RCPdb: An evolutionary classification and codon usage database for repeat-containing proteins

Noel G. Faux,[1,2,3] Gavin A. Huttley,[4] Khalid Mahmood,[1,2,3] Geoffrey I. Webb,[2,5] Maria Garcia de la Banda,[2,5,6] and James C. Whisstock[1,2,3,6]

[1]Protein Crystallography Unit, Department of Biochemistry and Molecular Biology, Monash University, Clayton Campus, Melbourne, Victoria 3800, Australia; [2]Victorian Bioinformatics Consortium, Monash University, Clayton Campus, Melbourne, Victoria 3800, Australia; [3]ARC Centre for Structural and Functional Microbial Genomics, Monash University, Clayton Campus, Melbourne, Victoria 3800, Australia; [4]John Curtin School of Medical Research, Australian National University, Canberra, Australian National Territory 0200, Australia; [5]School of Computer Science and Software Engineering, Monash University, Clayton Campus, Melbourne, Victoria 3800, Australia

Over 3% of human proteins contain single amino acid repeats (repeat-containing proteins, RCPs). Many repeats (homopeptides) localize to important proteins involved in transcription, and the expansion of certain repeats, in particular poly-Q and poly-A tracts, can also lead to the development of neurological diseases. Previous studies have suggested that the homopeptide makeup is a result of the presence of G+C-rich tracts in the encoding genes and that expansion occurs via replication slippage. Here, we have performed a large-scale genomic analysis of the variation of the genes encoding RCPs in 13 species and present these data in an online database (http://repeats.med.monash.edu.au/genetic_analysis/). This resource allows rapid comparison and analysis of RCPs, homopeptides, and their underlying genetic tracts across the eukaryotic species considered. We report three major findings. First, there is a bias for a small subset of codons being reiterated within homopeptides, and there is no G+C or A+T bias relative to the organism's transcriptome. Second, single base pair transversions from the homocodon are unusually common and may represent a mechanism of reducing the rate of homopeptide mutations. Third, homopeptides that are conserved across different species lie within regions that are under stronger purifying selection in contrast to nonconserved homopeptides.

[Supplemental material is available online at www.genome.org.]

Single amino acid repeats (homopeptides) are common in eukaryote proteins (4.3% of proteins in the GENPEPT database; Faux et al. 2005). The most common eukaryote homopeptides are poly-Q, poly-N, poly-A, and poly-E (Karlin et al. 2002; Faux et al. 2005). It is suggested that repeats are flexible regions important for mediating protein–protein interactions within large multiprotein complexes (Huntley and Golding 2002; Faux et al. 2005).

A key feature of trinucleotide repeats (TNRs) is their ability to undergo rapid changes in the number of repeat units during DNA replication and repair (Harding et al. 1992; Schlotterer and Tautz 1992; Tachida and Iizuka 1992; Richard et al. 2000; Ellegren 2002; Marcadier and Pearson 2003). As TNRs also represent codons, within the coding regions of DNA, homopeptides can thus change in length as a result of TNR length changes. Expansion and contraction of homopeptides can have positive as well as negative phenotypic effects. For example, repeat expansion and contraction within transcription factors is linked to major morphological changes in dogs (Fondon and Garner 2004). Conversely, expansion of the poly-Q tract within the large multidomain protein Huntingtin results in Huntington's Disease (HD) (OMIM 143100). Given the role of expansion and contraction of homopeptides in morphological variation and disease, two crucial questions arise: How are they controlled at the genetic level, and what are the evolutionary forces directing them? Regarding the former, studies to date have suggested that the G+C content of a genome underlies the frequency of amino acid repeats (Cocquet et al. 2003; Caburet et al. 2004), and that the accumulation of G or C at the third position of the codon may allow repeat expansion via replication slippage (Hancock et al. 2001). These studies, however, focus on organisms with G+C-rich transcriptomes. Regarding the latter question, Brock et al. (1999) have suggested that the local base composition and *cis*-elements are able to influence the expandability of homocodons (a codon repeat tract with a minimum length of three), and Rolfsmeier and Lahue (2000) have shown that interruption of homocodons dramatically reduces their rate of expansion.

In this study we comprehensively address these two questions by performing a detailed genomic analysis of all the genes encoding repeat-containing proteins (RCPs) of 13 eukaryotic species. To allow ready analysis of these data, we have developed an online searchable resource accessible at http://repeats.med.monash.edu.au/genetic_analysis/.

[6]Corresponding authors.
E-mail Maria.GarciadelaBanda@infotech.monash.edu.au; fax 61 3 9905 4699.
E-mail James.Whisstock@med.monash.edu.au; fax 61 3 9905 4699.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6255407. Freely available online through the *Genome Research* Open Access option.

Results

Data sets

To avoid functional bias and allow complete analysis of the genes encoding the RCPs, we analyzed the RCPs encoding genes from

17:1118–1127 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

species whose genomes have been completely sequenced. Moreover, to avoid any G+C content bias we included species whose transcriptomes are either G+C-rich (e.g., *Homo sapiens*) or A+T-rich (e.g., *Plasmodium falciparum*). In all, we analyzed the encoding RCP genes from 13 species: three mammals, two fish, a bird, three insects, a roundworm, baker's yeast, malaria, and a plant (see Methods for complete list). Here, we considered homopeptides as a run of seven or more identical amino acids (Faux et al. 2005).

Codon usage inside and outside of homopeptide repeats

As a general rule, within a particular species the most common codon that encodes a particular amino acid is also the most common codon involved in a homopeptide of the same amino acid. The codon usages inside a homopeptide, outside of the homopeptide, and within the species' transcriptome are all significantly different from each other (χ^2 $P << 0.001$ for all species, after correcting for multiple tests). This is clearly seen in *H. sapiens* and *P. falciparum* (Fig. 1A,B). The difference between the codon usage outside of the homopeptide and for the species' transcriptome is not unexpected, as RCPs are functionally biased (Karlin et al. 2002; Faux et al. 2005). Furthermore, the observed difference in codon usage is due to an overrepresentation of a small number of codons encoding homopeptides as indicated from the analysis of the Pearson's residuals from the χ^2 test between homopeptide codon usage and the transcriptome (data not shown).

If mutation was the sole agent generating and maintaining TNRs, we would expect equal frequencies of trinucleotides and their reverse complements in TNRs. This expectation was true for all species' genomic TNRs, which are predominantly drawn from non-protein-coding sequences; e.g., GAG and its complementary strand counterpart CTC occur at similar frequencies in *H. sapiens*. However, the expectation did not hold for the subset of TNRs that are homocodons, identifying a likely influence of natural selection. For instance, GAG (encoding E) is ~4.4 times more frequent than CTC (encoding L) within a homocodon in *H. sapiens* (see http://repeats.med.monash.edu.au/genetic_analysis/). This difference correlates with the frequency of E and L homopeptides (403 and 187, respectively). The difference in the frequency of trinucleotides in TNRs and homocodons was significant after correcting for multiple tests (χ^2 tests, $P << 0.001$) for all species. These data indicate that the interplay between selection and mutation is directing the types and frequency of homocodons in RCPs.

Previous studies of RCPs from *H. sapiens*, *Mus musculus*, *Rattus norvegicus*, and *Danio rerio* revealed an elevated G+C level,

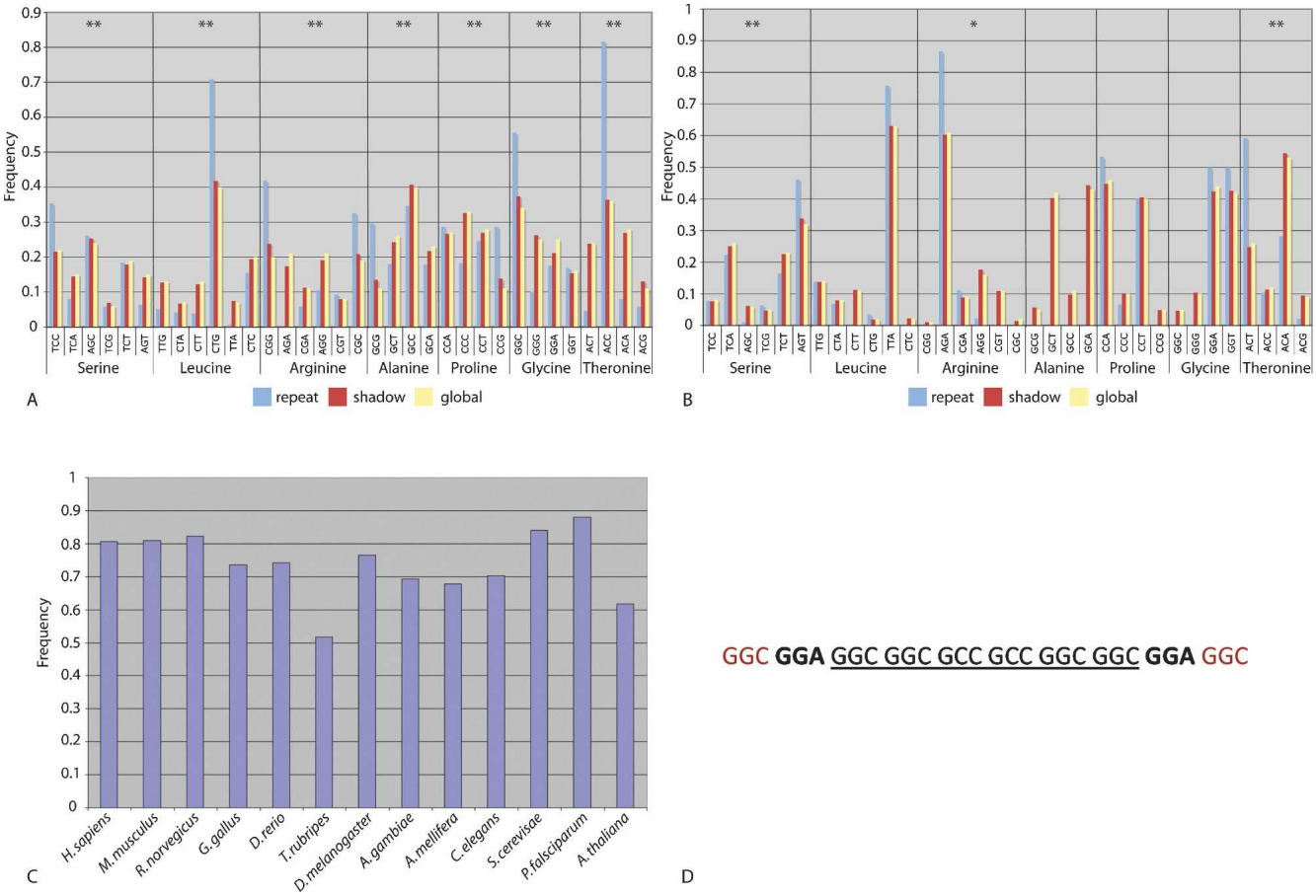

Figure 1. (*A*,*B*) Codon usage within a homopeptide (repeat), outside the homopeptide (shadow), and from the codon usage database (global). (*A*) *H. sapiens*, (*B*) *P. falciparum*. Significant differences between the codon usage within a homopeptide compared with the global codon usage: (*) $P \leq 0.01$, (**) $P \leq 0.001$. (*C*) The frequency of codons within a homopeptide that are immediately upstream and downstream of the 5′ and 3′ codons (the capping codons), respectively, which surround a homocodon and are synonymous to the homocodon. (*D*) An example of the encoding codons of a homopeptide, highlighting the codons mentioned in *C*. (Underlined) The homocodon, (red) the 5′ and 3′ capping codons, (bold) the immediate upstream and downstream codons of the capping codons.

leading to the suggestion that this elevation may be a vertebrate phenomenon (Cocquet et al. 2003; Alba and Guigo 2004; Veitia 2004; Siwach et al. 2006). However, this suggestion is not supported by our results. Firstly, there is no bias toward a specific species lineage, as assessed by a two-sided sign-test ($P = 1.0$) comparing the number of species that have an elevated G+C content outside of the homopeptide (in the RCP's encoding gene), with those that do not (seven of the 13 species). Secondly, we found that *D. rerio*, *Takifugu rubripes*, and *Gallus gallus* RCPs have reduced G+C content (Table 1). Furthermore, for those homopeptides that could be encoded by either G+C- or A+T-rich codons (S, L, R, T, Q, D, E, H, C, and V), there is no bias present that is not already present in the species' transcriptome (Table 2). This is consistent with the previous suggestion of Karlin et al. (2002). To further support this argument, within the A+T-rich transcriptome of *P. falciparum* there are no instances of an amino acid that is globally encoded by an A+T-rich codon being predominantly encoded by a G+C-rich codon within the repeat tract (see Supplemental Table 1 and http://repeats.med.monash.edu.au/genetic_analysis/ for this and other examples).

The RCPdb Web site (http://repeats.med.monash.edu.au/genetic_analysis/) facilitates comparison of codon usage and the TNR statistics. In particular, the user is able to choose which species they wish to compare, and is able to hide the homopeptides (codon sets) that are not of interest (Supplemental Fig. 1A).

Reiteration of identical codons (homocodons) within homopeptides

We examined the prevalence of homocodons within homopeptides, since long tracts of identical codons may suggest an underlying genetic event such as slippage. More than 55% of all amino acid repeats (across all species) contain a homocodon (Table 2). Of particular note, 87% of *P. falciparum* homopeptides contain a homocodon longer than four codons. If homopeptides originated by a mutagenic process (such as replication slippage), we would expect them to be comprised primarily of homocodon runs. These homocodons would be expected to be longer than the homocodons that have arisen purely by chance (based on the codon frequency outside of the homopeptide). We assessed this hypothesis via a runs test (see Methods), which showed that the homocodons present in the homopeptides were longer than expected purely by chance ($P << 0.001$ for all species, after correcting for multiple tests; Supplemental Table 2). These data, along with the χ^2 tests for the codon usage within a homocodon compared with the species transcriptome, imply that homopeptides, generally, contain a homocodon that is the most prevalent codon used for that homopeptide.

Table 1. The G+C content of each species' RCPs, regions of the RCPs excluding the homopeptide, and the transcriptome as a whole

Species	Median G+C content (whole RCPs)	Median G+C content (outside the homopeptide)	CUD G+C content[a]
H. sapiens	0.57	0.56	0.52
M. musculus	0.55	0.54	0.52
R. norvegicus	0.55	0.54	0.52
G. gallus	0.51	0.50	0.51
D. rerio	0.51	0.50	0.51
T. rubripes	0.56	0.54	0.55
D. melanogaster	0.57	0.55	0.54
A. gambiae	0.59	0.57	0.56
A. mellifera	0.50	0.48	0.43
S. cerevisae	0.40	0.39	0.40
C. elegans	0.46	0.44	0.43
A. thaliana	0.45	0.44	0.45
P. falciparum	0.22	0.22	0.24

Cells shaded in dark gray show an increase in G+C content, and cells shaded light gray show a decrease in G+C content by 0.01 compared with the species transcriptome.
[a]Values are from the codon usage database (CUD), http://www.kazusa.or.jp/codon/.

Table 2. The percentage of homopeptides that contain at least one or two codon reiterants longer than two, three, or four

	At least one			At least two	
Species	Length > 2	Length > 3	Length > 4	Length > 2	Length > 3
H. sapiens	77	56	38	19	6.8
M. musculus	76	53	36	21	6.6
R. norvegicus	70	46	30	18	5.2
G. gallus	70	44	28	17	3.4
D. rerio	76	52	29	22	5.5
T. rubripes	55	24	13	5.8	0.9
D. melanogaster	78	50	31	23	5.7
A. gambiae	74	53	35	10	1.5
A. mellifera	57	37	21	14	3.3
S. cerevisae	76	55	45	21	9.1
C. elegans	64	35	18	7.6	8.5
A. thaliana	84	63	45	13	3.0
P. falciparum	98	93	87	17	8.9

Cells shaded in gray have >50% of the repeats containing one or two codon reiterants.

Codon makeup within homopeptides

Despite the tendency for homopeptides to be encoded primarily by one codon, most homopeptides are not homogeneous in regard to their codon makeup (Table 3). Together with the above observation, these data suggest that the heterogenous codon makeup of RCPs is a result of both strand slippage and point mutations (Supplemental Table 2). Such repeat structures are also evident in noncoding repeats, where strand slippage is considered the primary mutational mechanism and point mutation the mechanism responsible for disrupting these repeats (Weber 1990; Mirkin 2006).

Globally across all species, for homopeptides that have greater than two encoding codons, the 5′ and 3′ capping codons (the codons directly 5′ and 3′ of a homocodon, Fig. 1D) are predominantly a single base transversion from the adjacent homocodon (R ↔ Y, Supplemental Table 3). However, there are a number of exceptions. For example, in serine homopeptides, the 5′ and 3′ codons in the warm-blooded species are predominantly transitions (R ↔ R or Y ↔ Y) (Supplemental Table 3). To investigate if these capping codons are potentially interrupting a longer homocodon, we analyzed the frequency that the next codon upstream or downstream from the capping codons was the same as the homocodon (Fig. 1D). We found that the majority of these codons are identical to the homocodon (Fig. 1C). We also investigated the prevalence of heterodi- and heterotri-codon repeats, to assess the level of higher order repeats. We found no major bias toward such repeat tracts (Supplemental Table 4). Together, these data reveal that most homopeptides are comprised of a homocodon interrupted by a single base transversion.

These data are easily accessible and comparable across ho-

Table 3. The frequency of the amino acid repeats encoded by a single codon

Species	Frequency
H. sapiens	0.089
M. musculus	0.075
R. norvegicus	0.068
G. gallus	0.073
D. rerio	0.040
T. rubripes	0.025
D. melanogaster	0.046
A. gambiae	0.106
A. mellifera	0.064
S. cerevisae	0.124
C. elegans	0.035
P. falciparum	0.550
A. thaliana	0.161

mopeptides and species on the online database, via the collapsible rows in the results table and the species checkboxes in the navigation menu (Supplemental Fig. 1B).

Selective pressures affecting RCPs

To investigate the evolutionary selective pressures, as measured by the ratio (ω) of nonsynonymous base substitutions (Ka) to synonymous base substitutions (K_s; i.e., $K_a/K_s = \omega$) that have affected homopeptides, we identified the *H. sapiens*, *M. musculus*, and *R. norvegicus* orthologs of each RCP. The amino acid sequences were aligned using Clustalw (Thompson et al. 1994), and the inferred gaps from the protein alignment were introduced into the encoding genes to ensure gaps in the DNA sequence alignment were placed respectful of codon boundaries and in multiples of three. These alignments were then broken into three groups based on the conservation of the homopeptides in these proteins (Supplemental Fig. 2). Group A: proteins that contain only conserved homopeptides (i.e., all homopeptides are present across all three species in the same positions; 314 proteins); Group B: proteins that contain only nonconserved homopeptides (1129 proteins); and Group C: proteins that contain a mixture of conserved and nonconserved homopeptides (86 proteins). For each group, we considered the selective pressure on three local regions: (1) The homopeptide itself (*r*); (2) the regions immediately surrounding the repeat "flanks" (*f*; defined as the 33 codons 5′ and 3′ of the homopeptide, truncated if a neighboring homopeptide is closer or at the 5′ and 3′ ends of the gene); and (3) the rest of the gene, or "shadow" (*s*) (Supplemental Fig. 3). The details of the statistical analyses of these regions are presented in the Methods section. We note here that the estimates of ω for the repeats (ω_r) should be viewed as unreliable; however, the parameter has little effect on statistical inference regarding the importance of separate ω parameters for the flanks (ω_f) and shadows (ω_s), as ω_r is a common parameter for the null and alternate hypotheses. Inclusion of ω_r improves accuracy of the ω_f and ω_s estimates (see the Methods section for a more detailed explanation).

Our analysis reveals that each of the local regions *r*, *s*, and *f* in groups A and B evolve under distinct evolutionary pressures, with *s* being under stronger selective pressure than *f* (likelihood tests showed statistical significance after correcting for multiple tests, $P << 0.001$). The selective pressure differences between *r* and the regions *s* and *f* most likely reflect the compositional difference of *r* compared with *s* and *f*, while the difference between *s* and *f* putatively is due to their evolutionary distinctiveness. Analysis of the distribution of the ω_f and ω_s for those genes that showed a nominally significant ($P \leq 0.05$) difference between the estimates for ω_f and ω_s (Fig. 2A) revealed that the ω_f of group A was significantly smaller than the ω_f of group B ($P << 0.001$, two-sample Kolmogorov–Smirnov test). Further, there was no significant difference between the distributions of ω_s of groups A and B ($P = 0.85$, two-sample Kolmogorov–Smirnov test, Fig. 2B), and the analysis of the distribution of the ratio ω_f/ω_s for both groups A and B (Fig. 2C) showed that the pattern of evolution of homopeptides and their flanks are related, with $\omega_f/\omega_s < 1$ for flanks adjacent to conserved homopeptides (group A) and >1 for flanks adjacent to nonconserved homopeptides (group B), indicating that the selective constraints applied upon the shadows of the genes in both groups A and B are similar but the flanks of group A are under greater purifying selection than those of group B.

Given that the flanks of conserved homopeptides in group A are under greater selective pressure than those of nonconserved homopeptides in group B (i.e., $\omega_{f(A)} < \omega_{f(B)}$), we investigated whether such a trend was apparent within group C (proteins that contain both conserved, *con*, and nonconserved, *noncon*, homopeptides). For this class, we further divided *f* into two classes f_{con} (conserved) and f_{noncon} (nonconserved) flanks. The significant ($P << 0.001$) result of the likelihood ratio test indicated that these two classes are under different selective pressures, and the analysis of the distribution of the ω values for the flanks and shadows, for the nominally significant results, confirms that the significant result observed was derived from the greater constraints applied upon the flanks adjacent to the conserved homopeptides (Fig. 2D). Thus, together these data show that the flanks of conserved homopeptides are generally under stronger selective pressure than the flanks of nonconserved homopeptides.

The protein and DNA alignments, the ω estimates, and the individual results of the hierarchical hypothesis tests are all available on the Web site. The user is able to search for specific proteins based on keywords or the refseq accession of the human, rat, or mouse protein. The user is also able to limit the search by the types of homopeptides present in the RCPs and by the conservation patterns as defined above (Supplemental Fig. 1C).

Biological process and functional distribution between RCPs containing solely conserved or nonconserved repeats across rodents and human

Fondon and Garner (2004) hypothesized that repeat expansion and contraction can provide a mechanism for rapid morphological evolutionary changes and that this process may be particularly important in those genes that are involved in development (such as transcription factors). Comparison of the rank order of the biological process and the functional classifications between groups A and B did not reveal any such bias (Fig. 3; Supplemental Tables 5, 6).

Discussion

We have developed a Web site for analysis of the genes encoding homopeptides (http://repeats.med.monash.edu.au/genetic_analysis/). The Web site allows investigators to easily search for, compare, and contrast RCPs from 13 different species. Furthermore, the Web site enables detailed comparison of the selective pressures affecting mammalian RCPs. We aim to extend this database to include other species as their genomes are com-

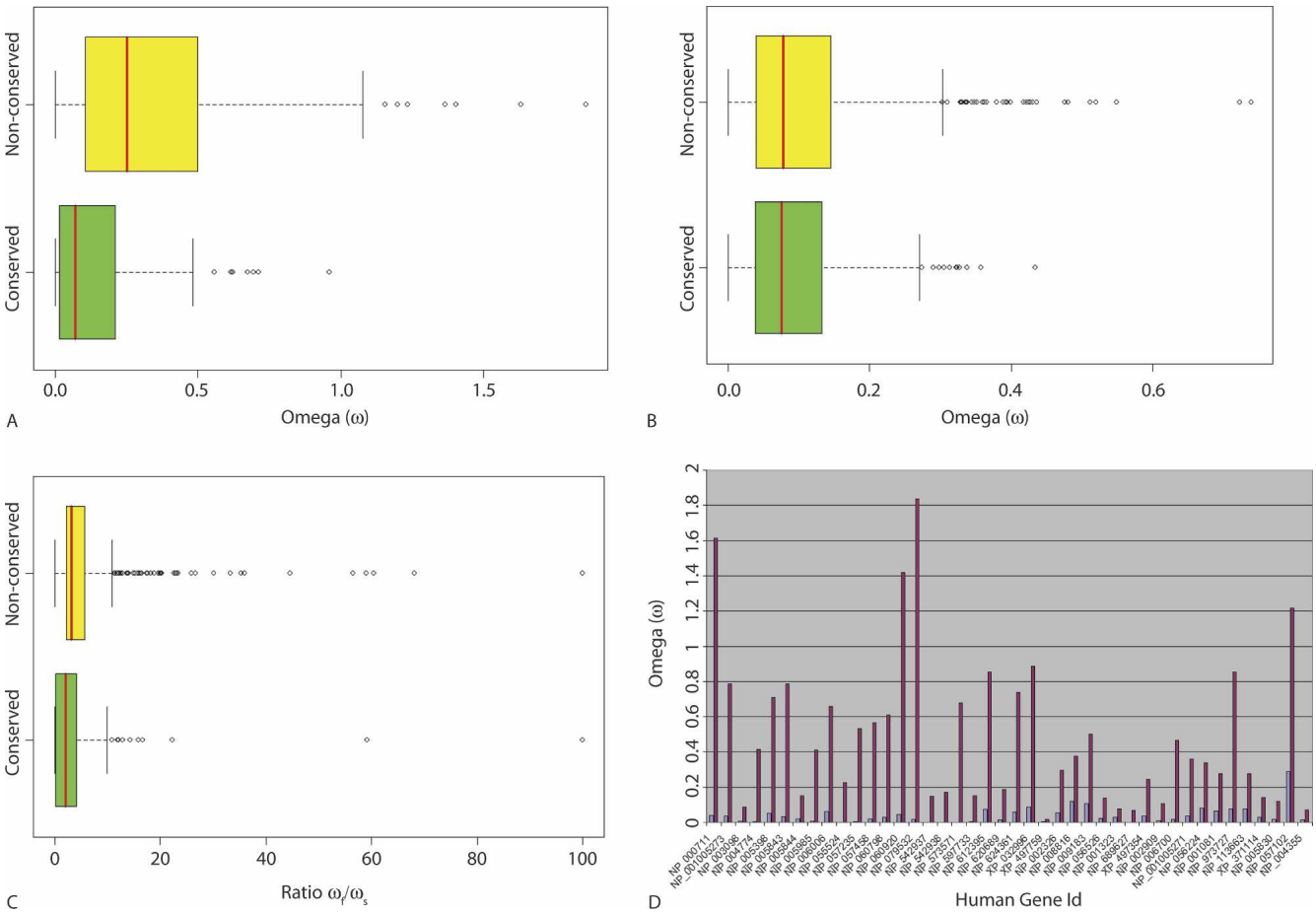

Figure 2. (*A*–*C*) Box plots representing the distribution of the ω values (*A*,*B*) and the ratio ω_f/ω_s (*C*). The first quartile right boundary is the *left* border of the box, the red line is the median, and the third quartile right boundary is the *right* border of the box. (*A*) The distribution of the flanks ω values, nonconserved, first quartile boundary: 0.105, median: 0.251, third quartile boundary: 0.505. Conserved, first quartile boundary: 0.015, median: 0.071, third quartile boundary: 0.211. There is a significant difference between the conserved and nonconserved groups ($P = 3.275 \times 10^{-12}$, two-sample Kolmogorov–Smirnov test). The nonconserved data set contains three ω values of 1.0×10^6; these values were removed from the plot to allow improved inspection/comparison of the two distributions. (*B*) The distribution of the shadow's ω values, nonconserved, first quartile boundary: 0.04, median: 0.078, third quartile boundary: 0.145. Conserved, first quartile boundary: 0.039, median: 0.075, third quartile boundary: 0.132. There is no significant difference between the conserved and nonconserved groups ($P = 0.961$, two-sample Kolmogorov–Smirnov test). (*C*) Nonconserved, first quartile boundary: 2.203, median: 3.254, third quartile boundary: 5.752. Conserved, first quartile boundary: 0.122, median: 2.115, third quartile boundary: 4.119. There is a significant difference between the conserved and nonconserved groups ($P = 9.115 \times 10^{-11}$, two-sample Kolmogorov–Smirnov test). (*D*) The ω values for the conserved and nonconserved flanking regions in the protein alignments that contained both conserved and nonconserved repeats. There is a significant difference between the conserved and nonconserved groups, and the nonconserved flanking regions are under less purifying selection compared with conserved flanking regions ($P = 3.7 \times 10^{-183}$). Of note are the four proteins whose nonconserved flanking regions may be undergoing adaptive selection ($\omega > 1$).

pletely sequenced and believe that the database will be of interest and utility in the research of triplet repeat expansion.

Taken as a whole, our analyses suggest that mutation alone does not dictate which codons are reiterated or their prevalence in homopeptides and that there is no bias toward G+C- or A+T-rich codons that is not already present in the transcriptome of the species (this general trend is most clearly seen in *H. sapiens* and *P. falciparum*). These analyses also show that the selective pressures at the nucleotide and amino acid level (such as the structural and functional constraints of a protein) influence the codons that are reiterated and their length within the homopeptides. This of course, does not preclude mutational events (such as single nucleotide substitutions and replication slippage) from being the primary driving forces in the origin of homopeptides.

In general, transitions substitutions are more common than transversions (Collins and Jukes 1994; Morton 1995; Kumar 1996; Moriyama and Powell 1997). Contrary to this observation, our results showed that the codons immediately 5′ and 3′ of a homocodon are typically single base transversions and are infrequently reiterated. Disruptions of homocodons are likely to reduce the rate of slippage mutation as shown in *Saccharomyces cerevisae* (Rolfsmeier and Lahue 2000) and implied by the lower (on average) heterozygosity of mammal imperfect repeats compared with perfect repeats (Weber 1990). Thus, we hypothesize that codon transversions may be selectively favored by both enhancing the genetic stability of the homopeptide and lowering their reversion rate.

Hancock et al. (2001) showed that CAG repeats in humans and mice lay within regions under weaker selective pressures than the rest of the gene (excluding the homopeptide). These

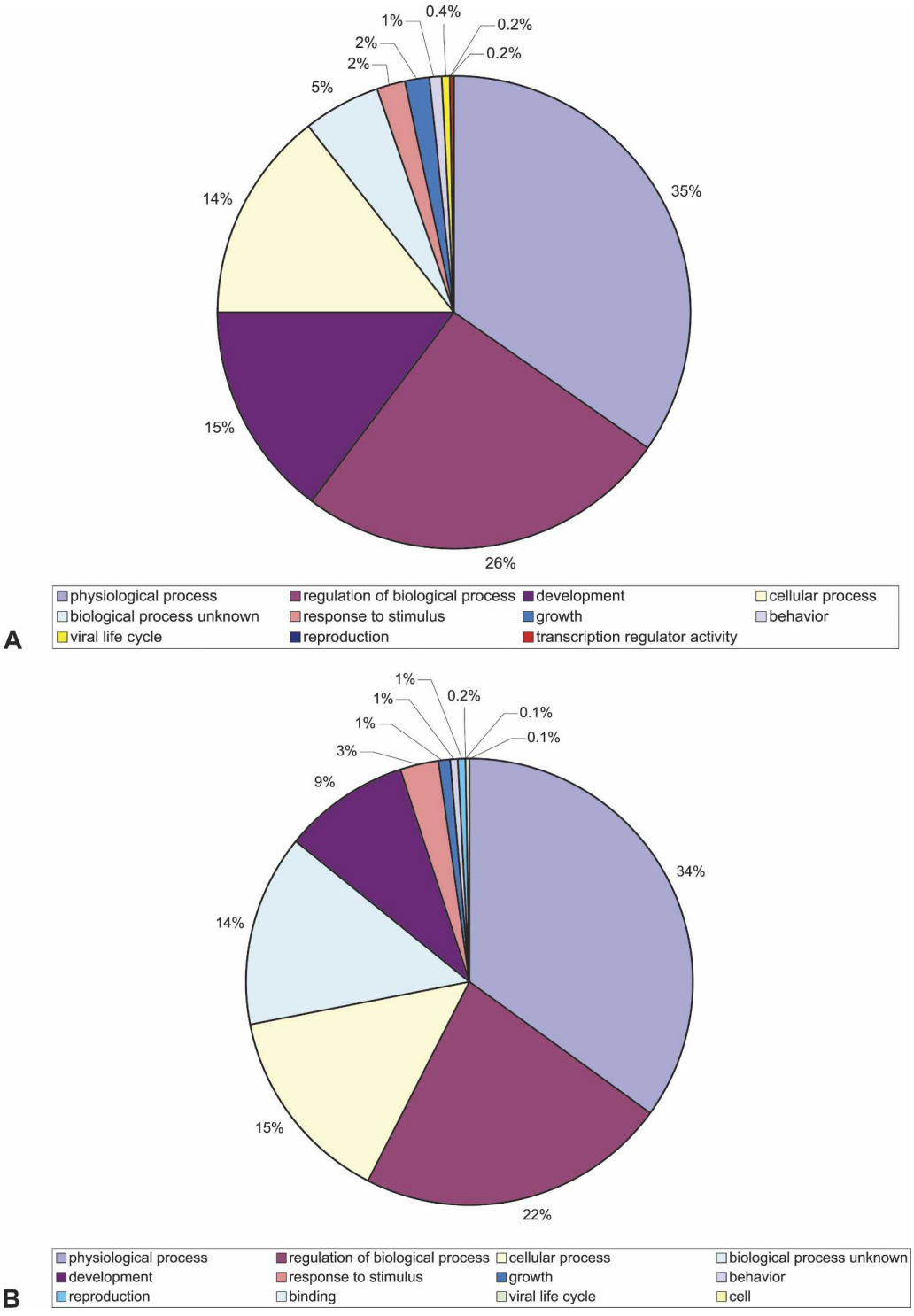

Figure 3. Distribution of the biological processes as defined in Gene Ontology (GO) for RCPs where their repeats are either conserved (*A*) or not conserved (*B*) across *H. sapiens*, *R. norvegicus*, and *M. musculus*.

investigators also showed that conserved homopeptides reside within regions under strong selective pressures compared with nonconserved homopeptides. Here, we sought to expand these observations across all RCPs. Our results indicate that, for the sampled mammal lineages at least, the flanking regions of nonconserved homopeptides have higher ω (i.e., K_a/K_s) rates than those flanking conserved homopeptides. These data suggest that nonconserved homopeptides lie in regions that are under lower purifying selection. This effect is also seen in RCPs that contain both conserved and nonconserved homopeptides. In compari-

son to the shadow regions of RCPs (i.e., the remainder of the RCP after the removal of the homopeptide(s) and the immediate flanking regions), the flanking regions are generally under less selective pressure, as evidenced by larger ω values, indicating that homopeptides lie in relatively neutrally evolving regions of the RCPs. These data are consistent with the work of Hancock et al. (2001) and the observations that homopeptides tend to be within functionally and structurally more evolutionarily active regions (Karlin and Burge 1996; Huntley and Golding 2002; Alba and Guigo 2004; Faux et al. 2005).

A recent study on the role of repeats in development focused on transcription factors involved in development (Fondon and Garner 2004). We and others (Karlin and Burge 1996; Karlin et al. 2002; Faux et al. 2005) have also shown that a large number of proteins with homo-amino acid repeats are involved in transcription regulation. In light of these studies, the question arises as to whether there are any functional or biological process biases or differences between those RCPs where the homopeptide is conserved across humans and rodents and in those RCPs where there is no conservation of the homopeptide. We were unable to detect any such biases.

In conclusion, the likelihood of a homopeptide to expand or contract depends on a number of factors: the species' genetic background, as previously suggested by Karlin et al. (2002); the selective pressures at the nucleotide (the ability to form stable nucleic acid secondary structures) and amino acid levels (the toxicity to both the protein structure and the cell); the influence of cofactors such as *cis* elements near the homocodon (Brock et al. 1999; Cleary and Pearson 2003); and the influence of other proteins (Richard et al. 2000; Feschenko et al. 2003; Owen et al. 2005), rather than a general trend driven by G+C content, as suggested by Alba and Guigo (2004), or just general mutational events. The predominant interruption of homocodons by transversions suggests there is selective pressure favoring repeat-stabilizing mutations. These data also suggest that most homopeptides originated as homocodons, and selection primarily at the amino acid level has dictated the type, prevalence, and length of the homopeptides and the length of the homocodons, whereas mutation at the nucleotide level has dictated the prevalence of homocodons within the homopeptides.

Methods

Previously, we and others have shown that prokaryotes have a paucity of homopeptide repeats in comparison to eukaryotes (Karlin and Burge 1996; Faux et al. 2005). Thus, for this study we chose to investigate homopeptide repeats in eukaryote organisms only. To avoid any sequence bias due to overrepresentation of particular protein families, organisms whose genomes have been completely sequenced or near completion as of January 2005 were chosen. These organisms are: *Homo sapiens* (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/), *Mus musculus* (ftp://ftp.ncbi.nih.gov/refseq/M_musculus/), *Rattus norvegicus* (ftp://ftp.ncbi.nih.gov/refseq/R_norvegicus/), *Gallus gallus* (ftp://ftp.ncbi.nih.gov/genomes/Gallus_gallus/), *Takifugu rubripes* (http://genome.jgi-psf.org/fugu6/fugu6.download.ftp.html), *Danio rerio* (ftp://ftp.ncbi.nih.gov/refseq/D_rerio/), *Drosophila melanogaster* (http://bugbane.bio.indiana.edu/data/seqs/genomic/), *Apis mellifera* (ftp://ftp.ensembl.org/pub/current_bee/data/fasta/), *Anopheles gambiae* (ftp://ftp.ensembl.org/pub/current_mosquito/data/fasta/), *Caenorhabditis elegans* (ftp://ftp.sanger.ac.uk/pub/databases/wormpep/wormpep136/), *Saccharomyces cerevisiae* (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/), *Plasmodium falciparum* 3D7 (http://plasmodb.org/restricted/data/P_falciparum/WG/), and *Arabidopsis thaliana* (ftp://ftp.arabidopsis.org/Sequences/blast_datasets/).

Detection of homopeptide repeats and trinucleotide repeats

For an average protein of 400 residues, a consecutive run of a single amino acid seven or more in length is statistically significant at the 0.1% level (Karlin 1995). On this basis we define the minimum length of a homopeptide to be seven and identified them as described in Faux et al. (2005). All trinucleotide repeats (TNRs) were similarly identified using a regular expression to find all TNRs with a minimum length of three.

Correction for multiple tests of a hypothesis

In several cases we are performing multiple tests of a single hypothesis, raising the problem of false positives. We adjusted our significance threshold to achieve an experiment-wide level of 0.05 per hypothesis using the Bonferroni correction. Dividing the specified experiment-wide significance by the number of tests conducted attains a corrected threshold. For instance, in the cases where the same hypothesis is tested separately for each of the 13 species, the experiment-wide threshold is $0.05/13 \approx 0.0038$. We report the raw probability estimates but indicate the significance relative to these adjusted thresholds.

Analysis of the codon usage within RCPs compared with the global codon usage

If homopeptides were derived primarily from DNA mutagenic processes, we would expect a significantly different synonymous codon usage for codons within (repeat) and those not within (shadow) homopeptides. We performed a χ^2 goodness-of-fit test on the codon frequencies of the repeat-containing genes. An extension of the aforementioned question is whether the involvement of the codon usage of the RCPs, for either the repeat or shadow codons, is proportional to their frequency in a species' transcriptome. χ^2 goodness-of-fit tests were also used for each of these cases. Goodness-of-fit tests were conducted using the R package (Team 2005), and a probability level of ≤ 0.001 (which is below the Bonferroni adjustment, i.e., 0.05/13[species] = 0.0038) was considered significant.

Statistical analysis of G+C elevation across the species investigated

To investigate if there was a species-lineage bias in the elevation of G+C content in the RCP's gene shadow, we used a two-sided sign test (where a successful event corresponds to a species with elevated G+C content in the RCP's gene shadow). This test was chosen as we have no a priori biochemical basis for suggesting that G+C should be increased in RCPs. While the assumption that the sampled lineages are statistically independent is incorrect, this violation increases the false-positive rate, making acceptance of the null hypothesis conservative.

Comparison of the codon usage for codon reiterants compared with the codon usage for nonreiterants

If homopeptides originated by a mutagenic process, we expect them to consist primarily of homocodon runs. The null to this alternate hypothesis holds that homopeptides consist of synonymous codons occurring proportional to their frequency in the shadow. We assessed the null hypothesis using a runs test, which

evaluates the probability of a random homocodon run of greater or equal length to the observed longest homocodon in the homopeptide. In order to do this, a random homopeptide sequence of length l was created by drawing with replacement l codons from the gene's shadow. The length of the longest homocodon run in the simulated homopeptide was determined. One thousand randomized sequences were generated, and the number of times that a randomized sequence contained a homocodon run of the same length or longer than that in the homopeptide was taken as an estimate of the probability (P) that the null hypothesis was correct. We performed a test of this hypothesis across all homopeptides in a transcriptome by calculating $-2\Sigma ln\mathrm{P}$, which is a distributed χ^2 with degrees of freedom (df) equal to twice the number of homopeptides (Sokal and Rohlf 1995). The probability level of ≤ 0.001 (which is below the Bonferroni adjusted threshold, i.e., 0.05/13[species] = 0.0038) was considered significant.

Discovery of the *H. sapiens*, *M. musculus*, and *R. norvegicus* putative homologs

Each *H. sapiens*, *M. musculus*, and *R. norvegicus* RCP was used as a query and searched against a database containing the proteomes of the three species using BLASTp (Altschul et al. 1997). For each search, the putative homolog of the other species (i.e., the sequence with the lowest e-value) was retained. Those RCPs for which a putative homolog was identified in the other two species were then aligned using ClustalW (Thompson et al. 1994). To ensure that placement of gaps was respectful of codon boundries, the protein alignment was used to guide the alignment of the cDNA. These alignments were then used to calculate the ratio (ω) of nonsynonymous codon substitution (K_a)/synonymous codon substitution (K_s), as stated below.

Assessing the association between the selective constraints of the homopeptides and their flanks

Hancock et al. (2001) showed that CAG repeats conserved across human and mouse proteomes lay within regions of higher purifying selection than those CAG repeats that were not conserved. We conjectured that the flanking regions (the 33 amino acids on either side of the repeat) of conserved homopeptides would exhibit a significantly lower rate of nonsynonymous substitutions than flanking regions of nonconserved homopeptides. Assuming synonymous substitutions are selectively neutral, ω indicates the type of natural selection affecting a sequence: $\omega < 1.0$ indicates that the sequences are undergoing purifying selection; $\omega = 1.0$ indicates neutral selection; and $\omega > 1.0$ indicates adaptive selection.

We tested these hypotheses concerning the correspondence between homopeptide classification (conserved or unconserved) and flanking sequence evolution with a series of nested models that differed by assigning independent values for ω to each annotated region of sequence. To address the hypotheses regarding the relative level of selective constraints, we used the codon substitution model of Yang (1998) in the calculations of ω. Elements of the matrix Q of instantaneous change are defined as:

$$q_{ij,i\neq j} = \begin{cases} 0, & \textit{more than one nucleotide difference} \\ \pi_j, & \textit{synonymous transversion} \\ \pi_j\omega, & \textit{nonsynonymous transversion} \\ \pi_j\kappa, & \textit{synonymous transition} \\ \pi_j\kappa\omega, & \textit{nonsynonymous transition} \end{cases}$$

where i, j are codons, κ is the ratio of transitions to transversions, ω is the ratio of nonsynonymous to synonymous substitutions, and the π_j is the equilibrium codon probabilities of j. We follow the convention of estimating the latter as the average frequency across aligned sequences. The full specification of Q requires its rows sum to 0, which is achieved by constraining $q_{ii} = -\Sigma_{ij,i\neq j}\, q_{ij}$. The ω estimates were calculated with the maximum likelihood phylogeny package PyEvolve (Butterfield et al. 2004).

Our baseline model (*W*) was that of a single ω value for each gene. Because repeat regions evolve by distinct evolutionary mechanisms (such as slipped-strand mispairing), we sought to capture those properties by specifying a second model (*RS*) where the repeat (*r*) and the remainder or shadow (*s*) regions of an alignment are assigned independent values of ω; i.e., $\omega_r \neq \omega_s$. The hypothesis that homopeptides are under significantly lower selective constraints compared with their shadows is tested using this model. The nonsynonymous rate within the repeat is expected to be low, which may affect the acceptances of this hypothesis. However, this will not affect the following hypotheses as both the null and alternative hypotheses share ω_r and thus reflect their difference (ω_f, the new K_a/K_s parameter for the flank). The hypothesis that repeat-flanks (*f*) experience distinct selective influence was represented by a third model (*RFS*) in which independent ω values were assigned to each of *r*, *f*, and *s*; i.e., $\omega_r \neq \omega_s \neq \omega_f$ (note the ω values for *s* here are different compared with the values of *s* from the *RS* model and are thus not comparable). For loci defined as having mixed repeat types, the hypothesis that flanks next to conserved homopeptides (*con*) would be more conserved than flanks next to nonconserved homopeptides (*uncon*) was represented by further breaking up ω_f, assigning independent ω values to *f* next to conserved homopeptides (ω_{fcon}) and *f* next to nonconserved homopeptides (ω_{funcon}). As our hypothesis is a one-sided hypothesis, in this model (*M*) ω_{fcon} was constrained to be $\leq \omega_{funcon}$. Maximum likelihood estimates were obtained for each model for each locus.

Assuming alignments evolve independently, the likelihood for each hypothesis given a specified set of sequence alignments was determined as the sum of log-likelihoods from each alignment in the set. The number of free parameters was similarly determined. We refer to these as the cumulative log-likelihood (lnL_{cum}) and degrees of freedom (df_{cum}) for each hypothesis. Likelihood ratio, $LR = 2(lnL_{cum}[alt] - lnL_{cum}[null])$, statistics were used to assess the contribution of the parameters that distinguish the null and alternate hypotheses. The probability of observing a *LR* of equal or greater value was determined using the χ^2 distribution with df equal to the difference in cumulative number of free parameters between the alternate and null hypotheses.

The order of testing these hypotheses was as follows: the *RS* hypothesis ($\omega_r \neq \omega_s = \omega_f$) was tested against the null hypothesis *W* ($\omega_r = \omega_s = \omega_f$) (test 1); the *RFS* ($\omega_r \neq \omega_s \neq \omega_f$) was tested against the *RS* hypothesis (test 2); the *M* hypothesis ($\omega_r \neq \omega_{fcon} \neq \omega_{funcon}$) was tested against the *RFS* hypothesis (test 3). The latter test was explicitly assessable using RCPs containing a mixed homopeptide conservation pattern. This hypothesis was indirectly assessed for the pure repeat alignments by comparing distributions of ω values using the Kolmogorov–Smirnov test, applied using R. As described in the results, the first two hypotheses tested were assessed only for the group of gene alignments where the homopeptides are either conserved or not across the three species; thus, the level of significance with the Bonferroni adjustment is 0.05/2 = 0.0025. As the final hypothesis test (*M* against *RFS*) was assessed only for the mixed homopeptide conservation pattern, only one a priori test was performed, and the threshold for significance does not need to be adjusted.

Functional assignment of the human RCPs with mouse and rat putative homologs

The biological process and molecular function of these human proteins were obtained from their NCBI (http://ncbi.nih.gov) GenBank record as stated by the Gene Ontology (GO) annotation. We chose to use the GO annotations, as they allow a consistent vocabulary and groupings of biological processes and molecular functions at various levels of generality. This collection was automated using several perl scripts and the eutils from NCBI. If there was no GO annotation available in the GenBank record, we attempted to obtain the biological process and functional information from the Human Protein Reference Database (HPRD: http://www.hprd.org/; Peri et al. 2004), Bioinformatic Harvester (EMBL Heidelberg: http://harvester.embl.de; Liebel et al. 2004), Mouse Genome Informatics (MGI: http://www.informatics.jax.org; Eppig et al. 2005), UniProt Knowledgebase (Swiss-Prot and TrEMBL: http://au.expasy.org; Bairoch et al. 2005), or SOURCE Search (http://genome-www5.stanford.edu/cgi-bin/source/sourceSearch; Diehn et al. 2003).

Assessment of higher order repeats

Homopeptides could also result from hexanucleotide or nonanucleotide repeats. The occurrence of such higher order DNA repeats can be detected as a prevalence of a heterodi- or heterotricodons. We investigated their prevalence by constructing a suffix tree and counting the occurrences of heterodi- and heterotricodons.

For each species, each three-letter codon in every repeat DNA sequence was substituted by a single unique character, and the resulting transformed repeat sequences were stored in a suffix tree (for an introduction to suffix trees, see Ukkonen 1995). Each node *N* in the suffix tree contained two fields: a string (of unique characters) and the list of positions (if any) in which string *S* appears in each (transformed repeat) sequence, where *S* is computed by concatenating the strings of all nodes from the root of the tree to *N*. The tree was then processed to obtain, for every string *S* occurring more than once in at least one sequence, the total number of sequences containing two or more occurrences of *S* divided by the total number of sequences with that amino acid repeat. Thus, this number represents the frequency of the particular sequence of codons represented by *S*, occurring more than once in the same repeat DNA sequence. For the purpose of this study we disregarded strings containing a single codon repeat, thus considering only those containing two or more unique codons. We also limited our study to strings of two and three codons in length, i.e., heterodi- and heterotricodons. The program was implemented in the Mercury language (http://www.cs.mu.oz.au/research/mercury/).

Acknowledgments

J.C.W. is a National Health and Medical Research Council (NHMRC) of Australia Principal Research Fellow and Monash University Senior Logan Fellow. We thank the NHMRC, the Australian Research Council, the Victorian Partnership for Advanced Computing (VPAC), and the State Government of Victoria for support. K.M. is an Australian Research Council Ph.D. student. We thank Sophie Katsabanis and Arthur Lesk for discussion and comment on the manuscript.

References

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33:** D154–D159.

Brock, G.J., Anderson, N.H., and Monckton, D.G. 1999. *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.

Butterfield, A., Vedagiri, V., Lang, E., Lawrence, C., Wakefield, M.J., Isaev, A., and Huttley, G.A. 2004. PyEvolve: A toolkit for statistical modelling of molecular evolution. *BMC Bioinformatics* **5:** doi: 10.1186/1471-2105-5-1.

Caburet, S., Vaiman, D., and Veitia, R.A. 2004. A genomic basis for the evolution of vertebrate transcription factors containing amino acid runs. *Genetics* **167:** 1813–1820.

Cleary, J.D. and Pearson, C.E. 2003. The contribution of *cis*-elements to disease-associated repeat instability: Clinical and experimental evidence. *Cytogenet. Genome Res.* **100:** 25–55.

Cocquet, J., De Baere, E., Caburet, S., and Veitia, R.A. 2003. Compositional biases and polyalanine runs in humans. *Genetics* **165:** 1613–1617.

Collins, D.W. and Jukes, T.H. 1994. Rates of transition and transversion in coding sequences since the human–rodent divergence. *Genomics* **20:** 386–396.

Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J.C., Hernandez-Boussard, T., Rees, C.A., Cherry, J.M., Botstein, D., Brown, P.O., et al. 2003. SOURCE: A unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.* **31:** 219–223.

Ellegren, H. 2002. Mismatch repair and mutational bias in microsatellite DNA. *Trends Genet.* **18:** 552.

Eppig, J.T., Bult, C.J., Kadin, J.A., Richardson, J.E., Blake, J.A., Anagnostopoulos, A., Baldarelli, R.M., Baya, M., Beal, J.S., Bello, S.M., et al. 2005. The Mouse Genome Database (MGD): From genes to mice—A community resource for mouse biology. *Nucleic Acids Res.* **33:** D471–D475.

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

Feschenko, V.V., Rajman, L.A., and Lovett, S.T. 2003. Stabilization of perfect and imperfect tandem repeats by single-strand DNA exonucleases. *Proc. Natl. Acad. Sci.* **100:** 1134–1139.

Fondon 3rd, J.W. and Garner, H.R. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci.* **101:** 18058–18063.

Hancock, J.M., Worthey, E.A., and Santibanez-Koref, M.F. 2001. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Mol. Biol. Evol.* **18:** 1014–1023.

Harding, R.M., Boyce, A.J., and Clegg, J.B. 1992. The evolution of tandemly repetitive DNA: Recombination rules. *Genetics* **132:** 847–859.

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

Karlin, S. 1995. Statistical significance of sequence patterns in proteins. *Curr. Opin. Struct. Biol.* **5:** 360–371.

Karlin, S. and Burge, C. 1996. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci.* **93:** 1560–1565.

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

Kumar, S. 1996. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics* **143:** 537–548.

Liebel, U., Kindler, B., and Pepperkok, R. 2004. 'Harvester': A fast meta search engine of human protein resources. *Bioinformatics* **20:** 1962–1963.

Marcadier, J.L. and Pearson, C.E. 2003. Fidelity of primate cell repair of a double-strand break within a (CTG).(CAG) tract. Effect of slipped DNA structures. *J. Biol. Chem.* **278:** 33848–33856.

Mirkin, S.M. 2006. DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. Struct. Biol.* **16:** 351–358.

Moriyama, E.N. and Powell, J.R. 1997. Synonymous substitution rates in *Drosophila*: Mitochondrial versus nuclear genes. *J. Mol. Evol.* **45:** 378–391.

Morton, B.R. 1995. Neighboring base composition and

transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. *Proc. Natl. Acad. Sci.* **92:** 9717–9721.
Owen, B.A., Yang, Z., Lai, M., Gajek, M., Badger 2nd, J.D., Hayes, J.J., Edelmann, W., Kucherlapati, R., Wilson, T.M., and McMurray, C.T. 2005. (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. *Nat. Struct. Mol. Biol.* **12:** 663–670.
Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K., Chandrika, K.N., Deshpande, N., Suresh, S., et al. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* **32:** D497–D501.
Richard, G.F., Goellner, G.M., McMurray, C.T., and Haber, J.E. 2000. Recombination-induced CAG trinucleotide repeat expansions in yeast involve the MRE11-RAD50-XRS2 complex. *EMBO J.* **19:** 2381–2390.
Rolfsmeier, M.L. and Lahue, R.S. 2000. Stabilizing effects of interruptions on trinucleotide repeat expansions in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **20:** 173–180.
Schlotterer, C. and Tautz, D. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20:** 211–215.
Siwach, P., Pophaly, S.D., and Ganesh, S. 2006. Genomic and evolutionary insights into genes encoding proteins with single amino acid repeats. *Mol. Biol. Evol.* **23:** 1357–1369.
Sokal, R.R. and Rohlf, F.J. 1995. *Biometry: The principles and practice of statistics in biological research*. W.H. Freeman, New York.
Tachida, H. and Iizuka, M. 1992. Persistence of repeated sequences that evolve by replication slippage. *Genetics* **131:** 471–478.
Team, R.D.C. 2005. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Ukkonen, E. 1995. On-line construction of suffix trees. *Algorithmica* **14:** 249–260.
Veitia, R.A. 2004. Amino acids runs and genomic compositional biases in vertebrates. *Genomics* **83:** 502–507.
Weber, J.L. 1990. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics* **7:** 524–530.
Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15:** 568–573.

Received January 2, 2007; accepted in revised form April 10, 2007.