Natural selection drives the accumulation of amino acid tandem repeats in human proteins

Loris Mularoni,[1,2,5] Alice Ledda,[1,5] Macarena Toll-Riera,[1,3] and M. Mar Albà[1,3,4,6]

[1]*Biomedical Informatics Research Programme (GRIB), Fundació Institut Municipal d'Investigació Mèdica (FIMIM), Barcelona 08003, Spain;* [2]*Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21298, USA;* [3]*Pompeu Fabra University (UPF), Barcelona 08002, Spain;* [4]*Catalan Institution for Research and Advanced Studies (ICREA), Barcelona 08010, Spain*

Amino acid tandem repeats are found in a large number of eukaryotic proteins. They are often encoded by trinucleotide repeats and exhibit high intra- and interspecies size variability due to the high mutation rate associated with replication slippage. The extent to which natural selection is important in shaping amino acid repeat evolution is a matter of debate. On one hand, their high frequency may simply reflect their high probability of expansion by slippage, and they could essentially evolve in a neutral manner. On the other hand, there is experimental evidence that changes in repeat size can influence protein–protein interactions, transcriptional activity, or protein subcellular localization, indicating that repeats could be functionally relevant and thus shaped by selection. To gauge the relative contribution of neutral and selective forces in amino acid repeat evolution, we have performed a comparative analysis of amino acid repeat conservation in a large set of orthologous proteins from 12 vertebrate species. As a neutral model of repeat evolution we have used sequences with the same DNA triplet composition as the coding sequences—and thus expected to be subject to the same mutational forces—but located in syntenic noncoding genomic regions. The results strongly indicate that selection has played a more important role than previously suspected in amino acid tandem repeat evolution, by increasing the repeat retention rate and by modulating repeat size. The data obtained in this study have allowed us to identify a set of 92 repeats that are postulated to play important functional roles due to their strong selective signature, including five cases with direct experimental evidence.

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

Amino acid tandem repeats, also known as homopeptides, are regions within proteins characterized by the consecutive recurrence of a single amino acid. In humans, ~15%–20% of the proteins contain at least one amino acid tandem repeat of size 5 or longer (Karlin et al. 2002; Albà and Guigo 2004). At the DNA level they might be encoded by a run of pure codons or by a mixture of synonymous codons. Runs of pure codons are more frequent than expected given the frequencies of each individual codon (Albà et al. 1999a, 2001), which reflects the influence of trinucleotide slippage in amino acid tandem repeat expansion. Due to the high mutation rates associated with slippage (Weber and Wong 1993), repeats are an important source of genetic variability (Wren et al. 2000; Mularoni et al. 2006) and may contribute to adaptive processes (Fondon and Garner 2004; Kashi and King 2006). It has also recently been shown that alternatively spliced genes are enriched in both the number and the length of amino acid tandem repeats (Haerty and Golding 2010). Not surprisingly, coding repeats often show high interspecies size variation. For example, ~17% of the repeats in orthologous human and chimpanzee proteins differ in size in the two species (Mularoni et al. 2008). In humans, the uncontrolled expansion of CAG triplets encoding glutamine tracts is associated with a number of neurological disorders, such as Huntington disease and several ataxias (Brown and Brown 2004; Gatchel and Zoghbi 2005). Mutations generating abnormally long alanine tracts also cause several human developmental diseases (Brown and Brown 2004).

[5]These authors contributed equally to this work.
[6]Corresponding author.
E-mail malba@imim.es; fax 34-93-3160550.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.101261.109.

The role of amino acid tandem repeats in protein function remains a matter of controversy. Repeats are generally located in protein regions that show low sequence conservation when compared to other parts of the protein (Hancock et al. 2001; Faux et al. 2007). It has been claimed that the high frequency of amino acid tandem repeats may simply reflect their high probability of expansion rather than their functional importance (Lovell 2003). In addition, the analysis of the codon composition of serine repeats has revealed that the majority of them are strongly influenced by slippage, with few cases showing clear evidence of selection (Huntley and Golding 2006). Although the vast majority of repeats do not have a known function, several examples exist in which modifications of the size of glutamine, proline, or alanine tracts can alter the capacity of the protein to activate or repress transcription (Gerber et al. 1994; Lanz et al. 1995; Janody et al. 2001; Galant and Carroll 2002; Buchanan et al. 2004; Brown et al. 2005). It has also been recently discovered that histidine repeats mediate nuclear speckle trafficking in several transcription factors and neural development proteins (Alvarez et al. 2003; Salichs et al. 2009). Finally, amino acid tandem repeats are often embedded in highly disordered protein regions (Huntley and Golding 2002; Tompa 2003; Simon and Hancock 2009), and thus may be involved in low-affinity or transient interactions that are hard to determine experimentally (Dunker et al. 2008).

Studies aimed at understanding vertebrate amino acid tandem repeat evolution have mainly focused on the comparison of orthologous proteins from pairs of related species (Albà et al. 1999b; Hancock et al. 2001; Faux et al. 2007; Mularoni et al. 2007, 2008; Simon and Hancock 2009). While strongly conserved repeats tend to be encoded by a mixture of synonymous codons, nonconserved repeats are more often encoded by runs of identical codons (Albà et al. 1999b; Albà and Guigo 2004; Mularoni et al.

2007). This is consistent with a predominant role of slippage in the formation of new, nonconserved, repeats. It has also been observed that well-conserved repeats tend to be embedded in protein regions that evolve more slowly than regions containing nonconserved repeats (Hancock et al. 2001; Faux et al. 2007; Mularoni et al. 2007; Simon and Hancock 2009). This bias suggests that repeat evolution is influenced by selection.

In order to measure the contribution of natural selection to amino acid tandem repeat evolution, we have compared the level of conservation of human amino acid repeats in orthologous proteins from 12 vertebrate species with the level of conservation of sequences of similar DNA composition but located in noncoding genomic regions, and thus expected to evolve neutrally. We find that the length and degree of phylogenetic conservation of amino acid tandem repeats is much higher than expected under neutrality, and that selection has played an important role in amino acid tandem repeat evolution.

Results

Amino acid repeat type and length depend on the protein functional class

We obtained a set of 6477 orthologous proteins from 12 different vertebrate species using the Ensembl database (Hubbard et al. 2007). Interestingly, the human was the species with the largest number of repeats in the data set (Fig. 1). The most commonly found human amino acid tandem repeats are shown in Table 1. We found an excess of polar and charged amino acids, and a paucity of hydrophobic residues, in comparison to the overall frequency of different amino acids in the proteins from the data set (Supplemental file 1, S1). These results were consistent with previous reports (Green and Wang 1994; Albà et al. 1999a; Karlin et al. 2002; Albà and Guigo 2004; Faux et al. 2005). In humans, the amino acids that showed the strongest overrepresentation in the repeats were glutamic acid, proline, and serine.

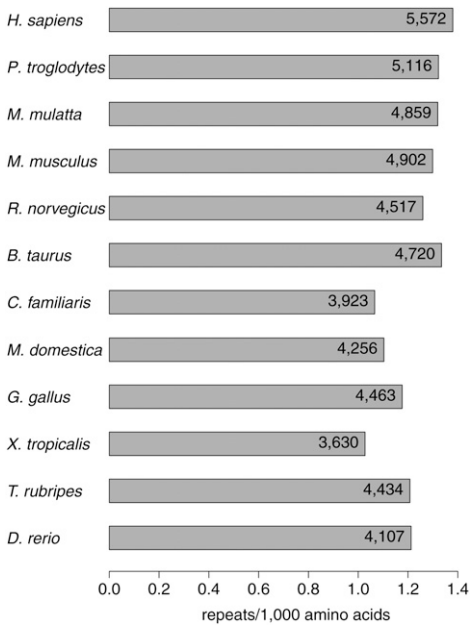

Figure 1. Amino acid repeat abundance in a vertebrate orthologous protein data set. The horizontal axis represents the average number of amino acid tandem repeats of size 4 or longer per 1000 amino acids, in a data set of 6477 one-to-one orthologous proteins. The values *inside* the bars indicate the total number of amino acid repeats in each species.

Table 1. Characteristics of human amino acid tandem repeats

	Amino acid tandem repeat size				
AA | triplet	*N*	Max	Mean	Median	Standard deviation
A | GCN	361	15	5.04	4	2.02
E | GAR	545	15	4.95	4	1.72
G | GGN	211	16	5.03	4	1.12
K | AAR	224	9	4.5	4	0.93
L | TTR CTN	402	11	4.34	4	0.75
P | CCN	446	21	4.88	4	1.83
Q | CAR	158	40	6.15	4	5
S | AGY TCN	160	42	4.67	4	2.26
Other	495	14	4.45	4	1.25
Total	3417	42	4.8	4	2.02

Repeats of size 4 or longer present in human proteins from a 6477 vertebrate protein orthologous data set.

Inspection of Gene Ontology (GO) terms (Harris et al. 2004) in the set of human proteins with repeats identified two main repeat-associated functional groups, "Transcription Factor and/or Development" and "Receptor and/or Membrane" (see Methods). This functional bias was consistent with the results obtained in other large-scale repeat analysis (Albà et al. 1999a; Karlin et al. 2002; Albà and Guigo 2004; Faux et al. 2005; Huntley and Clark 2007). Of special interest was determining which repeat types were most frequently associated with different functional groups. In this analysis, we also included for comparison a large functional group, "Metabolism," which was not overall significantly enriched in amino acid repeats.

The three functional groups mentioned above showed significant differences in the relative frequencies of repeats formed by different amino acids (Fig. 2, *P*-value < 10^{-5}; Supplemental file 1, S2). Leucine repeats were especially abundant in the "Receptor and/or Membrane" group, glutamine and alanine repeats in Transcription factor and/or Development, and lysine repeats in Metabolism. It has been previously observed that leucine repeats are typically located at the N-terminal end of proteins (Albà and Guigo 2004), where they may act as signal peptides (Karlin et al. 2002). This would explain the enrichment for leucine repeats in membrane proteins observed here. Several studies have shown that glutamine repetitive tracts fused to a DNA binding domain can activate transcription of a reporter gene, and that the addition of an alanine repeat represses transcription (Gerber et al. 1994; Janody et al. 2001; Galant and Carroll 2002), indicating that glutamine and alanine repeats can modulate transcriptional regulation. When we considered repeats of size 5 or longer, there was also an excess of glycine repeats in Transcription factor and/or Development (Supplemental file 1, S3). For repeats of this size, the excess of lysine repeats in Metabolism was no longer detectable, which is consistent with the fact that these repeats were among the shortest in the human proteome (Table 1).

Determination of the age of human amino acid tandem repeats

Human amino acid tandem repeats were classified into seven phylogenetic groups based on the existence of an overlapping repeat of size 4 or longer in a subset, or in the complete set, of aligned orthologous protein sequences (Fig. 3; Supplemental file 1, S4; Supplemental file 2). Given the very low proportion of repeats only

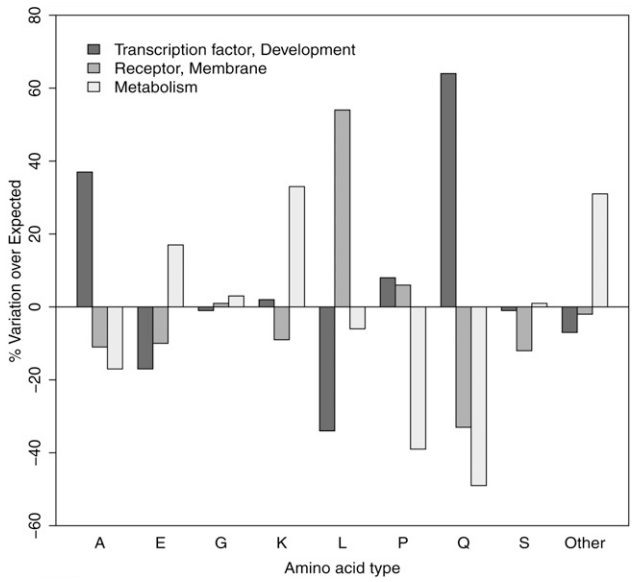

Figure 2. Distribution of amino acid repeat types in proteins from different functional classes. The classes were defined from Gene Ontology annotations (see text). The percentage difference with respect to the distribution of amino acids in all human repeats in the orthologous data set is shown.

found in humans (24 repeats) (Supplemental file 1, S5), the most recent phylogenetic group we considered was "Primata." This group was composed of human repeats that were also present in chimpanzee and rhesus macaque but absent from the rest of the mammalian and vertebrate orthologous proteins (157 repeats). Given their phylogenetic distribution, these repeats are likely to have originated, or to have expanded to a significant size (≥4 consecutive amino acids), in a primate ancestor. Similarly, we defined the groups "Euarchontoglires," "Eutheria," "Mammalia," "Amniota," "Tetrapoda," and "Vertebrata," containing increasingly older repeats (Supplemental file 2). To illustrate this, Figure 4 shows an example of an alanine repeat, present in the HOXD13 developmental transcription factor, which was classified as Amniota, as it was only found in mammals and chicken, and thus had probably originated in an Amniota ancestor. The phylogenetic group with the most repeats was Vertebrata, followed by Eutheria. The distribution was similar for repeats formed by amino acids of all types (Supplemental file 1, S4,S6). There were no major differences either in the conservation of repeats associated with different protein functional classes, with the exception of an under-representation of very recent repeats (Primata level) in Transcription Factor and/or Development (Supplemental file 1, S7). For the groups Tetrapoda or more recent, we could compare the number of repeats originated at that branch to the average number of genomic nucleotide substitutions in the same evolutionary interval, as published by Miller et al. (2007). The two parameters were roughly proportional. As the number of nucleotide substitutions can be taken as a proxy of evolutionary time, this indicated the lack of any sudden burst of repeat expansion and/or retention along the branches considered.

Excess of well-conserved amino acid tandem repeats over the neutral expectation

Comparison of human repeats with their vertebrate orthologs indicated that some of them were remarkably well conserved, with 28.3% of them being conserved in all vertebrate lineages examined (Vertebrata level, Fig. 3). A key issue in interpreting this result is to estimate how much conservation we expect in the absence of selection. Models of neutral evolution do not exist for amino acid tandem repeat sequences. The evolution of these sequences is affected by slippage, but a direct comparison with trinucleotide microsatellites is not appropriate, as amino acid tandem repeats might be encoded by complex mixtures of synonymous codons, making them less likely to undergo slippage than microsatellites (Fondon et al. 1998; Albà et al. 1999a; Wren et al. 2000; Mularoni et al. 2006). A more suitable model for comparison is sequences with the same DNA composition (consecutive repetition of 4 or more trinucleotides corresponding to synonymous codons) but located in noncoding genomic regions. These noncoding repeats will be subject to the same physical mutational process as repeats encoding tandem amino acids, but are expected to be free from selection, so that any conservation difference represents selective pressure on the coding repeats. We retrieved all such sequences from the human genome. Around one third were located in introns and the rest of them mostly in intergenic regions (Supplemental file 1, S8). The distribution of trinucleotide homogeneity values—the fraction of the repeat occupied by the longest pure trinucleotide repeat—was not significantly different from the set

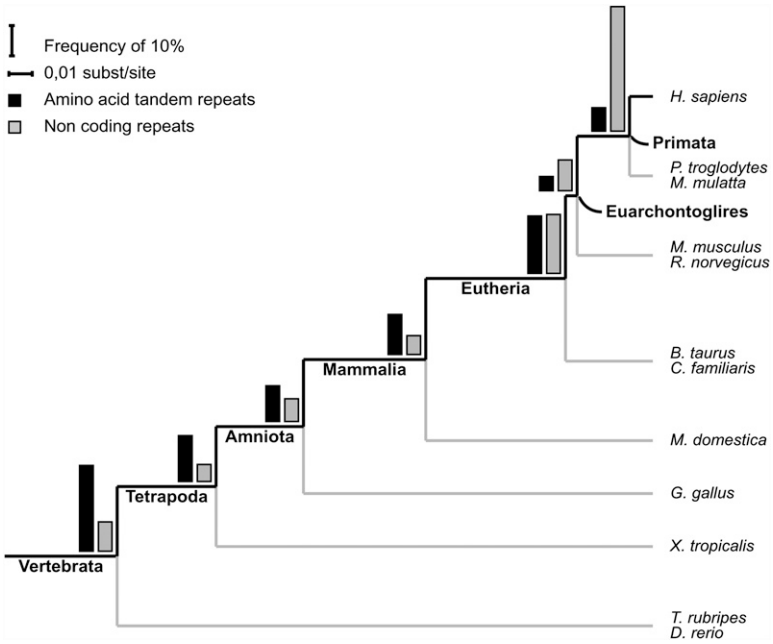

Figure 3. Phylogenetic depth of human repeat conservation. Relative number of human amino acid tandem repeats and of noncoding repeats of similar DNA composition that may have originated at different time periods according to their patterns of conservation.

```
H. sapiens        HSGRAAAAAAAAAAAAAAASGFA
P. troglodytes    HSGRAAAAAAAAAAAAAAASGFA
M. mulatta        HSGRAAAADAAAAAAAAAASGFA
M. musculus       HSGRAAAAAAAAAAAAAAASSFA
R. norvegicus     HSGRAAAAAAAAAAAAAAASSFA
B. taurus         HSGRAAAAAAAAAAAAAAASGFA
M. domestica      HSGRAAAAAAAAAAAAAAASGFS
G. gallus         HTGRAAAAAAAAA------SGFA
X. tropicalis     PSSR---------------TVPG
D. rerio          HSSR---------------CTSG
```

Figure 4. Alignment of the alanine repeat in HOXD13. The repeat was classified as Amniota as it was conserved in all mammalian species and in chicken (*G. gallus*), but not in frog (*X. tropicalis*) or zebrafish (*D. rerio*).

of amino acid tandem repeat encoding sequences (average 0.49 vs. 0.52 for amino acid repeats), which strengthened our comparisons. These repeats will be termed "noncoding repeats" for our purposes.

To compare the level of conservation of noncoding repeats to the level of conservation of amino acid repeats, we selected the human noncoding repeats in genomic regions showing conserved synteny with all the vertebrate species used to analyze amino acid repeat conservation. Synteny was determined by the availability of a genomic alignment comprising the complete set of species at the Genome Browser of the University of California Santa Cruz (Karolchik et al. 2008). Repeat conservation was defined as before, by the presence of an overlapping repeat of size 4 or longer in the other species. These data were employed to build the same phylogenetic groups as for amino acid repeats (Fig. 3; Supplemental file 3).

The level of conservation of the neutrally evolving, noncoding repeats, was strikingly lower than that of the amino acid tandem repeats (Fig. 3, P-value < 10^{-3}). For example, only 9.6% of the noncoding repeats were conserved at the deepest level (Vertebrata), in sharp contrast to 28.5% of the amino acid tandem repeats. Similar results were observed for different amino acid repeat types (Supplemental file 1, S4). Mutations disrupting the reading frame are in general highly disadvantageous in coding regions, but this kind of selective constraint does not exist in noncoding regions. To avoid any bias, we identified any such mutations in the noncoding alignments and discarded their effect on the classification of the noncoding repeats (see example in Supplemental file 1, S9). Although the number of deeply conserved noncoding repeats increased slightly (e.g., Vertebrata, 9.6% to 10.4%), the differences from the coding repeats remained highly significant (Supplemental file 1, S10, P-value < 10^{-3}). Altogether, these results indicate that selection strongly contributes to the evolution of many amino acid tandem repeats.

In addition, a significant fraction of the amino acid repeats classified as Vertebrata were conserved in other metazoans. In particular, we found that ~23% of them were already probably present in an ancestral metazoan, judging by the presence of repeats of the same nature and at the same protein location in close homologs from *Drosophila melanogaster* and/or *Caenorhabditis elegans*. An additional 7% were only found in *Ciona intestinalis*, indicating that they had probably originated in an ancestral chordate. Such phylogenetic depth of conservation reinforces the idea that a substantial fraction of the conserved repeats has been maintained by selection.

Identification of amino acid repeats under selection

The proportion of repeats under selection at each taxonomic level can be estimated if we assume that the vast majority of primate-specific repeats (Primata) are newly formed sequences upon which selection has not (yet) acted. Although this is a conservative assumption, as we cannot exclude that some primate-specific repeats might have been fixed by selection, it allows us to calculate the maximum number of repeats expected to be conserved at different phylogenetic depths under neutrality, using the frequencies of noncoding repeats. For example, the ratio between Eutheria and Primata noncoding repeats is 0.46. This means that, if amino acid tandem repeats were evolving neutrally, we should observe around 72 amino acid repeats at the Eutheria level (0.46 × 157 Primata amino acid repeats). However, we observed many more (384). This, to a first approximation, implies that 81% of the amino acid tandem repeats are conserved at the Eutheria level due to the action of selection. This percentage is 90% or higher for older repeats: 91% for Mammalia, 90% for Amniota, 94% for Tetrapoda, and 94% for Vertebrata (Supplemental file 1, S4). Thus, the vast majority of the repeats in these groups are more conserved than expected by mutational forces alone.

A second important difference concerned the size of the repeats. Whereas for recently formed repeats (Primata) the fraction of long repeats (size 6 or longer) was similar for amino acid repeats and for noncoding repeats (4.5% and 3%, respectively), for older repeats this fraction was significantly larger in amino acid repeats than in noncoding repeats (16% and 5%, respectively, P-value < 10^{-3}). Overall, the fraction of noncoding repeats of size 8 or longer was negligible (<0.1%). This means that conserved repeats of size 8 or longer are extremely rare in the absence of selection and thus are likely to be functionally important.

In light of these results, we compiled a list of 92 human amino acid repeats of size 8 or longer that are conserved at the Eutheria or deeper phylogenetic levels, and thus that are very likely to have been shaped by selection (Supplemental file 1, S11; Table 2 for a selection of repeats of size 10 or longer). Conservation meant that the size of the corresponding repeat in the other species was at least 4, although in most cases it was much longer than 4 and similar to the size of the human repeat. Many of these repeats are located in transcription factors where they may play roles in transcriptional regulation, as previously shown for alanine, glutamine, or proline tracts using reporter gene assays (Gerber et al. 1994; Lanz et al. 1995; Janody et al. 2001; Galant and Carroll 2002; Buchanan et al. 2004; Brown et al. 2005). Other proteins in the list contain histidine repeats. This type of repeat has no effect in transcriptional regulation but is required for translocation of several proteins to nuclear speckles (Salichs et al. 2009). We inspected the developmental expression of the set of proteins containing the 92 vertebrate conserved repeats conserved in the Edinburgh Mouse Atlas of Gene Expression (Venkataraman et al. 2008). Interestingly, the 10 top most frequent expression domains during development were related to the nervous system (Supplemental file 1, S12). Remarkably, expression in mouse developmental neural tissues was observed for 76% of the proteins with conserved repeats that had available expression data (Supplemental file 2, S11).

Experimental evidence of repeat functionality

We searched for any published data on repeat deletion/modifications experiments, or on naturally occurring mutations, for the 92 repeats listed previously. We found experimental data for five of them (5.4%). Although this is a small proportion of the repeats, reflecting the scarcity of repeat functional studies, in all five cases the expansion or deletion of the repeat had a measurable effect on the

Table 2. Selection of human amino acid tandem repeats conserved at different phylogenetic levels

Amino acid	Ensembl gene ID	Ensembl protein ID	HGNC	Position[a]	Size[b]	Ort. Size[c]	Description
Eutheria							
A	ENSG00000106689	ENSP00000362717	LHX2	186	10	7–10	LIM/homeobox protein Lhx2
E	ENSG00000110429	ENSP00000265651	FBXO3	423	10	7–10	F-box only protein 3
E	ENSG00000180592	ENSP00000326863	C10orf140	323	10	10–12	Novel protein (FLJ16611)
E	ENSG00000066248	ENSP00000264051	NGEF	215	11	4–11	Ephexin-1
E	ENSG00000155846	ENSP00000312649	PPARGC1B	433	11	11–14	Peroxisome proliferator-activated receptor γ
G	ENSG00000120093	ENSP00000308252	HOXB3	153	11	7–16	Homeobox protein HOXB3
G	ENSG00000153266	ENSP00000283268	FEZF2	101	16	7–16	Zinc finger protein 312
G	ENSG00000165655	ENSP00000361602	ZNF503	188	16	15–19	Zinc finger protein 503
P	ENSG00000170145	ENSP00000305976	SIK2	822	10	7–10	Serine/threonine-protein kinase SNF1-like
P	ENSG00000151612	ENSP00000281318	ZNF827	328	15	8–15	CDNA FLJ16555 fis
P	ENSG00000182742	ENSP00000328928	HOXB4	72	15	5–15	Homeobox protein HOXB4
Q	ENSG00000111676	ENSP00000349076	ATN1	483	19	5–19	Atrophin 1
S	ENSG00000038358	ENSP00000351811	EDC4	618	11	11–21	Enhancer of mRNA-decapping protein 4
S	ENSG00000205250	ENSP00000368686	E2F4	310	13	6–19	Transcription factor E2F4
S	ENSG00000012174	ENSP00000368798	MBTPS2	113	23	13–25	Membrane-bound transcription factor protease
Mammalia							
A	ENSG00000141448	ENSP00000269216	GATA6	172	11	11–12	Transcription factor GATA-6
A	ENSG00000163013	ENSP00000295133	FBXO41	50	11	4–11	F-box only protein 41
A	ENSG00000188620	ENSP00000350549	HMX3	60	13	11–3	Unknown function
A	ENSG00000106031	ENSP00000222753	HOXA13	37	14	7–14	Homeobox protein HOXA13
G	ENSG00000185129	ENSP00000332706	PURA	30	12	12–13	Transcriptional activator protein Pur-alpha
H	ENSG00000141448	ENSP00000269216	GATA6	323	10	4–10	Transcription factor GATA-6
S	ENSG00000111676	ENSP00000349076	ATN1	385	10	5–10	Atrophin 1
S	ENSG00000163635	ENSP00000295900	ATXN7	716	12	4–13	Ataxin 7
Amniota							
A	ENSG00000174279	ENSP00000312385	EVX2	397	11	4–11	Homeobox even-skipped homolog protein 2
A	ENSG00000128714	ENSP00000249505	HOXD13	48	15	10–15	Homeobox protein HOXD13
E	ENSG00000149970	ENSP00000368824	CNKSR2	876	10	4–11	Connector enhancer of kinase
H	ENSG00000116544	ENSP00000362444	DLGAP3	222	12	7–14	Disks large-associated protein 3
Q	ENSG00000183495	ENSP00000333602	EP400	2757	29	5–38	E1A-binding protein p400
S	ENSG00000148840	ENSP00000278070	PPRC1	1440	11	5–12	Peroxisome proliferator-activated receptor γ
Tetrapoda							
A	ENSG00000170549	ENSP00000305244	IRX1	28	10	7–13	Iroquois-class homeodomain protein IRX-1
A	ENSG00000164107	ENSP00000352565	HAND2	20	12	4–12	Heart- and neural crest derivatives-expr. prot.2
A	ENSG00000143970	ENSP00000337250	ASXL2	670	13	9–13	Additional sex combs like 2
P	ENSG00000136848	ENSP00000362887	DAB2IP	919	10	8–10	Disabled homolog 2-interacting protein
P	ENSG00000164944	ENSP00000297591	KIAA1429	138	11	11	KIAA1429
S	ENSG00000148840	ENSP00000278070	PPRC1	1479	20	5–20	Peroxisome proliferator-activated receptor γ
Vertebrata							
A	ENSG00000174279	ENSP00000312385	EVX2	358	12	9–13	Homeobox even-skipped homolog protein 2
A	ENSG00000087095	ENSP00000262393	NLK	58	13	11–13	Serine/threonine kinase NLK
G	ENSG00000164548	ENSP00000297071	TRA2A	215	16	7–18	Transformer-2 protein homolog
H	ENSG00000077458	ENSP00000351631	FAM76B	170	10	9–10	Protein FAM76B
H	ENSG00000157540	ENSP00000340373	DYRK1A	597	13	11–13	DYRK1A
L	ENSG00000115155	ENSP00000369878	OTOF	1968	11	4–11	Otoferlin
P	ENSG00000184922	ENSP00000327442	FMNL1	601	11	4–9	Formin-like protein 1
P	ENSG00000157827	ENSP00000334991	FMNL2	552	21	6–22	Formin-like protein 2
S	ENSG00000070495	ENSP00000302916	JMJD6	345	10	10	JmjC domain-containing protein 6
S	ENSG00000100941	ENSP00000216832	PNN	576	11	5–16	Pinin

[a]Position of the repeat within the protein.
[b]Repeat length (≥10).
[c]Repeat length in the orthologous proteins (≥4).

function of the protein (Table 3). Therefore, it seems very likely that the rest of the repeats showing similar strong selective signatures also influence their proteins' functions.

The first example is the poly-alanine tract located at position 48 of the HOXD13 human protein (Tables 2, 3). This tract shows a conserved length of 15 in the seven examined mammalian species, a length of 9 in chicken, and is completely absent from frog and fishes (Fig. 4); hence, it is classified as Amniota. An expansion of 7–14 alanine residues in this tract results in synpolydactyly, a limb abnormality involving duplication of the fingers (Muragaki et al. 1996). In support of a function of this repeat in limb development, it has been observed that transgenic mice lacking the 15-alanine tract present alterations in the formation of limb sesamoid bones (Anan et al. 2007).

The second well-studied case is the paired mesoderm homeobox 2B protein (PHOX2B), which contains a perfectly conserved alanine repeat of size 9 in all vertebrate species analyzed. PHOX2B mutants containing a further expansion of 5–13 alanines are associated with a disorder of the autonomic nervous system, congenital central hypoventilation syndrome (Amiel et al. 2003). In this particular case, recombination rather than slippage seems to be associated with the pathogenic expansions (Amiel et al. 2004).

Table 3. Mutation data supporting functionality of a set of phylogenetically conserved amino acid tandem repeats

Amino acid	Ensembl protein ID	Protein name	Position[a]	Size[b]	Mutation/system	Effect	Reference
A	ENSP00000249505	HOXD13	48	15			
					+7–14 Alanines/occurs in nature	Synpolydactyly	Muragaki et al. 1996
					Δ Alanine tract/transgenic mouse	Alterations in sesamoid bone formation	Anan et al. 2007
A	ENSP00000371160	PHOX2B	158	9			
					+5–13 Alanines/occurs in nature	Congenital central hypoventilation syndrome	Amiel et al. 2003
					+5 Alanines/cell transfection assays	Interaction between PHOX2B and CREBBP impaired	Wu et al. 2009
Q	ENSP00000349076	Atrophin 1	483	19			
					+30–69 Glutamines/occurs in nature	Dentatorubral-pallidoluysian atrophy	Igarashi et al. 1998
					+63 Alanines/yeast two-hybrid, cell transfection assays	Interacts with coactivator TAF10 interfering with cellular transcription	Shimohata et al. 2000
H	ENSP00000351631	FAM76B	170	10			
					Δ Histidine-rich tract/cell transfection assays	Translocation to nuclear speckles is severely impaired	Salichs et al. 2009
H	ENSP00000340373	DYRK1A	597	13			
					Δ Histidine-rich tract/cell transfection assays	Translocation to nuclear speckles is severely impaired	Salichs et al. 2009

[a]Position of the repeat within the protein.
[b]Repeat length.

In transfection assays, the interaction of PHOX2B with CREBBP to activate gene expression is severely impaired when the PHOX2B alanine tract is elongated by five additional alanine residues (Wu et al. 2009), indicating that the size of the tract is functionally constrained.

Atrophin 1 (dentatorubral-pallidoluysian atrophy) contains a tract of 19 glutamines in the human genome reference protein, although its size may vary from seven to 34 residues in healthy individuals (Gatchel and Zoghbi 2005). This tract is also present in the orthologous proteins from other eutherian mammals, but is absent from more distant species. An expanded repeat tract of 49 or longer is associated with dentatorubral-pallidoluysian atrophy in humans (Igarashi et al. 1998; Gatchel and Zoghbi 2005). Atrophin 1 containing the expanded glutamine tract interacts with transcription coactivators CREBBP and TAF10 and inhibits the CREB-mediated gene transcription (Shimohata et al. 2000).

The histidine repeat tract in the protein kinase DYRK1A contains 13 consecutive histidines in all species except in frog and zebrafish (11 and 12, respectively). The C-terminal region of DYRK1A in which this tract is embedded was found to be required for targeting to nuclear speckles, using transfection experiments of deletion mutants in human cells (Alvarez et al. 2003). This finding showed, for the first time, that repetitive tracts could be important for subcellular compartment targeting.

The protein FAM76B is a nuclear protein of unknown function that contains a histidine tract of size 9 or 10 in all vertebrates except fishes, with repeat size 5. This repeat is embedded in a 21-residue histidine-rich region, which has also been shown to be required for nuclear speckle targeting (Salichs et al. 2009). Interestingly, a very similar paralogous protein exists in humans, FAM76A, which lacks the histidine repeat, and which does not colocalize with speckles.

Discussion

Although amino acid tandem repeats, and simple sequences in general, are the most commonly shared patterns between eukaryotic proteins (Huntley and Golding 2000), their role in protein function is still poorly understood. Amino acid tandem repeats can rapidly accumulate in proteins as a result of trinucleotide repeat expansion (microsatellites) caused by slippage. In the absence of selection, these repeats may later degenerate by the accumulation of point mutations and repeat contraction also via slippage. In fact, the evolution of genomic microsatellites can be modeled as a balance between slippage and point mutation (Kruglyak et al. 1998). However, in coding repeats, selection may perturb this balance. We have attempted to quantify the effect of selection by measuring the differences in the phylogenetic conservation of amino acid repeats and of sequences of similar DNA composition that are located in noncoding genomic regions, and thus are expected to evolve neutrally. This approach has its foundation on a large body of comparative studies that use sequence conservation over the neutral expectation to infer the existence of functional constraints (Ponting 2008). We have found that the level of conservation of human coding repeats in vertebrate orthologous sequences is significantly higher than the level of conservation of equivalent noncoding sequences. Interestingly, the differences were not only due to the more distant comparisons, which may pose problems when aligning noncoding sequences, as we obtained similar estimates of the fraction of repeats under selection in the Mammalia group (repeats conserved only in mammals) and the Vertebrata group (repeats conserved in all vertebrates). The results presented here strongly indicate that selection has played a more important role than previously suspected in the preservation and evolution of many amino acid tandem repeats.

Selection acting on mutations can be purifying (also known as negative selection) when the mutation is selected against, or adaptive (positive selection) when the mutations confer an advantage that increases the chance that it will become fixed in the population. We observed that amino acid repeats conserved in several species tended to be longer than recently formed repeats, a trend that was not observed in the control set of neutrally evolving repeats. This suggests that repeat expansion is likely to have been, at least in some instances, adaptive. This is consistent with the observation that a minimum repeat size is required for the repetitive tract to have a functional effect and that increasing repeat size leads to more dramatic functional alterations (see below). In other words, long repeats are more likely to disrupt the function, and thus be selected against, if they are not beneficial. Once a repeat of a certain length has been formed, some amino acid replacements might be deleterious and may be eliminated by purifying selection. This is supported by a previous analysis on human amino acid repeat variants that showed that, in most cases, the polar/charged nature of the residue was preserved and that some amino acid replacements were never observed (Mularoni et al. 2006). By measuring the frequency of nonsynonymous versus synonymous substitutions in repeats and adjacent regions, this study also indicated that the strength of negative selection was surprisingly similar within repeats and in the surrounding sequences.

A well-studied case is that of histidine repeats. In recent work we have shown that the histidine repetitive tracts present in several nuclear proteins are both sufficient and necessary to translocate the protein to the nuclear subcompartment known as nuclear speckles (Salichs et al. 2009). These structures are involved in the storage and regulation of splicing factors and other nuclear proteins (Lamond and Spector 2003). The experiments performed demonstrated that a minimum repeat size of 6 was required for efficient protein translocation to this compartment and that the relative amount of translocated protein increased with increasing repeat size. Interestingly, human histidine repeats tended to be longer than sequences of similar DNA composition located in noncoding regions, suggesting positive selection for increased amino acid repeat size. Another example of size-dependent repeat functionality is the classical work of Gerber et al. (1994), in which the investigators showed that the size of glutamine or proline repeats fused to a GAL4 DNA-binding domain was positively correlated with the strength of transcriptional activation of a reporter gene containing GAL4 binding sites.

Trinucleotide slippage can result in repeat expansions or contractions, but there is evidence that, when the repeats are short, expansion dominates over contraction (Xu et al. 2000). This bias is expected to produce new repeats of moderate size at a high frequency, and, not unexpectedly, newer species-specific repeats are often encoded by pure codon runs, whereas older repeats that are highly conserved between human and mouse tend to contain a higher number of synonymous mutations (Albà et al. 1999b; Albà and Guigo 2004). Taken together, the data indicate that repeats that are well conserved across species tend to be longer and less uniform than recently formed repeats. Interestingly, some proteins appear to be particularly prone to accumulating repeats. In our data set, about one third of the human proteins contained more than one amino acid repeat, and, in most cases, the repeats belonged to different age groups. The acquisition of additional repeats in a protein may help fine-tune its function and/or modify some of its properties. One illustrative example is Bicoid in *Drosophila*. This protein contains several repetitive regions that act as regulatory domains, including a region containing several glutamine repeats and a region rich in alanine (Janody et al. 2001). The glutamine-rich region acts as a transcriptional activator domain that, in contrast to other activator domains in the protein, is not down-regulated by Torso. The alanine-rich region, however, reduces the activity of the glutamine-rich domain, restoring down-regulation by Torso (Janody et al. 2001).

A question of interest is whether the frequency distribution of different repeat types reflects the genomic GC (or AT) content of the species. In *Plasmodium* species, a correlation between genomic AT richness and AT-rich encoded repeats (such as asparagine or lysine repeats) has been observed (Dalby 2009). However, some of the species-specific differences could not be explained by the background nucleotide composition, indicating lineage-specific amino acid selection. In the present study, we found an important compositional bias in coding versus noncoding repeats, with a clear over-representation of GC-rich trinucleotides in the former type of sequences (Supplemental file 1, S4). This bias is consistent with the high GC content of exons when compared to introns and intergenic regions. It has been noted that the high GC content in mammalian genes should facilitate the formation of novel GC-rich repeats, and the newly formed repeats should, in turn, lead to an increase in gene GC content (Sumiyama et al. 1996; Nakachi et al. 1997). In fact, the fact that mammalian genes encoding amino acid repeats do, indeed, tend to have an unusually high GC content points in the same direction (Albà and Guigo 2004). Several alanine, glycine, and proline repeats in various transcription factors, such as HOX and class III POU proteins, have been previously described as being well conserved in mammals but completely absent from other vertebrates (Sumiyama et al. 1996; Nakachi et al. 1997; Mortlock et al. 2000; Lavoie et al. 2003; Anan et al. 2007). As these repeats are formed by amino acids encoded by GC-rich codons (alanine, GCN; glycine, GGC; and proline, CCN), it has been argued that this pattern may be due to the pressure for increased gene GC content in mammals with respect to cold-blooded vertebrates (Sumiyama et al. 1996). However, our data do not support this interpretation, as there is no significant enrichment of alanine, glycine, or proline residues at the Eutheria or Mammalia levels with respect to other phylogenetic groups (Supplemental file 1, S6).

In summary, we have compared the level of phylogenetic conservation of human amino acid repeats with that of sequences of similar DNA composition but located in genomic noncoding regions. The results strongly indicate that, although repeats probably may originally arise by a probabilistic slippage mechanism, they can be subsequently shaped by selection. We have identified a set of 92 human amino acid repeats showing a highly significant selective signature as evidenced by their strong phylogenetic conservation and their length. The majority of proteins containing these repeats show neural developmental expression. Mutations in five of these tracts alter the function of the protein, providing additional evidence that these repeats are likely to be involved in important cellular functions.

Methods

Sequences

Protein and cDNA sequences for human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), rhesus macaque (*Macaca mulatta*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), dog (*Canis familiaris*), cow (*Bos taurus*), opossum (*Monodelphis domestica*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis*), zebrafish (*Danio rerio*), and pufferfish

(*Takifugu rubripes*) were retrieved from Ensembl database release 46 (Hubbard et al. 2007). For each gene the longest protein was selected for further study. The genomic sequences were obtained from the University of California Santa Cruz (golden path 200603, NCBI36).

Identification of amino acid repeats in vertebrate proteins

We obtained 6477 groups of orthologous proteins from the species listed above, using the orthology definitions from Ensembl Compara (Vilella et al. 2009). Only one-to-one orthologous genes were considered, with the exception of bonefishes, which underwent a whole-genome duplication ~350 million yr ago (Panopoulou and Poustka 2005), and for which we considered all possible co-orthologs (one-to-many relationships with respect to the human protein).

We used an in-house Python program for the detection of all amino acid tandem repeats of size 4 or longer, their positions, and the DNA sequences encoding the repeats, in all proteins from the orthologous protein data set. We calculated codon homogeneity (CH) as the fraction of the repeat occupied by the longest pure codon run (Mularoni et al. 2007).

Identification of noncoding repeats

We defined the noncoding repeats as arrays of synonymous triplets located in noncoding genomic regions. For example, an array of three GAAs and one GAG in a noncoding sequence would be a noncoding repeat equivalent to a repeat of four glutamic acid residues in a protein. We identified all such sequences in the human genome, discarding those that were located within gene coding regions, using the coding sequence genomic coordinates available from Ensembl. In cases in which several repeats overlapped, we considered all of them (for GAAGAAGAAGAAGAA this would be equivalent to five glutamic acid residues when reading from +1, to four lysines when reading from +2, and to four arginines when reading from +3). However, such cases were rare (~7%) because most repeats were short (of size 4) and interrupted, and thus could only be read once.

Amino acid repeat conservation in orthologous sequences

We determined whether repeats present in human proteins from the orthologous protein data set were also present in other vertebrates. In cases in which the protein looked incomplete in Ensembl, we attempted to obtain the full sequence from GenBank using BLASTP with default parameters (Altschul et al. 1997). We built orthologous protein sequence alignments using T-Coffee (Notredame et al. 2000). The presence of a conserved repeat in another species was inferred by a procedure described previously (Mularoni et al. 2007). Briefly, for each repeat of the reference species (human), an equivalent repeat existed in the other species if it was formed by the same amino acid, it showed an overlapping position in the alignment, and it was above the size cutoff (4 or longer). If multiple repeats overlapped, we selected the longest one. We did not consider terminal parts of the alignment or internal regions with very long gaps (≥35 amino acids), as these regions could represent exons not yet annotated in some of the species.

Definition of groups of repeats conserved at different phylogenetic levels

Human amino acid tandem repeats were classified in different phylogenetic groups on the basis of their pattern of conservation in different vertebrate species. These groups were useful to determine the period over which a given repeat had been formed, or expanded to a significant length (≥4). The first group was Primata, containing repeats that, besides human, were found in chimpanzee and rhesus macaque orthologous proteins, but not in the rest of the vertebrate orthologs. Similarly, we defined the classes Euarchontoglires (Primata plus Rodents), Eutheria (Euarchontoglires plus Laurasiatheria), Mammalia (Eutheria plus *M. domestica*), Amniota (Mammalia plus *G. gallus*), Tetrapoda (Amniota plus *X. tropicalis*) and Vertebrata (Tetrapoda plus *D. rerio*/*T. rubripes*). In the case of Rodents, it was sufficient if one of the two species, mouse or rat, contained the repeat, to consider that the repeat had been present in a common ancestor. The same criterion was applied in the case of Laurasiatheria (dog and cow); this allowed us to circumvent possible problems in repeat annotation in some of the species, such as dog, that contained a suspiciously low number of repeats. In the case of fishes, we considered the presence of the repeat in any of the co-orthologs from *D. rerio* or *T. rubripes* as positive evidence that the repeat was ancestral. Following the above criteria, 2024 human repeats were classified in seven different taxon-specific groups: Vertebrata (574), Eutheria (384), Tetrapoda (307), Mammalia (268), Amniota (239), Primata (157), and Euarchontoglires (95).

In the case of noncoding repeats, we recovered genome multiple alignments from University of California Santa Cruz containing the same set of species described above (Karolchik et al. 2008). We mapped all previously identified human noncoding repeats onto these alignments. We obtained 799 repeats that could be classified in a well-defined phylogenetic group, following the same criteria than for amino acid repeats (see above). The level of conservation of noncoding repeats was significantly lower than of coding repeats (Supplemental file 1, S4). To evaluate how the lack of frame restrictions in noncoding regions could bias the results, we identified all insertions and deletions of size other than a multiple of 3 (1, 2, 4, 5, etc.), resulting in "frameshifts," as well as any substitutions or insertions resulting in "stop codons," in the regions aligning the human noncoding repeat (see example in Supplemental file 1, S9). The elimination of such mutations, which would not have been accepted in coding regions, moved 78 of the 799 human noncoding repeats to older classes (9.8%). This did not significantly alter the results (Supplemental file 1, S10), indicating that the stronger conservation of the coding repeats with respect to the noncoding ones was not simply due to constraints to preserve the reading frame.

Homologous protein search in other eukaryotes

We searched for homologs in non-vertebrate eukaryotic species using BLASTP sequence similarity searches (Altschul et al. 1997) of the human protein against species-specific protein libraries downloaded from Ensembl. An *E*-value < 10^{-4} was considered indicative of homology. A small percentage of proteins (~5%) lacked homologs in the six non-vertebrate species tested (*Anopheles gambiae*, *Arabidopsis thaliana*, *C. elegans*, *D. melanogaster*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*), indicating that they are likely to be vertebrate-specific proteins.

For proteins containing repeats classified as Vertebrata, we determined if similar repeats existed in metazoan homologs. For this, we focused on the three top-scoring BLASTP hits for *C. intestinalis*, *C. elegans*, and *D. melanogaster*. In 36 cases a similar repeat existed only in *C. intestinalis*, suggesting it was a chordate-specific repeat, and in 135 cases in *C. elegans* and/or *D. melanogaster*, indicating that the repeat was already present in a metazoan ancestor.

Gene Ontology annotation

We extracted all GO terms (Harris et al. 2004) for human proteins from the vertebrate orthologous data set containing repeats using

Biomart at Ensembl (Kasprzyk et al. 2004). The most common words found in the GO annotation of the proteins studied were: "metabolic," "response," "genesis," "transferase," "receptor," "transcription," "transport," "membrane," "development," and "differentiation." We grouped together those terms that were functionally related and tended to co-occur in the same protein and created three nonredundant data sets. The first data set, Transcription factor and/or Development, contained development-, genesis-, and transcription-related functions. The second data set, Receptor and/or Membrane, contained membrane-, receptor-, and transport-related functions. The third data set, Metabolism, contained proteins involved in metabolic processes and/or annotated as having transferase activity. To ensure lack of redundancy between data sets, genes already classified in the first group were eliminated from the rest. Subsequently, genes already classified in the second group were eliminated from the third group.

Statistical analysis

Statistical analysis was performed with the R statistical package (R Development Core Team 2007). The χ^2 distribution was used for all statistical tests unless stated otherwise.

Acknowledgments

We thank Roderic Guigó, Francesc Calafell, Eduardo Eyras, and Sarah Wheelan for useful discussions on this project. We received financial support from Infobiomed EU grant (L.M.), Ministerio de Ciencia y Tecnología (FPU fellowship (M.T.-R.), Plan Nacional BIO2009-08160), Regione Autonoma della Sardegna (A.L.), and Fundació ICREA (M.M.A.).

References

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

Albà MM, Santibanez-Koref MF, Hancock JM. 1999a. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. *J Mol Evol* **49:** 789–797.

Albà MM, Santibanez-Koref MF, Hancock JM. 1999b. Conservation of polyglutamine tract size between mice and humans depends on codon interruption. *Mol Biol Evol* **16:** 1641–1644.

Albà MM, Santibáñez-Koref MF, Hancock JM. 2001. The comparative genomics of polyglutamine repeats: Extreme differences in the codon organization of repeat-encoding regions between mammals and *Drosophila*. *J Mol Evol* **52:** 249–259.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Alvarez M, Estivill X, de la Luna S. 2003. DYRK1A accumulates in splicing speckles through a novel targeting signal and induces speckle disassembly. *J Cell Sci* **116:** 3099–3107.

Amiel J, Laudier B, Attie-Bitach T, Trang H, de Pontual L, Gener B, Trochet D, Etchevers H, Ray P, Simonneau M, et al. 2003. Polyalanine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome. *Nat Genet* **33:** 459–461.

Amiel J, Trochet D, Clement-Ziza M, Munnich A, Lyonnet S. 2004. Polyalanine expansions in human. *Hum Mol Genet* **13:** R235–R243.

Anan K, Yoshida N, Kataoka Y, Sato M, Ichise H, Nasu M, Ueda S. 2007. Morphological change caused by loss of the taxon-specific polyalanine tract in Hoxd-13. *Mol Biol Evol* **24:** 281–287.

Brown LY, Brown SA. 2004. Alanine tracts: The expanding story of human illness and trinucleotide repeats. *Trends Genet* **20:** 51–58.

Brown L, Paraso M, Arkell R, Brown S. 2005. In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: Alanine tract expansion modulates DNA binding and transactivation. *Hum Mol Genet* **14:** 411–420.

Buchanan G, Yang M, Cheong A, Harris JM, Irvine RA, Lambert PF, Moore NL, Raynor M, Neufing PJ, Coetzee GA, et al. 2004. Structural and functional consequences of glutamine tract variation in the androgen receptor. *Hum Mol Genet* **13:** 1677–1692.

Dalby AR. 2009. A comparative proteomic analysis of the simple amino acid repeat distributions in Plasmodia reveals lineage specific amino acid selection. *PLoS One* **4:** e6231. doi: 10.1371/journal.pone.0006231.

Dunker AK, Silman I, Uversky VN, Sussman JL. 2008. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* **18:** 756–764.

Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG, Whisstock JC. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res* **15:** 537–551.

Faux NG, Huttley GA, Mahmood K, Webb GI, de la Banda MG, Whisstock JC. 2007. RCPdb: An evolutionary classification and codon usage database for repeat-containing proteins. *Genome Res* **17:** 1118–1127.

Fondon JW III, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci* **101:** 18058–18063.

Fondon JW III, Mele GM, Brezinschek RI, Cummings D, Pande A, Wren J, O'Brien KM, Kupfer KC, Wei MH, Lerman M, et al. 1998. Computerized polymorphic marker identification: Experimental validation and a predicted human polymorphism catalog. *Proc Natl Acad Sci* **95:** 7514–7519.

Galant R, Carroll SB. 2002. Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* **415:** 910–913.

Gatchel JR, Zoghbi HY. 2005. Diseases of unstable repeat expansion: Mechanisms and common principles. *Nat Rev Genet* **6:** 743–755.

Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, Schaffner W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263:** 808–811.

Green H, Wang N. 1994. Codon reiteration and the evolution of proteins. *Proc Natl Acad Sci* **91:** 4298–4302.

Haerty W, Golding GB. 2010. Genome-wide evidence for selection acting on single amino acid repeats. *Genome Res* (this issue). doi: 10.1101/gr.101246.109.

Hancock JM, Worthey EA, Santibanez-Koref MF. 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.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32:** D258–D261.

Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, et al. 2007. Ensembl 2007. *Nucleic Acids Res* **35:** D610–D617.

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

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

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

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

Igarashi S, Koide R, Shimohata T, Yamada M, Hayashi Y, Takano H, Date H, Oyake M, Sato T, Sato A, et al. 1998. Suppression of aggregate formation and apoptosis by transglutaminase inhibitors in cells expressing truncated DRPLA protein with an expanded polyglutamine stretch. *Nat Genet* **18:** 111–117.

Janody F, Sturny R, Schaeffer V, Azou Y, Dostatni N. 2001. Two distinct domains of Bicoid mediate its transcriptional downregulation by the Torso pathway. *Development* **128:** 2281–2290.

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

Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36:** D773–D779.

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet* **22:** 253–259.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. 2004. EnsMart: A generic system for fast and flexible access to biological data. *Genome Res* **14:** 160–169.

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

Lamond AI, Spector DL. 2003. Nuclear speckles: A model for nuclear organelles. *Nat Rev Mol Cell Biol* **4:** 605–612.

Lanz RB, Wieland S, Hug M, Rusconi S. 1995. A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. *Nucleic Acids Res* **23:** 138–145.

Lavoie H, Debeane F, Trinh QD, Turcotte JF, Corbeil-Girard LP, Dicaire MJ, Saint-Denis A, Page M, Rouleau GA, Brais B. 2003. Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. *Hum Mol Genet* **12:** 2967–2979.

Lovell SC. 2003. Are non-functional, unfolded proteins ('junk proteins') common in the genome? *FEBS Lett* **554:** 237–239.
Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, et al. 2007. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17:** 1797–1808.
Mortlock DP, Sateesh P, Innis JW. 2000. Evolution of N-terminal sequences of the vertebrate HOXA13 protein. *Mamm Genome* **11:** 151–158.
Mularoni L, Guigo R, Albà MM. 2006. Mutation patterns of amino acid tandem repeats in the human proteome. *Genome Biol* **7:** R33. doi: 10.1186/gb-2006-7-4-r33.
Mularoni L, Veitia RA, Albà MM. 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* **89:** 316–325.
Mularoni L, Toll-Riera M, Albà MM. 2008. Comparative genetics of trinucleotide repeats in the human and ape genomes. In *Handbook of human molecular evolution* (ed. DN Cooper, H Kehrer-Sawatzki), pp. 677–686. Wiley, Chichester, UK.
Muragaki Y, Mundlos S, Upton J, Olsen BR. 1996. Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. *Science* **272:** 548–551.
Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S. 1997. Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. *Mol Biol Evol* **14:** 1042–1049.
Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302:** 205–217.
Panopoulou G, Poustka AJ. 2005. Timing and mechanism of ancient vertebrate genome duplications—the adventure of a hypothesis. *Trends Genet* **21:** 559–567.
Ponting CP. 2008. The functional repertoires of metazoan genomes. *Nat Rev Genet* **9:** 689–698.
R Development Core Team. 2007. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. http://www.r-project.org/
Salichs E, Ledda A, Mularoni L, Albà MM, de la Luna S. 2009. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet* **5:** e1000397. doi: 10.1371/journal.pgen.1000397.
Shimohata T, Nakajima T, Yamada M, Uchida C, Onodera O, Naruse S, Kimura T, Koide R, Nozaki K, Sano Y, et al. 2000. Expanded polyglutamine stretches interact with TAFII130, interfering with CREB-dependent transcription. *Nat Genet* **26:** 29–36.
Simon M, Hancock JM. 2009. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol* **10:** R59. doi: 10.1186/gb-2009-10-6-r59.
Sumiyama K, Washio-Watanabe K, Saitou N, Hayakawa T, Ueda S. 1996. Class III POU genes: Generation of homopolymeric amino acid repeats under GC pressure in mammals. *J Mol Evol* **43:** 170–178.
Tompa P. 2003. Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays* **25:** 847–855.
Venkataraman S, Stevenson P, Yang Y, Richardson L, Burton N, Perry TP, Smith P, Baldock RA, Davidson DR, Christiansen JH. 2008. EMAGE—Edinburgh Mouse Atlas of Gene Expression: 2008 update. *Nucleic Acids Res* **36:** D860–D865.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19:** 327–335.
Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Hum Mol Genet* **2:** 1123–1128.
Wren JD, Forgacs E, Fondon JW III, Pertsemlidis A, Cheng SY, Gallardo T, Williams RS, Shohet RV, Minna JD, Garner HR. 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am J Hum Genet* **67:** 345–356.
Wu HT, Su YN, Hung CC, Hsieh WS, Wu KJ. 2009. Interaction between PHOX2B and CREBBP mediates synergistic activation: Mechanistic implications of PHOX2B mutants. *Hum Mutat* **30:** 655–660.
Xu X, Peng M, Fang Z. 2000. The direction of microsatellite mutations is dependent upon allele length. *Nat Genet* **24:** 396–399.

Received September 30, 2009; accepted in revised form March 17, 2010.

Natural selection drives the accumulation of amino acid tandem repeats in human proteins

Loris Mularoni, Alice Ledda, Macarena Toll-Riera, et al.

Genome Res. 2010 20: 745-754 originally published online March 24, 2010
Access the most recent version at doi:10.1101/gr.101261.109