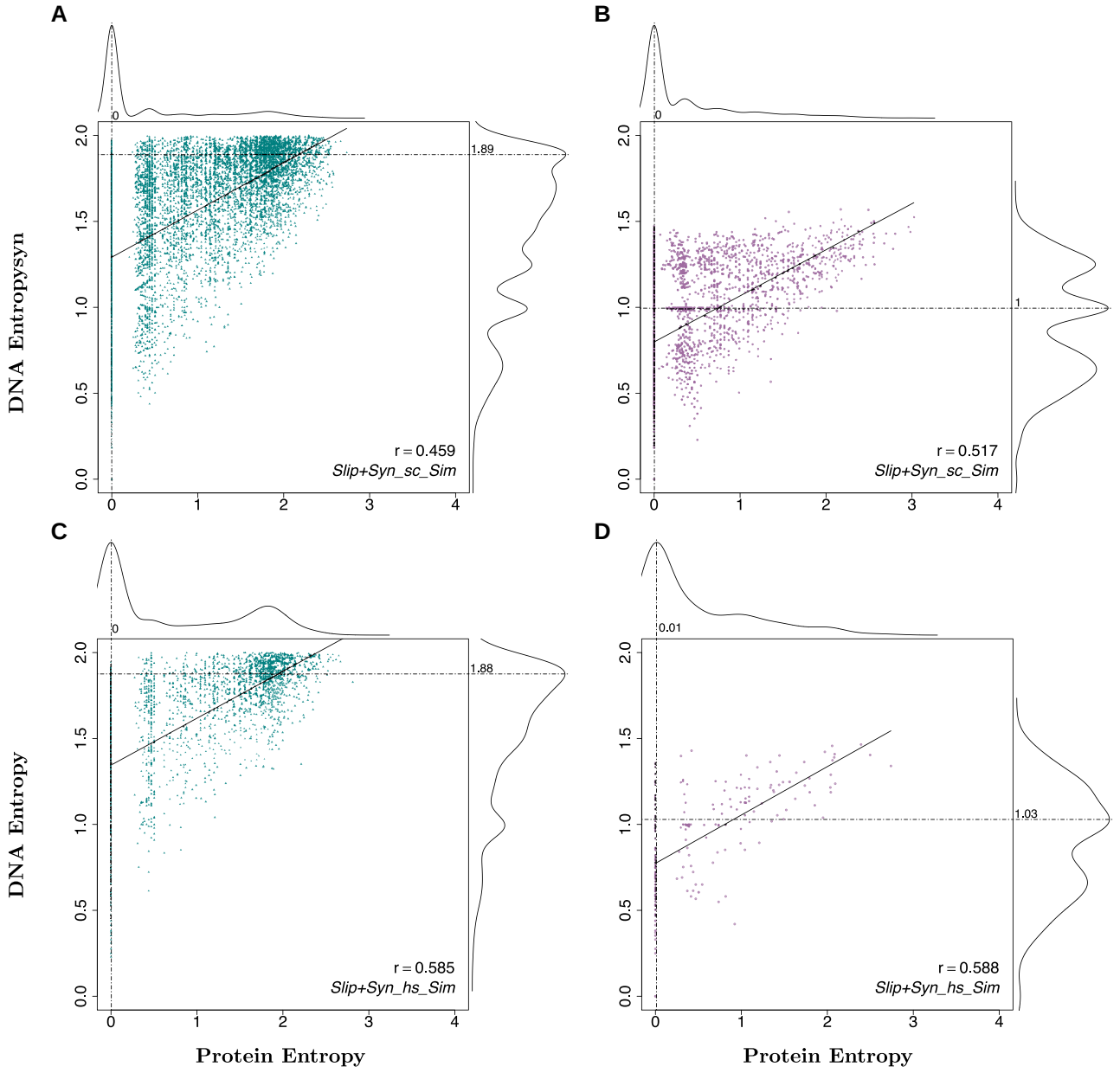

Fig. 6. Entropy of LCRs from the Slip+Syn simulated proteomes specific to *S. cerevisiae* and *H. sapiens*. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. LCRs were identified from a sample of 100,000 sequences. (*A*) 17,239 LCRs were identified from protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.459$). (*B*) 3,315 LCRs were identified from coding sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.517$). (*C*) 3,765 LCRs were identified from protein sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.585$). (*D*) 260 LCRs were identified from coding sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.588$).

simulation, protein LCRs and their corresponding sequences had a significantly lower correlation compared with the Slip model ($r = 0.585$; 95% CI: 0.561–0.608; fig. 6C). However, DNA LCRs and their corresponding sequences were slightly higher although this was not significant ($r = 0.588$; 95% CI: 0.492–0.670; fig. 6D). This helped confirm that we could expect LCR sequence entropy correlation to be lower if a homogeneous run of codons was broken up by synonymous mutations and the LCR had thus been conserved or selected for. When comparing Slip+Syn simulations to the biological sequences, there were varying results depending on the organism and LCR sequence type. For protein LCRs and corresponding sequences in *S. cerevisiae*, the correlation was insignificantly lower. For DNA LCRs and their encoded protein sequences, the correlation was insignificantly higher. In *H. sapiens*, the correlation for protein LCRs and corresponding sequences had a significantly higher correlation and DNA LCRs and corresponding sequences had an insignificantly higher correlation. Thus, the biological sequences have LCR sequence entropy correlation more similar to the LCRs generated from slippage followed by synonymous mutations although this model may still be limited in its ability to describe and predict LCR evolution.

The Slip and Slip+Syn simulations did have a greater tendency to generate mono-codon runs of LCRs as made evident by the large fraction of both DNA and protein LCRs with a protein entropy of 0. Since this trend was not observed in the biological sequences, this suggested that perhaps biology has a preference against runs of identical codons in LCRs. Thus, a preference for codon nucleodicity class was investigated in each organism. In the coding sequences for protein LCRs a strong nucleodicity class bias was observed for *S. cerevisiae* ($\chi^2 = 9.749 \times 10^{-30}$) and *H. sapiens* ($\chi^2 = 1.138 \times 10^{-280}$). Sequences in both organisms showed a greater number of codons with a nucleodicity of two and fewer codons with a nucleodicity of one and three. The fourth proteome, Slip+CC was generated taking this codon nucleodicity bias into account. Correlation coefficients from this model specific to *S. cerevisiae* were $r = 0.699$ (95% CI: 0.690–0.707; fig. 7A) for Protein LCRs and corresponding coding sequences. This was significantly higher than the corresponding correlation coefficient for the *S. cerevisiae* Slip simulation as well as the *S. cerevisiae* biological sequences. A correlation coefficient of $r = 0.553$ (95% CI: 0.534–0.572; fig. 7B) was observed for DNA LCRs and their encoded protein sequences which was significantly lower than in the corresponding Slip simulation and insignificantly higher than for the biological *S. cerevisiae* DNA LCRs. For the *H. sapiens* specific model, the correlation coefficient for protein LCRs and their corresponding sequences was $r = 0.808$ (95% CI: 0.795–0.820; fig. 7C). Again, this was significantly higher than in the corresponding Slip simulation as well as for the *H. sapiens* biological sequences. The correlation coefficient for DNA LCRs and their corresponding sequences was $r = 0.546$ (95% CI: 0.479–0.607; fig. 7D) which was insignificantly lower than in the Slip simulation as well as for the biological *H. sapiens* DNA LCRs. Thus, the Slip+CC simulation did not model the LCR sequences as anticipated and result in higher correlations for protein LCRs and their corresponding sequences but have less of a discernible effect on the correlation of DNA LCRs and their corresponding sequences.

Lastly, 1,000 synonymous mutations were added into the coding sequences from the Slip+CC simulations to produce a fifth proteome, Slip+CC+Syn. Similarly to the Slip and Slip +Syn simulations, the correlations from Slip+CC+Syn compared with Slip+CC were lower for both the *S. cerevisiae* and *H. sapiens* in both LCR sequence types. This was significant for all values except DNA LCRs and corresponding sequences in *H. sapiens*. In the *S. cerevisiae* specific version of this simulation, the correlation for protein LCRs and their corresponding sequences was close compared with the biological *S. cerevisiae* protein LCRs at $r = 0.490$ (95% CI: 0.477–0.503; fig. 8A). The correlation for DNA LCRs and their corresponding sequences was $r = 0.430$ (95% CI: 0.397–0.462; fig. 8B) which was also close to the biological *S. cerevisiae* DNA LCRs. For the *H. sapiens* specific simulation, the correlation between protein LCRs and their corresponding sequences was significantly higher than in the biological sequences at $r = 0.620$ (95% CI: 0.598–0.641; fig. 8C). The correlation between DNA LCRs and their corresponding sequences was insignificantly lower at $r = 0.536$ (95% CI: 0.446–0.615; fig. 8D). Thus, while this simulation seemed a good model for *S. cerevisiae*, this was not the case for *H. sapiens* suggesting length dependent slippage and incorporation of codon nucleodicity preferences followed by synonymous mutations is not sufficient to explain the correlations observed and therefore the mode of LCR evolution. Figure 9 summarizes the entropy correlations from the five simulations and compares them to the entropy correlations from the biological sequences. A summary table of the main results can be found in supplementary table S11, Supplementary Material online.

Comparing Correlations Between LCRs Categorized as Periodic or Cryptic Repeats in Biological Sequences

The original analysis of LCRs and corresponding sequences of the five model organisms looked at sequence entropy correlations of LCRs as a whole. However, some studies have suggested that different types of LCRs, particularly protein LCRs with tandem periodic amino acid repeats may evolve differently than cryptic repeat LCRs with periodic repeat LCRs being more likely to have evolved through DNA polymerase slippage (Battistuzzi et al. 2016). To test this theory, LCRs from the biological sequences were divided into two categories, those with periodic amino acid repeats and those without periodic amino acid repeats (cryptic repeats). Sequence entropy correlations between these two LCR classes were only investigated for protein LCRs as this seemed more biologically relevant for repeats at the amino acid level and because tri- or hexa- repeats at the DNA level would lead to repeats at the corresponding amino acid level and thus would likely not be informative of LCR evolution. For all organisms, LCR sequence entropy correlations were always significantly higher in periodic repeat LCRs compared with cryptic repeat LCRs (fig. 10). Correlations for LCRs with

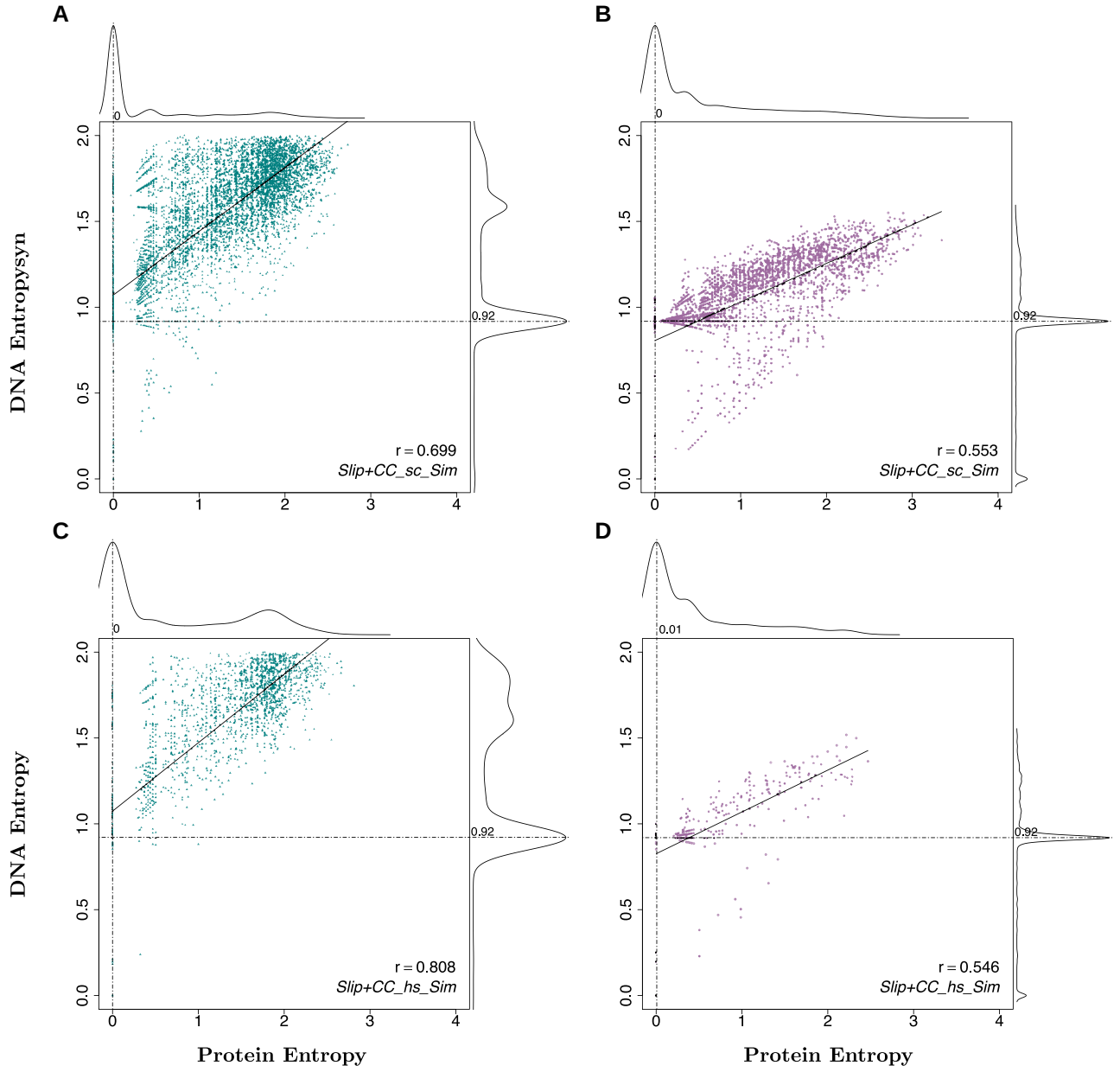

Fig. 7. Entropy of LCRs from the Slip+CC simulated proteomes specific to *S. cerevisiae* and *H. sapiens*. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. LCRs were identified from a sample of 100,000 sequences. (*A*) 17,255 LCRs were identified from protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.699$). (*B*) 6,505 LCRs were identified from coding sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.553$). (*C*) 3,767 LCRs were identified from protein sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.808$). (*D*) 579 LCRs were identified from coding sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.546$).

periodic repeats were also either higher or significantly higher than the correlation for both LCR types combined, and correlations in cryptic repeats were significantly lower than the correlation for both LCR types combined. For *S. cerevisiae*, the correlation between protein LCRs and corresponding sequences with periodic repeats and cryptic repeats was $r = 0.573$ (95% CI: 0.481–0.652) and $r = 0.322$ (95% CI: 0.248–0.392), respectively. For *H. sapiens*, the correlations for periodic repeats versus cryptic repeats were $r = 0.488$ (95% CI: 0.402–0.492) and $r = 0.242$ (95% CI: 0.207–0.277), respectfully (fig. 10). Overall, these results suggest that LCRs containing periodic amino acid repeats are more likely to evolve via DNA polymerase slippage whereas cryptic repeat LCRs are more likely to be selected for. Thus, of the five slippage and substitution models, the Slip simulation should be the most accurate model for the evolution of the periodic repeat LCRs. Overall, the Slip simulation did bear the closest resemblance in correlation to the biological sequences with the correlation being similar between Slip and periodic LCRs in *S. cerevisiae* but significantly higher compared with periodic LCRs in *H. sapiens* (fig. 10). On the contrary, the cryptic LCRs did

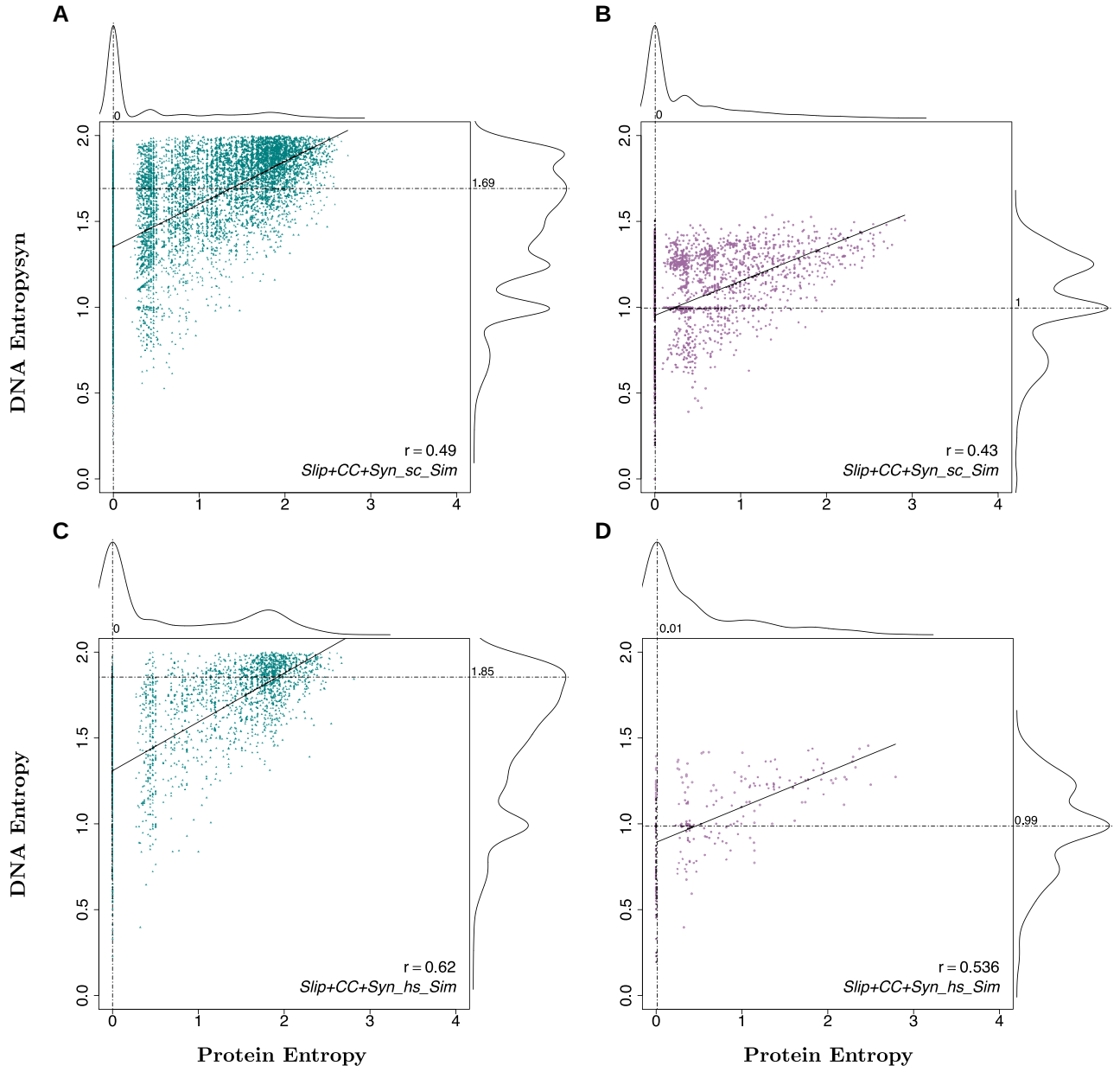

Fig. 8. Entropy of LCRs from the Slip+CC+Syn simulated proteomes specific to *S. cerevisiae* and *H. sapiens*. LCRs were identified from a sample of 100,000 sequences. Distributions of sequence entropies can be found along the vertical and horizontal axes with values indicating the mode. (*A*) 17,255 LCRs were identified from protein sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.490$). (*B*) 3,004 LCRs were identified from coding sequences in *S. cerevisiae* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.430$). (*C*) 3,767 LCRs were identified from protein sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding coding sequences ($r = 0.620$). (*D*) 340 LCRs were identified from coding sequences in *H. sapiens* and their entropies were plotted against the entropies of their corresponding protein sequences ($r = 0.536$).

not bear resemblance to any of the slippage or substitution models and were significantly below the correlations for all species-specific models in all organisms (fig. 10).

Discussion

At the outset, the fact that information flows from coding sequence to protein sequence might lead one to have a naïve expectation that the entropies of LCRs at each level would be highly correlated. However, for each of the genomes from the model organisms examined the correlations observed were low to moderate.

To ensure that the choice of parameters chosen for the measurement of LCRs via `Seg` was not the cause of this unusual effect, LCRs were identified using varying sets of parameters. To identify protein LCRs, parameters were chosen based on previous studies which found these parameters to work well for identifying highly repetitive LCRs while avoiding sequences that had higher complexity (Huntley and Golding 2002; Haerty and Golding 2010;

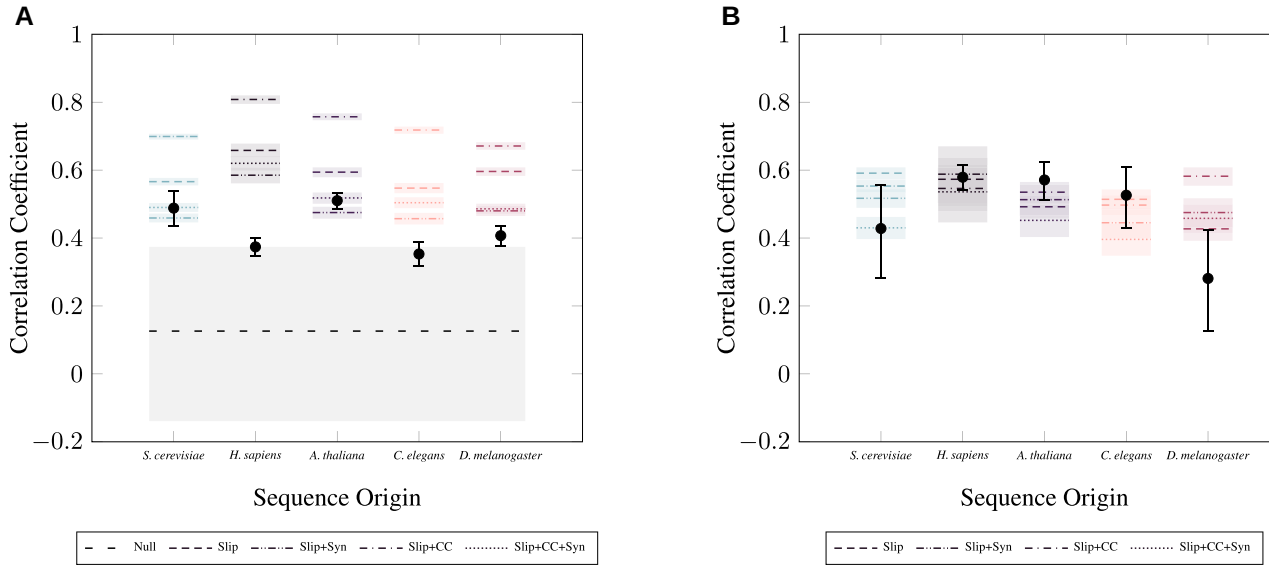

Fig. 9. A summary of the LCR entropy correlation coefficients and 95% confidence intervals for each model organism and species-specific simulated proteome. Exact correlation coefficient values can be found in supplementary table S11, Supplementary Material online. (*A*) Correlation coefficients are from protein LCRs and corresponding coding sequence linear regressions. The non-species-specific Null proteome is also included (the shaded area which spans all five species). (*B*) Correlation coefficients are from DNA LCRs and corresponding protein sequence linear regressions. A linear regression could not be performed for DNA LCRs from the Null simulation due to lack of data points.

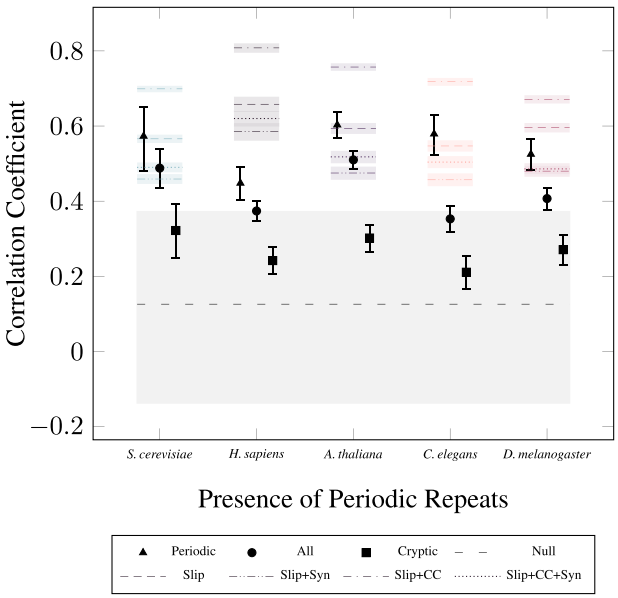

Fig. 10. A summary of the correlation coefficients with 95% confidence intervals for protein LCRs with periodic amino acid repeats, all LCR types combined, and cryptic repeats for the five model organisms. The correlations for the four species-specific slippage and substitution models and the Null model with 95% confidence intervals are also given for comparison. Exact correlation coefficient values can be found in supplementary table S11, Supplementary Material online.

Battistuzzi et al. 2016). These were then increased and decreased to explore the effect of the parameters as shown in supplementary table S2, Supplementary Material online. It was difficult to determine the biologically equivalent parameters for finding DNA LCRs as there are no known studies which have used `Seg` for DNA LCRs previously. We therefore chose to use the parameters suggested in the `Seg` manual and varied the parameters around these values as shown in supplementary table S3, Supplementary Material online. As can be seen in these tables, adjusting window length, low cut, and high cut parameters overall had little impact on the degree of correlation for the biological sequences (supplementary tables S2 and S3, Supplementary Material online). The value of the high cut parameter (*K*2) had the greatest impact on correlation. In general, and particularly in the DNA-based results, increasing the values of the three parameters (window length, low, and high cutoffs) resulted in higher correlation coefficients but at an extreme cost of many fewer regions considered to be low complexity. This could be because low entropies at long window lengths are less common, and higher low cut and high cut parameters results in less extreme LCRs being considered. A visual summary of the excess correlation can be found in supplementary fig. S4, Supplementary Material online.

The analysis of the simulated proteomes Null, Slip, and Slip+Syn were also performed with the `Seg` parameter sets described above (supplementary tables S4–S8, Supplementary Material online). Regardless of the set of parameters used, organism examined, and sequence type in which the LCRs were identified in, the results were qualitatively the same: The correlations observed in biological data were lower than the correlations produced in any simulation. In most cases, the biological correlations were significantly lower, however the exceptions were concentrated at the high complexity extreme of `Seg` parameters tested. For example, in the instances where the biological correlation was significantly higher than the

correlations in Slip and Slip+Syn for LCRs identified in proteins, all occurred at the settings with highest K2 values. K2 values much higher than K1 cause `Seg` to degenerate to finding the least probable subsequence in a protein regardless of entropy values. We observed that the excess correlation in the simulations was greatest when the definition of an LCR was strictest. This indicates that the evolutionary mechanisms embodied in the simulations: replication slippage as well as nucleodicity and synonymous substitutions insufficiently explain the observed distribution of protein LCRs, especially for highly repetitive LCRs.

Even when the correlations seen in the simulations were not significantly different from the biological sequences, the distribution of sequence entropies was very different. All of the simulations which included slippage produced lower entropy LCRs both at the DNA and protein level (figs. 5 and 7). Slippage produced far more mono-amino acid repeats than observed biologically, indicating that slippage alone doesn't explain the abundance of less compositionally biased LCRs. Non-synonymous mutations which break up the amino acid tract would shift the distribution away from homopolymers towards the more commonly observed entropies (figs. 6 and 8). The distribution of DNA entropies in simulations which did not include synonymous mutations was also biased towards lower entropies, but was also multimodal with peaks near the entropies corresponding to the nucleodicity of the plurality codon in the sequence (figs. 5 and 7). The addition of synonymous mutations brings the DNA entropy distribution more in line with what is seen biologically: unimodal with a peak at higher entropy (figs. 6 and 8). There may be a sample size effect, as we see the most dissimilarity between simulation and biology in *H. sapiens* which had the most LCR containing proteins at 5,005 while *S. cerevisiae* had only 1,034.

Comparing LCRs identified in proteins or in DNA coding sequences, the correlations for LCRs found in coding sequences were usually significantly higher in biological sequences. The exceptions are *S. cerevisiae* and *D. melanogaster* which both had the fewest LCRs identified in coding sequences at 176 and 187, respectively. The patterns are also unclear for the simulated proteomes: the Slip model often had higher correlation at the protein level, but the Slip+Syn model often showed the reverse. Without considering the bias in codon usage each organism has, correlations would be expected to be higher when identifying LCRs in coding sequences as some information is lost during the translation process. That is, the total entropy of the DNA sequence can never be lower than the protein sequence it encodes. The maximum entropy for a nucleotide is 2 bits, while the maximum entropy for an amino acid is roughly 4.32 bits. However, each amino acid is encoded by three nucleotides, for a maximum entropy of 6 bits. Thus, if the nucleotide variation is substantially constrained, as seen in DNA LCRs, the amino acids which can be encoded are limited to a select few. On the contrary, if the amino acid variation is limited, as seen with protein LCRs, there is still a possibility to have up to all four nucleotides comprising its coding sequence. Essentially, limiting DNA information content will limit protein information content, but the same is not necessarily true in the reverse direction. Hence, LCR sequences taken in one direction might be less correlated than LCR sequences taken in the other direction. Biological biases in codon usage, as well as the codons and amino acids tolerated in LCRs may modify this effect, and lead to the inconsistent pattern we observe.

It is interesting that DNA entropy for both protein and DNA LCRs rarely goes below one bit as there is evidence suggesting a bias toward the use of two nucleotides to drive particular codon usage which is thought to be associated with the presence of LCRs (Knight et al. 2001; Xue and Forsdyke 2003; Albà and Guigó 2004; DePristo et al. 2006; Li et al. 2015). In coding regions, there are rarely subsequences of DNA containing two or fewer nucleotides for 45 or more consecutive nucleotides. The percentage of protein LCRs with a corresponding DNA sequence entropy under 1 bit was 1.16% and 1.04% for *S. cerevisiae* and *H. sapiens*, respectively. None of the simulated proteomes had proportions as low, with the exception of the Null simulated proteome where no LCRs had DNA entropy less than 1. The Slip and Slip+CC models both produced proteomes where 31.6 (in *H. sapiens* Slip) to 53.2% (in *C. elegans* Slip+CC) of LCRs had DNA entropies less than 1.0, while the addition of synonymous mutations brought this proportion down to 14.6 (in *H. sapiens* Slip+Syn) to 21.2% (in *S. cerevisiae* Slip+Syn). Because the simulated values are much higher than observed in nature, it might indicate that biological sequences tend toward coding sequences with a higher variation of nucleotides than would be expected if both polymerase slippage resulting in codon repeats or a more heterogeneous mixture of codons was present.

It was possible that examining entropy at the codon level may have provided insights into LCR evolution. Each nucleotide triplet could be considered its own distinct character, especially considering that strong selection against frameshift mutations results in only replication slippage of whole codons being tolerated in coding regions (Metzgar et al. 2000). Because of this, we also calculated entropy at the codon level for each LCR. However, when comparing protein sequence entropy to codon entropy, there are mathematical constraints which force LCRs in the scatter plot to lie within a very constrained minimum and maximum threshold value, forcing a more linear relationship (supplementary fig. S5, Supplementary Material online). The maximal codon entropy for a homopolymeric sequence would occur if the amino acid had a six codon degeneracy at $\log_2 6$ which is 2.58 bits (supplementary fig. S5, Supplementary Material online). While the minimal codon entropy is zero for the repetitive usage of a single codon. Due to these tight mathematical constraints, entropy at the codon level was not further considered.

The mathematical constraints on possible entropies is the result of codon nucleodicity. Nucleodicity bias would therefore play a role in the correlation between DNA and protein level entropies. In all organisms examined,

there was a significant bias against the formation or maintenance of mononucleic codon repeats. This definitely impacts the correlation between protein and DNA sequence entropies, as exemplified in fig. 3. However, including this bias in the simulations increases the correlation between DNA and protein sequence entropies, while leaving the distribution of protein entropies largely unaffected. Therefore, the nucleodicity bias not only doesn't explain the low biological correlation, but seems to drive the correlation up. This indicates that other mechanisms must be acting as well.

One possible explanation for these low correlation results is that these LCRs are formed under a high selective pressure rather than through just polymerase slippage with low selective constraints. This result is surprising because there is a great deal of evidence to suggest that LCRs are the product of polymerase slippage at microsatellites, resulting in either the expansion or contraction of a repeat sequence (Levinson and Gutman 1987; Wierdl et al. 1997; Viguera et al. 2001; Tompa 2003; Hannan 2018). As well, evidence suggests that the polymorphic nature of LCRs is a result of this instability in combination with low selective pressure acting on either the protein as a whole, or this specific region of the protein (Fan and Chu 2007; Mularoni et al. 2007; Behura and Severson 2012). *In vitro* studies have also shown polymerase slippage can explain the observed microsatellite distributions within a genome (Madsen et al. 1993). Sequences consisting of pure codon repeats are more likely to undergo slippage than codon tracts with synonymous mutations. Such mutations break up trinucleotide repeats and have been shown to help stabilize and conserve LCRs (Albà et al. 1999). Still other studies suggest that LCR expansion is in an equilibrium between insertions which decrease tract stability and point mutations which increase tract stability (Kruglyak et al. 1998; Brandström and Ellegren 2008). The length of repeat has also been shown to have a positive correlation with the chance of slippage (Lai and Sun 2003). Together this led us to hypothesize that as protein LCRs came nearer to a perfect repeat, and as the length of an uninterrupted tandem amino acid repeat became longer, that the chance of it being a result of slippage and having a corresponding DNA sequence consisting of pure codons also would increase.

Of course, if LCRs were a product of high selective pressures forcing an irregularly biased amino acid composition, the length and biochemical properties of the amino acids at the protein level would be the major important factor in determining the repetitiveness and ordering of an LCR. This suggests that the codon choice at the DNA level would be unimportant and could consist of any codons so long as they encoded for the correct amino acid. In this case, variety in codon usage would increase, likely resulting in a greater variety of nucleotides used. Thus, a high DNA entropy could encode for a wide range of protein LCR entropies, ultimately resulting in a lower correlation between sequence entropies. The correlations from the Slip+CC proteome was significantly higher than those observed for the biological sequence comparisons in both *S. cerevisiae* and *H. sapiens* (fig. 9). If slippage were the predominant LCR driving force with low selective constraints, it would be expected that the correlation coefficients for these organisms would be closer to those observed from the simulations. Instead, sequence entropy correlations were significantly different from that observed int the Slip and Slip+Syn proteomes. As well, the decrease in correlation coefficient between Slip and Slip+CC suggests that if LCRs were created predominantly by polymerase slippage and contained more pure codon repeats, the DNA and protein sequence entropies would be more highly correlated than if they were a product of selection and contained a greater mixture of synonymous codons (fig. 9).

Overall, the selective retention or loss or an LCR may be dependent on the location of the LCR within the protein (Huntley and Clark 2007; Coletta et al. 2010), the LCR type (Kobe and Kajava 2001; Radó-Trilla et al. 2015), the protein function (Ekman et al. 2006; Coletta et al. 2010), and the organism itself (Karlin et al. 2002). Battistuzzi et al. (2016) and Zilversmit et al. (2010) suggest that LCR type and periodicity may constitute factors which affect the mode of LCR evolution. They propose that slippage may be a more prominent mechanism as an LCR gets closer to a perfect repeat, and selection may be more important for LCRs with a higher complexity and lower periodicity. This would make sense, as slippage is thought to occur at higher rates at longer continuous repeat sequences (Lai and Sun 2003; Leclercq et al. 2010). When looking only at highly repetitive LCRs, our analyzes agree. The results of the slippage-based simulations do closely resemble the biological sequences for LCRs which are mainly periodic repeats (fig. 10). The differences seen when looking at all LCRs are driven by the non-repetitive LCRs. These compositionally biased domains with a lack of periodicity (cryptic repeats) would be less likely to undergo slippage and their presence would be more reasonably attributed to selection. But this discounts any suggestion that the cryptic repeats were at one point a periodic repeat which degenerated over time (Radó-Trilla and Albà 2012). Similarly, Zilversmit et al. (2010) showed that in *P. falciparum* compositionally biased, aperiodic LCRs are less variable and evolve slower, whereas regions with long asparagine tracks are more variable and thought to evolve via replication slippage. We investigated the effect of periodicity on LCR entropy correlation using a range of minimum repeat lengths and found that when only sequences containing periodic repeats were compared, sequence entropy correlations were significantly higher in *H. sapiens* and higher, although not significantly, in *S. cerevisiae* compared with correlations of all LCR types combined (fig. 10). While correlations for periodic repeat LCRs were always significantly higher when comparing correlations for only cryptic repeat LCRs (fig. 10). The overall higher LCR sequence entropy correlation for periodic amino acid repeats is consistent with the findings of Battistuzzi et al. (2016) and Zilversmit et al. (2010).

The data presented here demonstrate that there is an unusually low correlation between the entropies of LCRs within proteins and their corresponding DNA coding sequences. This is largely driven by LCRs with cryptic rather than periodic

repeats. Although LCRs are thought to be primarily created via polymerase slippage, the simulations conducted suggest that the correlations would be higher if this were the sole mechanism. Continuing evolution of cryptic LCRs via synonymous substitutions cannot reduce the size the correlation and still maintain the size and entropy of the observed LCRs. Instead, the data suggest that these protein LCRs are maintained by genome wide, pervasive selection which acts to reduce the correlation by favoring synonymous substitutions that lower the correlation by lowering the repetitiveness of the LCRs at the DNA level and hence increasing the stability of the LCR. This may be partially facilitated through a bias against mononucleic codon repeats as we observe significantly fewer of these than codon usage frequencies would suggest. This would make the LCRs less prone to potentially deleterious slippage mutations. In the future, it is necessary to know if the correlations in sequence entropies change with the age of the proteins. If this hypothesis is true we would expect a higher correlation in more recently evolved proteins (Toll-Riera et al. 2012) compared with older, more highly conserved proteins. Future investigations should also determine how preferred amino acid residue, protein function, protein age, and role of the LCR within the protein influences the relation between protein and DNA sequence entropies in LCRs.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution online*.

Acknowledgements

We thank Natural Sciences and Engineering Research Counsel for funding for this project (grants RGPIN-202-05733 to G.B.G., PGSD3-547476-2020 to Z.W.D., and USRA-526761 to J.M.E).

Competing Interests

The authors declare there are no competing interests.

Funding Statement

This research was supported by Natural Sciences and Engineering Research Council (grants RGPIN-202-05733 to G.B.G., PGSD3-547476-2020 to Z.W.D., and USRA-526761 to J.M.E.).

Data Availability

The most up-to-date data, code, and supplemental information for this research is publicly available on `GitHub` at https://github.com/JohannaEnright/LCREntropyProject/

References

Albà M, Guigó R. 2004. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res*. **14**:549–554.

Albà M, Santibáñez-Koref M, Hancock J. 1999. Conservation of polyglutamine tract size between mice and humans depends on codon interruption. *Mol Biol Evol*. **16**:1641–1644.

Barik S. 2017. Amino acid repeats avert mRNA folding through conservative substitutions and synonymous codons, regardless of codon bias. *Heliyon*. **3**:e00492.

Battistuzzi F, Schneider K, Spencer M, Fisher D, Chaudhry S, Escalante A. 2016. Profiles of low complexity regions in *Apicomplexa*. *BMC Evol Biol*. **16**:47.

Behura S, Severson D. 2012. Genome-wide comparative analysis of simple sequence coding repeats among 25 insect species. *Gene* **504**:226–232.

Brandström M, Ellegren H. 2008. Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. *Genome Res*. **18**:881–887.

Brown L, Brown S. 2004. Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends Genet*. **20**: 51–58.

Coletta A, Pinney J, Solís D, Marsh J, Pettifer S, Attwood T. 2010. Low-complexity regions within protein sequences have position-dependent roles. *BMC Syst Biol*. **4**:43.

David Shen ZL. 2006. Computation of Correlation Coefficient and Its Confidence Interval in SAS. volume 170 of *SUGI 31*.

DePristo M, Zilversmit M, Hartl D. 2006. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene* **378**:19–30.

Dere R, Napierala M, Ranum L, Wells R. 2004. Hairpin structure-forming propensity of the (CCTG.CAGG) tetranucleotide repeats contributes to the genetic instability associated with myotonic dystrophy type 2. *J Biol Chem*. **279**:41715–41726.

Dosztányi Z, Chen J, Dunker A, Simon I, Tompa P. 2006. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res*. **5**:2985–2995.

Dyson H, Wright P. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. **6**:197–208.

Ekman D, Light S, Björklund A, Elofsson A. 2006. What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol*. **7**:R45.

Everett C, Wood N. 2004. Trinucleotide repeats and neurodegenerative disease. *Brain*. **127**:2385–2405.

Fan H, Chu J. 2007. A brief review of short tandem repeat mutation. *Genom Proteom Bioinform*. **5**:7–14.

Faux N, Bottomley S, Lesk A, Irving J, Morrison J, Whisstock J. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res*. **15**:537–551.

Gragg H, Harfe B, Jinks-Robertson S. 2002. Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in *Saccharomyces cerevisiae*. *Mol Cell Biol*. **22**:8756–8762.

Haerty W, Golding G. 2010. Low-complexity sequences and single amino acid repeats: not just junk peptide sequences. *Genome*. **53**:753–762.

Hannan A. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet*. **19**:286–298.

Huntley M, Clark A. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol Biol Evol*. **24**:2598–2609.

Huntley M, Golding G. 2000. Evolution of simple sequence in proteins. *J Mol Evol*. **51**:131–140.

Huntley M, Golding G. 2002. Simple sequences are rare in the Protein Data Bank. *Proteins*. **48**:134–140.

Huntley M, Golding G. 2006. Selection and slippage creating serine homopolymers. *Mol Biol Evol*. **23**:2017–2025.

Jeronimo C, Collin P, Robert F. 2016. The RNA polymerase II CTD: the increasing complexity of a low-complexity protein domain. *J Mol Biol*. **428**:2607–2622.

Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles A. 2002. Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci USA*. **99**:333–338.

Kebede A, Tadesse F, Feleke A, Golassa L, Gadisa E. 2019. Effect of low complexity regions within the PvMSP3alpha block II on the tertiary structure of the protein and implications to immune escape mechanisms. *BMC Struct Biol*. **19**:6.

Knight R, Freeland S, Landweber L. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol*. **2**:RESEARCH0010.

Kobe B, Kajava A. 2001. The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol*. **11**:725–732.

Koonin E, Novozhilov A. 2009. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*. **61**:99–111.

Kruglyak S, Durrett R, Schug M, Aquadro C. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci USA*. **95**:10774–10778.

Lai Y, Sun F. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol Biol Evol*. **20**:2123–2131.

Leclercq S, Rivals E, Jarne P. 2010. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome Biol Evol*. **2**:325–335.

Lenz C, Haerty W, Golding G. 2014. Increased substitution rates surrounding low-complexity regions within primate proteins. *Genome Biol Evol*. **6**:655–665.

Levinson G, Gutman G. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol*. **4**:203–221.

Li J, Zhou J, Wu Y, Yang S, Tian D. 2015. GC-content of synonymous codons profoundly influences amino acid usage. *G3 (Bethesda)*. **5**:2027–2036.

Madsen C, Ghivizzani S, Hauswirth W. 1993. In vivo and in vitro evidence for slipped mispairing in mammalian mitochondria. *Proc Natl Acad Sci USA*. **90**:7671–7675.

Metzgar D, Bytof J, Wills C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res*. **10**:72–80.

Millard P, Bugge K, Marabini R, Boomsma W, Burow M, Kragelund B. 2020. IDDomainSpotter: compositional bias reveals domains in long disordered protein regions-insights from transcription factors. *Protein Sci*. **29**:169–183.

Monahan Z, Ryan V, Janke A, Burke K, Rhoads S, Zerze G, O'Meally R, Dignon G, Conicella A, Zheng W, *et al*. 2017. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *EMBO J*. **36**:2951–2967.

Moore H, Greenwell P, Liu C, Arnheim N, Petes T. 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. *Proc Natl Acad Sci USA*. **96**:1504–1509.

Mularoni L, Veitia R, Albà M. 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics*. **89**:316–325.

Murat P, Guilbaud G, Sale J. 2020. DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. *Genome Biol*. **21**:209.

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

Radó-Trilla N, Arató K, Pegueroles C, Raya A, Albà M. 2015. Key role of amino acid repeat expansions in the functional diversification of duplicated transcription factors. *Mol Biol Evol*. **32**:2263–2272.

R Core Team. 2022. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richard G, Paques F. 2000. Mini- and microsatellite expansions: the recombination connection. *EMBO Rep*. **1**:122–126.

Schug M, Hutter C, Wetterstrand K, Gaudette M, Mackay T, Aquadro C. 1998. The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. *Mol Biol Evol*. **15**:1751–1760.

Shannon C. 1948. A mathematical theory of communication. *Bell Syst Tech J*. **27**:379–423, 623–656.

Tautz D, Trick M, Dover G. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature*. **322**:652–656.

Toll-Riera M, Radó-Trilla N, Martys F, Albà M. 2012. Role of low-complexity sequences in the formation of novel protein coding sequences. *Mol Biol Evol*. **29**:883–886.

Tompa P. 2003. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays*. **25**:847–855.

Van Rossum G, Drake FL. 2009. *Python 3 reference manual*. Scotts Valley (CA): CreateSpace.

Velasco A, Becerra A, Hernández-Morales R, Delaye L, Jiménez-Corona M, Ponce-de Leon S, Lazcano A. 2013. Low complexity regions (LCRs) contribute to the hypervariability of the HIV-1 gp120 protein. *J Theor Biol*. **338**:80–86.

Verstrepen K, Jansen A, Lewitter F, Fink G. 2005. Intragenic tandem repeats generate functional variability. *Nat Genet*. **37**:986–990.

Viguera E, Canceill D, Ehrlich S. 2001. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J*. **20**:2587–2595.

Wierdl M, Dominska M, Petes T. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics*. **146**:769–779.

Wootton J. 1994a. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem*. **18**:269–285.

Wootton J. 1994b. Sequences with "unusual" amino acid compositions. *Curr Opin Struct Biol*. **4**:413–421.

Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computers Chem*. **17**:149–163.

Xue H, Forsdyke D. 2003. Low-complexity segments in *Plasmodium falciparum* proteins are primarily nucleic acid level adaptations. *Mol Biochem Parasitol*. **128**:21–32.

Zilversmit M, Volkman S, DePristo M, Wirth D, Awadalla P, Hartl D. 2010. Low-complexity regions in *Plasmodium falciparum*: missing links in the evolution of an extreme genome. *Mol Biol Evol*. **27**:2198–2209.