Mol Genet Genomics (2016) 291:1851–1869
DOI 10.1007/s00438-016-1219-7

ORIGINAL ARTICLE

Selection pressure on human STR loci and its relevance in repeat expansion disease

Makoto K. Shimada[1,2,3] · Ryoko Sanbonmatsu[3] · Yumi Yamaguchi-Kabata[2,4] · Chisato Yamasaki[2,3] · Yoshiyuki Suzuki[5] · Ranajit Chakraborty[6] · Takashi Gojobori[2,7] · Tadashi Imanishi[2,8]

Received: 7 November 2015 / Accepted: 21 May 2016 / Published online: 11 June 2016

Abstract Short Tandem Repeats (STRs) comprise repeats of one to several base pairs. Because of the high mutability due to strand slippage during DNA synthesis, rapid evolutionary change in the number of repeating units directly shapes the range of repeat-number variation according to selection pressure. However, the remaining questions include: Why are STRs causing repeat expansion diseases maintained in the human population; and why are these limited to neurodegenerative diseases? By evaluating the genome-wide selection pressure on STRs using the database we constructed, we identified two different patterns of relationship in repeat-number polymorphisms between DNA and amino-acid sequences, although both patterns are evolutionary consequences of avoiding the formation of harmful long STRs. First, a mixture of degenerate codons is represented in poly-proline (poly-P) repeats. Second, long poly-glutamine (poly-Q) repeats are favored at the protein level; however, at the DNA level, STRs encoding long poly-Qs are frequently divided by synonymous SNPs. Furthermore, significant enrichments of apoptosis and neurodevelopment were biological processes found specifically in genes encoding poly-Qs with repeat polymorphism. This suggests the existence of a specific molecular function for polymorphic and/or long poly-Q stretches. Given that the poly-Qs causing

Communicated by S. Xu.

Makoto K. Shimada and Ryoko Sanbonmatsu contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-016-1219-7) contains supplementary material, which is available to authorized users.

✉ Makoto K. Shimada
mshimada@fujita-hu.ac.jp

Ryoko Sanbonmatsu
nagon2323@yahoo.co.jp

Yumi Yamaguchi-Kabata
yamaguchi@megabank.tohoku.ac.jp

Chisato Yamasaki
cyamasak01@gmail.com

Yoshiyuki Suzuki
yossuzuk@nsc.nagoya-cu.ac.jp

Ranajit Chakraborty
Ranajit.Chakraborty@unthsc.edu

Takashi Gojobori
takashi.gojobori@kaust.edu.sa

Tadashi Imanishi
imanishi@tokai.ac.jp

[1] Institute for Comprehensive Medical Science, Fujita Health University, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan

[2] National Institute of Advanced Industrial Science and Technology, 2-3-26 Aomi Koto-ku, Tokyo 135-0064, Japan

[3] Japan Biological Informatics Consortium, 10F TIME24 Building, 2-4-32 Aomi, Koto-ku, Tokyo 135-8073, Japan

[4] Tohoku Medical Megabank Organization, Tohoku University, 2-1 Seiryo-machi, Aoba-ku, Sendai 980-8573, Japan

[5] Graduate School of Natural Sciences, Nagoya City University, 1 Yamanohata, Mizuho-cho, Mizuho-ku, Nagoya, Aichi 467-8501, Japan

[6] Health Science Center, University of North Texas, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA

[7] Computational Bioscience Research Center, King Abdullah University of Science and Technology, Ibn Al-Haytham Building (West), Thuwal 23955-6900, Kingdom of Saudi Arabia

[8] Department of Molecular Life Science, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan

expansion diseases were longer than other poly-Qs, even in healthy subjects, our results indicate that the evolutionary benefits of long and/or polymorphic poly-Q stretches outweigh the risks of long CAG repeats predisposing to pathological hyper-expansions. Molecular pathways in neurodevelopment requiring long and polymorphic poly-Q stretches may provide a clue to understanding why poly-Q expansion diseases are limited to neurodegenerative diseases.

Keywords STR polymorphism · Single amino-acid repeat · Human evolution · Triplet-repeat expansion disease · Database for human polymorphism (VarySysDB)

Abbreviation

STR	Short tandem repeat
SAR	Simple amino acids repeat
UTR	Untranslated region
CDS	Coding sequence region
c-triSTR	Coding trinucleotide short tandem repeat
INSDC	The international nucleotide sequence databases collaboration
H-GOLD/GDBS	Human-gene diversity of life-style related diseases/gene diversity database system
GO	Gene ontology
AHG2	Annotation data set for All Human Genes version 2
H-InvDB	H-invitational database
HIT	H-InvDB transcript
HIX	H-InvDB gene cluster defined by mapping of transcripts on genome sequence
GC3	Percentage of G or C at the third codon

Introduction

Nearly half of the human genome is occupied by repetitive sequences including interspersed and tandem repeats (Richard et al. 2008), in which tandem repeats are distributed every two kilobases (International Human Genome Sequencing Consortium 2001). Tandem repeats are classified according to the length of repeat unit as satellites (approximately 1000 bp or more), minisatellite (approximately 10–1000 bp), and microsatellite (approximately 1–9 bp, virtual synonyms: "simple sequence repeat", or "short tandem repeat (STR)", respectively (Gemayel et al. 2010; Tompa 2003). The simplest are the STRs, which represent a remarkably high mutability and high diversity (10^{-7}–10^{-3} mutations per locus per generation in eukaryotes) (Buschiazzo and Gemmell 2006). Replication slippage is a major contributor to the high variability in the number of repeating units of STRs, whereas that of minisatellites is caused by meiotic recombination events (Gemayel et al. 2010). Some STRs are known to be located within coding regions and to be related to neurodegenerative disease. Although the mechanisms of the existence or persistence of STRs in the genome have been studied in neutral processes (King 2012; Takezaki and Nei 2009), no systematic analysis of the relationship between evolution and the molecular function of STRs under selection pressure has been conducted. To understand the maintenance mechanisms of various STRs including disease-causing ones, we intend to exploit a fast-evolving feature and the genome-wide distribution of STRs in this study.

Generally, the numbers of repeating units (repeat length) of STRs are limited to a certain range with a bell-shaped distribution, although multimodal repeat size distributions are found for some STRs (Shriver et al. 1993). Several models explain the mechanisms that maintain the repeat length within a certain range (Buschiazzo and Gemmell 2006; Haasl and Payseur 2013; Kashi and King 2006; Rado-Trilla and Alba 2012). Briefly, the mutation rate in repeat length increases in a given locus as the repeat length expands, which also results in more length polymorphism in longer alleles (Gemayel et al. 2010). Contrary to the elongation process, the rate of contraction increases exponentially as the repeat length increases. Consequently, longer alleles tend to contract. Moreover, the rate of breakage of long STRs by a point mutation or a short insertion increases in longer allele (Buschiazzo and Gemmell 2006). Generally, STRs with long repeat tend to be highly variable in their number of repeating units. Occasionally, longer STRs change into an unstable mode of rapid evolution that results in hyper-expansion of repeat length and is known as repeat expansion disease (Laffita-Mesa et al. 2012; McIvor et al. 2010). The relationship between hyper-expansion and disease is complex given that they are not always correlated (Deka et al. 1999b). The mechanism of the hyper-expansion is thought to differ from that of STRs within the normal range, which originates from a slip-out of one DNA strand caused by transcription-coupled RNA-DNA hybrid formation (Grabczyk et al. 2007; Salinas-Rios et al. 2011). Furthermore, simultaneous transcription of sense and antisense RNA through a repeat track is also considered to involve these hyper-expansion of STRs, which enhances repeat instability leading to cell death via apoptosis (Lin et al. 2010, 2014; Lin and Wilson 2012).

It is not known why triplet repeat expansion diseases are generally limited to neurodegenerative and neuromuscular diseases (Paulson 2000). Rapid mutability of STRs is considered to promote evolutionary changes through changes in the binding manner of splicing factors (Hui et al. 2005), conformation of nucleic acid (Zhang et al. 2013), and methylation (Fukuda et al. 2013), which is known as the evolutionary tuning knob hypothesis [reviewed in (King 2012) and (Trifonov 1989)].

Given that the number of repeating units of the STRs causing triplet repeat expansion diseases are shown to be slightly longer on average with a larger variance for healthy human alleles than those of apes (Andrés et al. 2004), rapid mutability of these STRs has been intuitively considered in relation to human encephalization during human evolution, which has been awaiting an empirical examination (King 2012).

Difference in the strength of functional constraint among STRs within the human genome might be a clue to address the longstanding questions regarding demography and selection on STRs in human population (Deka et al. 1999a). Considering that STRs arise ubiquitously from simple sequences by the accumulation of substitution mutations (Ananda et al. 2013; Buschiazzo and Gemmell 2006), the present state of STRs (e.g., density, purity, and variation) in the human population is an evolutionary consequence of functional constraint. To understand the effect of functional constraint on the maintenance of STRs in various regions of the human genome, we comprehensively surveyed the human genome for STR density and presented that STRs in genic regions have evolved under different selective pressures than those in intergenic regions. Then, we focused on the fact that a limited number of STRs in coding regions are polymorphic in the number of repeating units. We showed that the properties of the repeat polymorphisms in trinucleotide STRs in coding regions differ from that expected based on their coding simple amino acid repeats (SARs). Namely we observed two different evolutionary stable states according to repeating amino-acid residues, one was represented by a poly-glutamine (poly-Q) repeat, and the other by a poly-proline (poly-P) repeat. We explained the origin of the two patterns by the relation between repeat length and selection pressure using SNP density difference. In particular, there was a confliction in the long repeats between functional advantage at the protein level and disadvantage at the DNA level in length polymorphism of the poly-Q stretch. We indicated that poly-Qs causing expansion diseases were longer even in healthy people than other poly-Qs. Moreover, genes containing longer poly-Qs or that have high length variation were enriched in genes with neuronal functions. These data suggest an advantage of long and/or variable poly-Q repeat length in neuronal function, which may be a clue to explain why the triplet repeat expansion diseases are associated with the nervous system.

Results

STR distribution and selection in the genome

To understand the differences in selection pressure on STRs between exonic regions and other genomic regions, we examined density of STR sites in the human genome using the polymorphic STR database, H-GOLD/GDBS (Tamiya et al. 2005). The H-GOLD/GDBS contained STRs that were demonstrated to be polymorphic in the Japanese population through empirical screening (i.e., pooled PCR system). The H-GOLD/GDBS is based on PCR primer sets designed to amplify genomic regions (marker regions) that contain at least one polymorphic STR marker and that are located approximately every 100 kilobases evenly throughout the whole genome to be used as genomic markers. We used the genomic distribution of these polymorphic STR loci to evaluate selection pressure (See Supplementary Files for the appropriateness of this method). Out of the 22,216 STR marker regions in the whole genome, we found in total 141 (0.63 %) overlapping m-RNA/cDNA (H-Inv transcript) regions, excluding hypothetical genic regions (see Supplementary Files for detail). Considering that the H-InvDB exonic region is 3.7 % (=112,525,104/3,076,781,887 bp) of the total genomic region, this suggests that fewer STRs are located in exonic region than expected from a random distribution. A comparison of exonic and whole genomic regions suggests a significant increase in exonic trinucleotide repeats (27 out of 141, 19.15 %, included in transcripts vs. 1505 out of 22,216, 6.77 %, in whole genome, $p < 10^{-6}$ two-sided binominal) and a significant decrease in exonic tetranucleotide repeats (7 out of 141, 4.96 %, connected to transcripts vs. 3891 out of 22,216, 17.51 %, in the whole genome, $p < 10^{-4}$ two-sided binomial, Fig. S1a). Di- and tri- nucleotide repeats have increased during mammalian evolution although the details of the evolutionary mechanism remain unknown (Astolfi et al. 2003; Guo et al. 2009). We confirmed the increased proportion of trinucleotide repeats in exonic regions, which suggests selection pressure against frame-shifting polymorphisms in the coding regions (Fig. S1b).

Focusing on exonic STRs and SARs sequences

To focus on the evolutionary features of exonic STRs and SARs, apart from the foregoing whole-genome assessment, we extracted repetitive sequences (≥5 repeating units) located in the published human transcriptome sequences. To collect transcript sequences comprehensively, three transcriptome databases were mined: H-InvDB 3.8, RefSeq, and EnsEMBL; this is called the All Human Genes 2 (AHG2) data set. Then, we compared SARs (≥5 repeating units) obtained by translating the transcriptome sequences.

Database search for STRs

In the AHG2 data set, we identified 56,743 STR-transcript pairs that were redundant with each other, including 12,171 non-redundant STRs. Among them, 6350 STRs were observed in validated transcripts with protein evidence (i.e., Category I to III in H-InvDB, See Methods, hereafter "validated STRs").

Database search for SARs

We identified 42,378 SARs in amino acid sequences deduced from the AHG2 data set. Among them, 5858 SARs were observed in amino acid sequences deduced from non-redundant representative transcripts, including 4984 SARs found in those with protein evidence (i.e., Category I to III, See "Methods").

Prediction of variable STRs

We determined polymorphism in the number of the repeating units (repeat-length polymorphism) by detecting repeat-length differences among repeat sequences extracted from transcript sequences, as well as by mining public international databases for human polymorphism, which is a combination of the following three detection methods via: (1) alignment of the transcript sequences, (2) a known polymorphic STR database in human, H-GOLD/GDBS, and (3) deletion/insertion polymorphism (DIP) in a polymorphism information database of NCBI, dbSNP (Fig. S2, See "Methods" for details). We aimed to detect all globally reported STR polymorphisms in human exonic regions using this combination. As a result, within the validated 6350 STRs, 939 were polymorphic (14.8 %) and identified in 787 gene loci defined by H-InvDB (HIXs).

Distribution difference between polymorphic and monomorphic STRs

Comparison of the 939 polymorphic STRs and remaining (i.e., monomorphic) 5411 STRs, showed a distinctive feature in the proportion of STRs located in CDS regions in that there were fewer polymorphic STRs (19.4 %, 182/939) than monomorphic STRs (36.2 %, 1958/5407, excluding 4 STRs overlapping the UTR-CDS boundary from 5411 monomorphic STRs, Fig. S3). Focusing on trinucleotide STRs, we identified 1410 trinucleotide STRs located in CDS (coding trinucleotide STRs, c-triSTR) including 161 polymorphic and 1249 monomorphic c-triSTRs (Fig. S3). Remarkably, there was a high proportion of c-triSTRs among the polymorphic STRs in CDS (88.5 %, 161/182), compared with that among monomorphic STRs (63.8 %, Fig. S3). This proportional difference in polymorphic STRs is consistent with the expectation that variation in the number of repeating units involves codon-frame changes except for trinucleotide repeats.

Distribution of repeats in CDS

We detected 1410 c-triSTR, whereas the number of SARs was 4984, when we counted them with the same criterion, more than four tandem repeating units in the human genome (Fig. 1). The identification of 3.5 times more

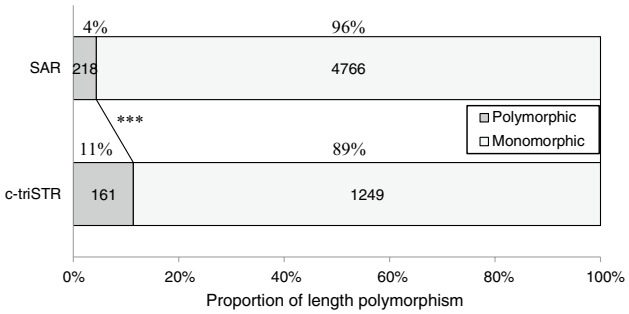

Fig. 1 Proportion of length polymorphism and number of repeat sites of trinucleotide STRs in coding regions (c-triSTRs) and SARs. The *asterisk* marks indicate a significant difference in the proportion of polymorphic repeats between SARs and c-triSTRs ($P < 10^{-16}$, Fisher's exact test). The numbers of c-triSTRs and SARs (4 times repeats<) are represented in *bars*

SARs than c-triSTRs, suggests that a substantial number of SARs are composed of multiple synonymous codons or trinucleotide repeats of less than five repeats. For example, HIT000001279 is a transcript containing five SARs in the deduced amino acid sequences, however only one of the five is translated from an STR. Each of the remaining four SARs is translated from a consecution of synonymous codons or trinucleotide repeats of less than five repeats (Example 1 in Table 1, http://h-invitational.jp/varygene/gene.htm?id=191780). This raises a hypothesis that SARs have not been simply produced as consequences of STR generation and as neutral elements, but some of the long SARs have been maintained during evolution by combining multiple shorter trinucleotide repeats, due to the functional roles of long SARs. It is already known that at least some long SARs have special functions at the protein level (Mularoni et al. 2007) (See "Discussion").

We also found a suggestive feature in length polymorphism. We identified 218 polymorphic SARs out of 4984 SARs (4.4 %, 218/4984). Contrastingly, the proportion of polymorphic repeats in c-triSTRs was 11.4 % (161/1410). This is 2.6 times higher than that of SARs and the proportional difference is significant ($P < 2.2 \times 10^{-16}$ by a two-sided Fisher's exact test, Fig. 1). This suggests an alternation of selective pressure in at least part of polymorphic c-triSTRs with a special role in an evolutionary context.

Homogeneous codon usage of poly-P stretches

We found different tendencies in codon usage between poly-glutamine (poly-Q) and poly-proline (poly-P) stretches in the human transcriptome. In total, poly-Q stretches in our SAR data include 470 codons consisiting of 278 (59 %) CAG and 192 (41 %) CAA codons, respectively (Fig. 2). This percentage is consistent not only with previous work (Siwach et al. 2006) but also with the average

Table 1 Examples of STR and SAR composed of mixed codons

Example	HIT-ID, HUGO Gene Name	SAR (AA seq.)			Transcript sequence[a]	STR (cDNA seq.)		
		Position in transcript	Repeat unit	# Repeat unit		Position in transcript	Repeat unit	# Repeat unit[b]
1	HIT000001279 CHD8	1342–1356	E	5	(gaa)4 + gag			<5
		4876–4890	R	5	cgg + cga + agg + cgg + cgt			<5
		5794–5835	S	14	(tct)4 + (tcc + tca)2 + (tcc + agc)3			<5
		7030–7047	H	6	cat + (cac)5	7033–7047	cac	5
		7057–7077	H	7	cac + cat + (cac)2 + cat + c ac + cat			<5
2	HIT000260818 RNF44	1391–1423	P	11	ccc + (ccg)2 + ccc + cca + ccc + cca + ccc + (cca)2 + ccc			<5

[a] Repeat unit (i.e., codon) sequences are depicted in parentheses followed by number of repeating unit

[b] Less than five times repeat (indicated as " <5") were not counted as tandem repeat sequences in this study

GC content at the third-codon position (GC3) in human protein-coding genes (Elhaik et al. 2009; Tatarinova et al. 2013). In contrast, poly-P stretches use the four codons almost homogenously; 601 (26 %) CCC, 479 (21 %) CCG, 570 (25 %) CCT, 654 (28 %) CCA, respectively (Fig. 2). This is the least GC3 among SARs composed of >100 codons. This tendency in the difference between poly-P and poly-Q was manifested both in polymorphic and in monomorphic SARs (Fig. S4).

Poly-P with mixed codon vs. Poly-Q with pure repeats

We compared the number of loci between STRs and SARs (i.e., number of SARs over that of STRs, #SAR/#STR) for each repeating amino-acid residue. This comparison suggests that there are two types in repeats regarding length polymorphism (Fig. 3). One type, such as the poly-Q repeat, shows little difference between the numbers of SARs and c-triSTRs, [#SAR/#STR <1 for polymorphic loci and #SAR/#STR <2 for monomorphic loci (Q-type)]. However, the poly-P repeat shows a converse relationship with significant differences between SARs and c-triSTRs (P-type). We found that the number of synonymous codons coding an amino acid associate with this type of classification; i.e., amino acids of the Q-type are coded by two synonymous codons, while three or more synonymous codons for those of the P-type.

The monomorphic rich tendency and vast difference in counts between SARs and c-triSTRs in poly-P repeat (Fig. 3) are consistent with the even proportion of the four codons in poly-P repeats (Fig. 2). The poly-P repeats tend to be encoded by mixed codons (Table 1, Example 2). This tendency of the P-type suppresses slippage of tandem repeats at the DNA level, because such SARs are composed of a mixture of short STRs. On the contrary, poly-Q repeats tend to be encoded by repeats of the same codon or pure repeats, which have a higher potential for replication slippages resulting in more length polymorphism than those in poly-P repeats. Consequently, polymorphic poly-Q repeats are frequently observed (Fig. 3).

Length of STRs and SARs in poly-Q and poly-P stretches

The median numbers of the repeating unit for all SARs, poly-Q SARs, and poly-P SARs are equally 6 repeat counts in each category, and those of c-triSTRs show little differences among them, (5, 6, and 5 repeat counts, respectively). The polymorphic and monomorphic poly-Q SARs showed significantly longer repeats than those of other amino acids (Two–tailed Mann–Whitney-Wilcoxon test, $P < 0.0001$, Fig. 4). Those of c-triSTRs coding poly-Q were also significant, but the tendencies were weaker than those of SARs ($P < 0.001$ and $P < 0.01$, respectively, Fig. 4). Taking into account the positive correlation between number of repeating units and diversity of repeat polymorphisms (Ogasawara et al. 2005), highly variable with longer alleles is considered to be an advantageous state in poly-Q stretches in humans. In poly-P stretches, a slightly shorter tendency was observed in monomorphic c-triSTR than that coding for other amino acids. This represents the tendency of poly-P stretches to be composed of short STRs of mixed codons. The relationships between repeat length and polymorphism in the repeats coding other than Q or P are shown in supplementary material (Fig. S5).

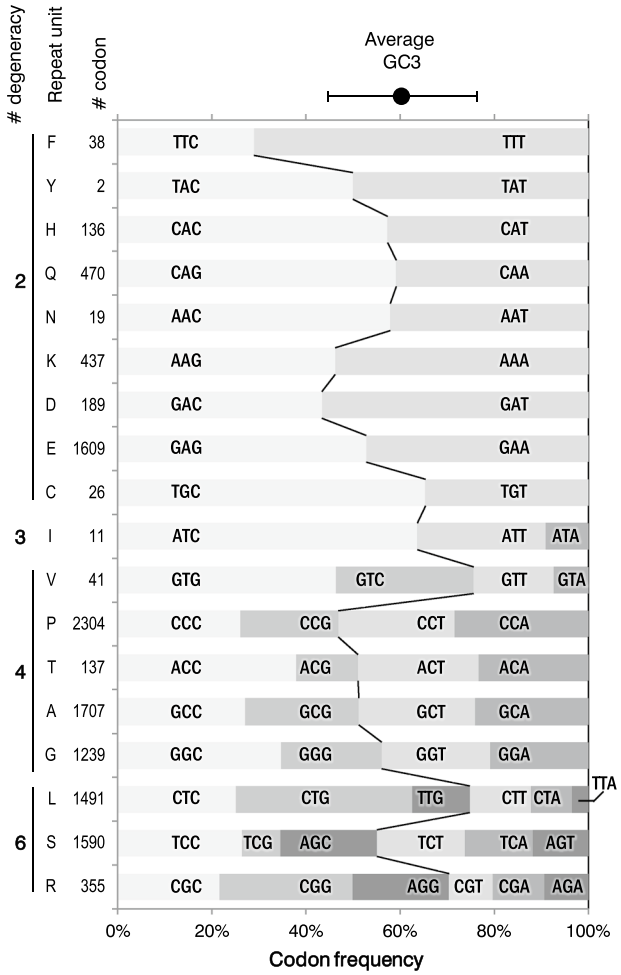

Fig. 2 Codon usage of repeat units of SARs. Frequencies of codon usages are shown with total codon counts (# codon) for each repeat unit (i.e., amino-acid residue). The line divides between GC (*left*) and AT (*right*) for third codon positions. *Bold numbers* indicate numbers of degeneracy in each codon (#degeneracy). The *circle with horizontal bars* shows the mean (60 %) and SD (17 %) of the GC composition of 3rd codon (GC3) of human genes (Elhaik et al. 2009). Poly-W was not observed in our data. Three codons specifying Poly-M were observed and omitted because only ATG codes for Methionine

Length of SARs containing SNPs

SNP density

The poly-Q SARs that were polymorphic in repeat length contained notably more synonymous SNP sites than those of other amino acids (Fig. 5a, two-tailed Fisher's exact tests, P value = 0.03474). However, the monomorphic poly-Q SARs did not show such differences. Because synonymous changes divide a STR into two shorter STRs without changing the length of coding poly-Q SARs, this increase of synonymous SNPs indicates that the lengths of poly-Q c-triSTRs are constrained to being short at the DNA level, while poly-Q SARs are favored to be long at the protein level. This difference in selection pressures on repeat lengths between the DNA and protein levels enables long poly-Q SARs to exist without long STRs (i.e., CAA and CAG repeats), which are prone to hyper-expansion.

Number of the repeating unit (repeat length)

As intuitively expected, SARs containing synonymous SNPs are longer than those without any SNPs as a whole (Two–tailed Mann–Whitney-Wilcoxon test, $P = 7.57 \times 10^{-10}$, triangles in Fig. 4). Notably, all poly-Q SARs containing synonymous SNP sites are not shorter than the medians in both of the polymorphic and monomorphic classes (Fig. 4, Two–tailed Mann–Whitney-Wilcoxon test, $P = 0.0001679$ and $P = 0.0001093$ for polymorphic and monomorphic classes, respectively). This is a characteristic property of poly-Q SARs, which may suggest a stronger tendency of synonymous SNP interruption in long poly-Q SARs than that in other SARs (e.g., the synonymous rs473915 SNP changes between CAA and CAG, Fig. 5b).

GO analyses

To find clues about maintenance mechanisms of repeat polymorphisms, we investigated the molecular function of poly-P and poly-Q using GO analysis by following three approaches. First, we manually collected GO terms (Molecular Functions) representing glutamine-rich and proline-rich 'motifs' from the InterPro database (GO-I, "Q-motif" and "P-motif" in Fig. 6a). Second, GO terms of 'genes' containing poly-Q and poly-P stretches were searched similarly (GO-II, "Q-gene" and "P-gene" in Fig. 6a). Third, we conducted gene set enrichment analyses using the DAVID database (Huang et al. 2009a, b) to collect overrepresented annotations of the genes containing poly-Q and poly-P stretches (GO-III, Tables 2, S1–S4). In GO-I results, the most frequently observed GO terms of glutamine rich motif was "binding to nucleic acids" and "transcription regulation" ("Q-motif" in Fig. 6a). Searching for GO terms of genes containing poly-Q stretches (GO-II) showed that "binding to nucleic acids" was the most frequently observed GO term ("Q-gene" in Fig. 6a). The GO-II result for poly-Q-containing genes was consistent with those of our gene set enrichment analysis (GO-III) representing molecular functions regarding regulation of transcription (Table 2). These results suggest that poly-Q stretches play roles in binding nucleic acids and endow host genes with functions to regulate transcription and metabolism. Contrary to poly-Q, the most remarkable GO

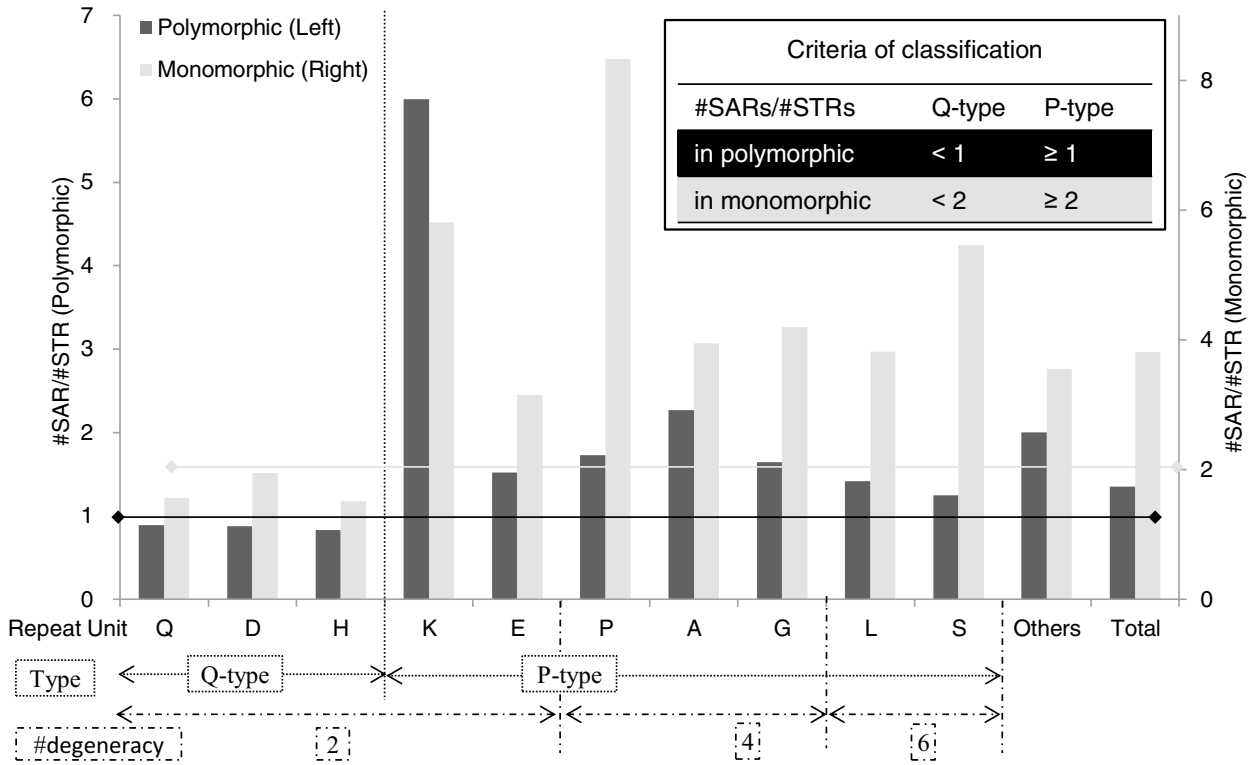

Fig. 3 Ratios of SAR and STR sites for the top 10 amino-acid residues. The ratios of the number of SAR sites to c-triSTR sites are coded by the presence (*black*) or absence (*grey*) of repeat length polymorphism. *Q-type*: the ratio in polymorphic <1 and the ratio in monomorphic <2, *P-type*: the ratio in polymorphic ≥1 and the ratio in monomorphic ≥2. Numbers of synonymous codons (#degeneracy) and type classification are depicted by the broken and *dotted lines*, respectively. The repeats composed of ten amino-acid residues (R, T, C, I, V, F, M, N, Y, W) are summarized and shown as "Others." The numbers of SARs/STRs for polymorphic and monomorphic sites are as follows; Q: 48/54, 230/147; D: 7/8, 107/55; H: 5/6, 74/49; K: 6/1, 250/43; E: 38/25, 828/263; P: 19/11, 791/95; A: 34/15, 616/156; G: 23/14, 466/111; L: 17/12, 588/154; S: 15/12, 546/1006; others: 6/3, 270/76; total: 218/161, 4766/1249

term of proline (P)-rich motif (GO-I) was "binding to proteins" ("P-motif" in Fig. 6a), which seems discordant with the annotations of genes containing poly-P, "binding to nucleic acids" (GO-II, "P-gene" in Fig. 6a, GO-III, Tables S3 and S4). One explanation is that the most well-known primary molecular function of poly-P stretches is protein binding, however proline-stretch motifs are also found in genes that bind to nucleic acids via other nuclear-binding domains such as poly-Q stretches. Thus, we examined the cases in which two sites of SARs were juxtaposed with each other in our dataset (Table 3). As a result, we observed 38 cases of adjacent SARs pairs among the 4984 SARs in total, in which 13 cases were a combination of poly-P and poly-Q SARs. This is unlikely to occur in random combinations ($P < 10^{-14}$, assuming a binominal distribution) selected from the given SARs frequencies observed (810 poly-P and 278 poly-Q were observed in 4984 SARs in total). This analysis also showed excess combinations of poly-E and poly-D ($P < 10^{-7}$), excluding combinations with one pair. This is also a combination of P-type and Q-type. These results suggest that specific combinations of different SARs, especially P-type and Q-type, play roles as functional domains.

Biological function specific to polymorphic poly-Q stretches

We compared the enriched GO terms of genes containing polymorphic and monomorphic poly-Q stretches to uncover differences in biological function. In spite of no significant differences between them in GO terms of molecular function, we found biological processes enriched only in genes with polymorphic poly-Q repeats (bold in Table 2). These biological processes are related to apoptosis and regulation of nervous systems. The association between triplet repeat instability and apoptosis presented here (Table 2) is

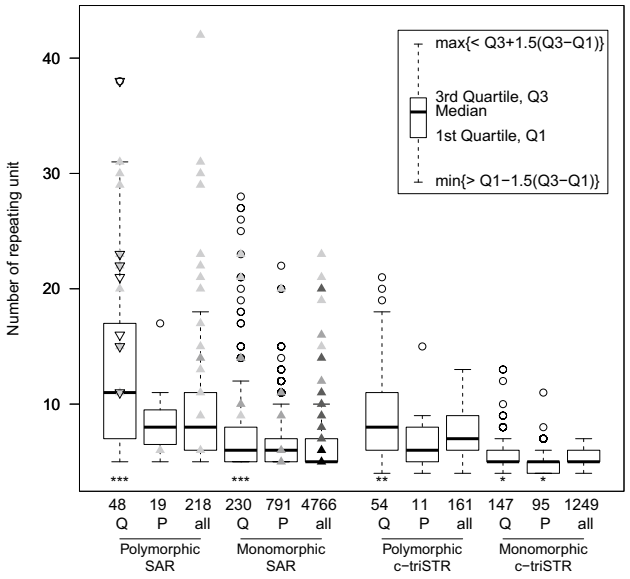

Fig. 4 Distribution of the number of repeating units of poly-Q and poly-P with all repeats. The numbers of repeating units (repeat length) of SARs and trinucleotide STRs in coding regions (c-triSTRs) for repeats of poly-Q (Q), poly-P (P), and over all amino-acid residues (all) are shown in the boxplot. *Asterisks* indicate significantly different distributions from the repeats of other amino-acid residues within the same classes (Two–tailed Mann–Whitney-Wilcoxon test, $*P < 0.01$, $**P < 0.001$, $***P < 0.0001$). The Q-SARs with synonymous SNP sites (*closed gray triangles* whose *darkness* represent degrees of multiple hits in the range of [1, 38]) and those known to cause repeat expansion diseases (*inverted open triangles*) were not shorter than the medians of all Q-SARs. The counts of repeats in each category are displayed numerically below the *horizontal axes*. Outliers (*open circles*) are not displayed in the "all" categories, and their maximum numbers of repeating units are 42, 65, 21, and 50 for polymorphic SARs, monomorphic SARs, polymorphic STRs, and monomorphic STRs, respectively

congruent with the experimental results that simultaneous sense and antisense transcription at CAG-repeats enhances instability and induces apoptosis (Lin et al. 2014; Lin et al. 2010; Lin and Wilson 2012).

Discussion

Two types of selection pressure on STRs

Our comprehensive database searches of human transcript sequences indicate two types of adaptation to prevent unfavorable hyper-extension of STRs at the nucleotide level. One is typically observed in the poly-Q stretches characterized by increased frequencies of synonymous SNPs within repeats (Q-type, Figs. 4 and 5a). The other is represented by the poly-P stretches that are prone to be constituted by a succession of synonymous codons (P-type, Figs. 2, 3, and 4). Furthermore, we discovered that selective pressures at the amino acid and DNA levels are different from each other in each type of repeat. At the amino acid sequences (SARs) level, SARs forming poly-Q stretches were characterized by long and high variability in repeat number (Fig. 4). Given that the tendency of poly-Q SAR loci is conserved between human and mouse genomes (Mularoni et al. 2007), the property of long repeats in poly-Q stretches seemed to be based on an advantageous protein function. However, long poly-Q stretches are prone to incur hyper-expansion at the DNA (c-triSTRs) level. Since glutamine is encoded by two synonymous codons (i.e., CAA and CAG), the repeats of each codon are necessarily prone to be longer than those specified by four synonymous codons such as poly-P. Long CAG repeats form hairpin structures that are frequently harmful; however, those of CAA repeats do not (Sobczak et al. 2010). Nevertheless, 278 out of 470 codons specifying poly-Q stretches (59.1 %) are CAG, which is approximately an equal proportion of GC3 (Fig. 2). This preference of CAG to CAA is congruent with previous studies that showing a higher GC content of whole STR regions than of other exonic regions (Alba and Guigo 2004; Siwach et al. 2006). This also indicates that selection pressure for codon usage does not exist when new poly-Q SARs are generated, but arises after long poly-Q stretches have grown. Our results showing significantly increased frequencies of synonymous mutations at poly-Q stretches are explained by interruptions of long CAG repeats by a CAA codon leading to the avoidance of harmful hyper-expansion during replication, as well as avoidance of hairpin structure formation after transcription (Fig. 5a). This explanation is congruent with knowledge that interruptions by synonymous mutations in poly-Q-coding regions prevent instability and hyper-expansion of STRs (Choudhry et al. 2001; Kurosaki et al. 2006; Sobczak and Krzyzosiak 2004). Given this, the selection pressure to maintain short CAG repeats at the DNA level is an adaptation to prevent harmful events such as hyper-expansion and hairpin formation of CAG repeats. On the contrary, we showed that poly-P stretches are shorter than other SARs on average at the amino acid sequence level (Fig. 4). These results are consistent with a known tendency that hydrophobic SARs are frequently shorter, while hydrophilic SARs are longer (Faux 2012). At the DNA sequence level, we demonstrated that the poly-P regions have been selected to maintain the repeat number as short as possible (Fig. 4) by composing an even mixture of four codons (CCU, CCC, CCA, CCG) coding for proline, in contrast to glutamine encoded by two codons that are occupied proportional to the GC3 of human genes (Fig. 2). Our result shows that poly-P STRs have been maintained to be around a minimum repeat number of c-triSTRs, with a cutoff of five-times (Fig. 4). This is consistent with the fact that a shift in mutation mechanism from nucleotide insertion to repeat expansion by slippage is considered to occur at approximately four-times repeat for trinucleotide repeats (Ananda et al. 2013). On these bases, poly-P stretches at the protein level are maintained by the avoidance of STR formation, which results in no

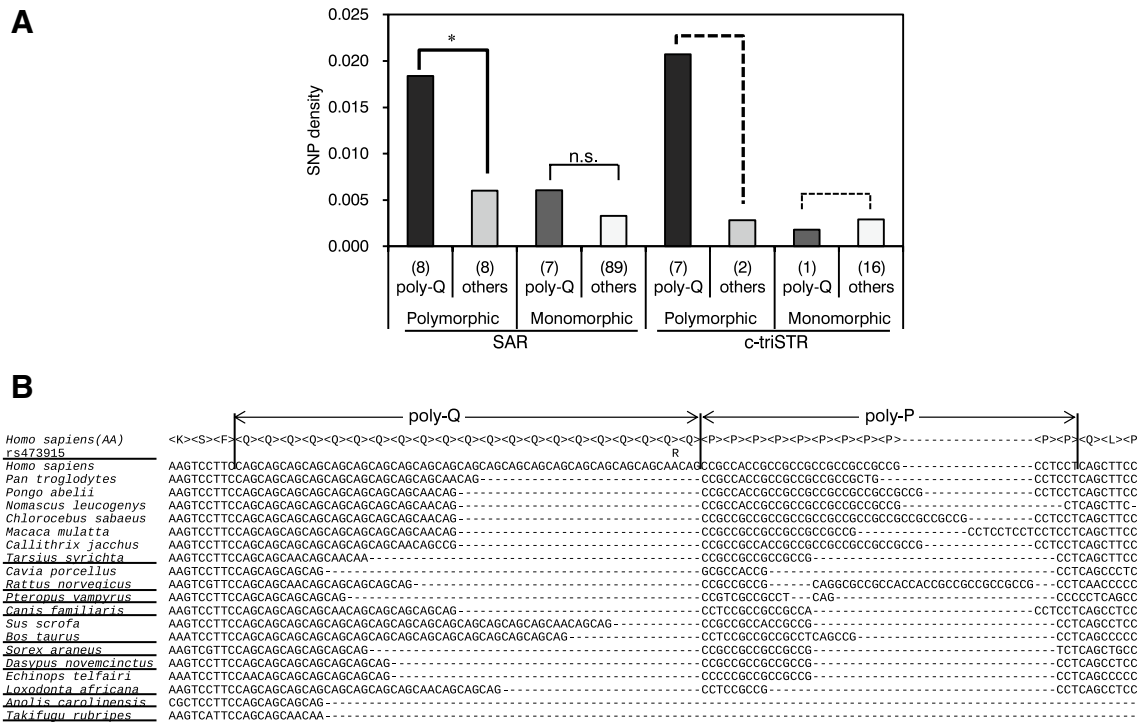

Fig. 5 **a** Frequencies of SARs and trinucleotide STRs in CDS (c-triSTR) overlapping synonymous SNPs. The densities of synonymous SNPs overlapping SARs and c-triSTRs were compared between poly-Q repeats and others. The synonymous SNP densities of polymorphic poly-Q repeats (Q) are significantly higher than those of non-poly-Q (others) repeats in SARs; Fisher's exact test, $P < 0.05$ (*). This tendency is also observed in c-triSTRs although few synonymous SNP sites interrupt c-triSTR. The numbers of synonymous SNPs are shown in parentheses. **b** Alignment of protein and DNA sequences around poly-Q and poly-P stretches of the Huntingtin gene (exon 1 partial). The partial sequence of the huntingtin gene (HTT) for human protein (*top*) and vertebrate DNA sequences (*below*) with a human SNP (rs473915) were aligned. *Horizontal lines* indicate the following taxonomic groups; Primates, Glires, Chiroptera, Carnivora, Cetartiodactyla, Insectivora, Xenarthra, Afrotheria, Reptilia, Osteichthyes. Human poly-Q and poly-P stretches are indicated by *vertical lines*

trigger for triplet expansion in the DNA level. Thus, we demonstrated that the selection to avoid hyper-expansion of c-triSTRs exists in all kinds of amino-acid forming SARs as the two contrasting mechanisms, Q-type and P-type, respectively. Thus, our comprehensive searches of human transcriptome sequences have uncovered two different mechanisms to solve the conflict in selection pressure between SARs and STRs.

Association of poly-Q and poly-P motifs

We showed functional differences between poly-Q and poly-P stretches, by linking a discrepancy in motif functions between the poly-Q and poly-P (i.e., "binding to nucleic acid" for "Q-motif" vs. "binding to protein" for "P-motif" in GO-I) and a consistency in gene functions between genes containing poly-Q and those containing poly-P ("binding to nucleic acid" at "Q-gene" and "P-gene" in GO-II) (Fig. 6a). This suggested that poly-P stretches assist the function of flanking poly-Q stretches. Short poly-Qs (three or six glutamine residues) tend to form a poly-P type II (PPII)-like helix, but long poly-Qs (nine or 15 glutamine residues) form a β-sheet structure that causes the formation of amyloid-like aggregates (Perutz et al. 1994; Takahashi et al. 2010). The adjacent poly-P inhibits β-sheet formation, which can increase the length of the adjacent poly-Q (Bhattacharyya et al. 2006; Darnell et al. 2007; Mishra et al. 2012; Siwach et al. 2011). Evolutionary studies on huntingtin gene (*HTT*) sequences are consistent with this explanation. The lengths of poly-Q stretches have increased during vertebrate evolution, which followed by the insertion of poly-P stretches to the C-terminus of the poly-Q stretches (Tartari et al. 2008) (Fig. 5b). A study using model animals showed that *HTT* became to play important roles in neuronal development in the vertebrate lineage (Lo Sardo et al. 2012). Moreover in healthy humans, an increase of grey matter within the pallidum with increasing long CAG-repeats of *HTT* is known (Mühlau et al. 2012). The juxtaposition between poly-Q and poly-P stretches frequently observed by our comprehensive survey of human genes is likely a consequence of the benefit of adjacent poly-P to poly-Q stretches (Table 3). Our study shows that this juxtaposition is not limited to the *HTT*

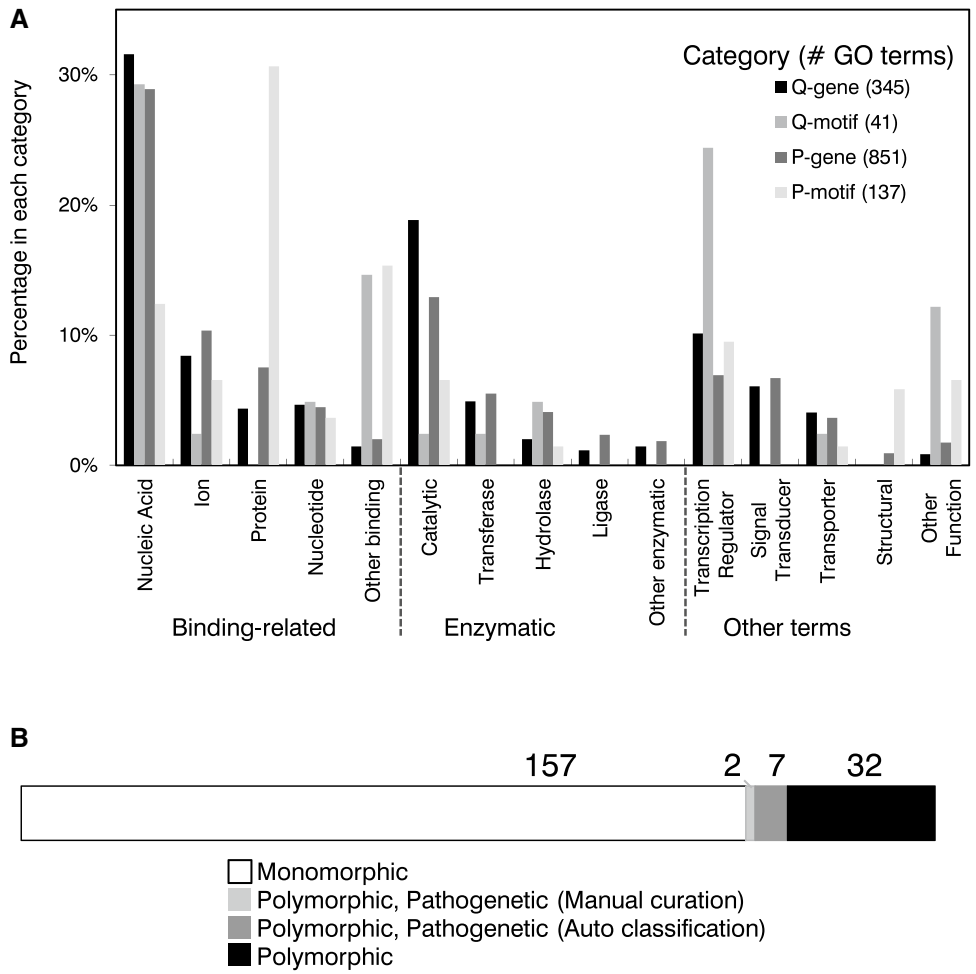

Fig. 6 **a** GO-terms associated with SAR motifs (GO-I) and genes containing SARs (GO-II). The GO-terms associated with glutamine or proline rich motifs obtained by manual curation of keyword searches in InterPro (GO-I) are shown as "Q-motif" and "P-motif," respectively. The GO-terms that are associated with genes containing poly-Q or poly-P repeats (GO-II) are shown as "Q-gene" and "P-gene," respectively. The total numbers of GO terms (Molecular functions) in each category are shown in parentheses. **b** Number of genes encoding poly-Qs. Those with length polymorphism are rare but include all of the genes with pathogenetic poly-Qs. This is significantly unlikely under random occurrence ($P < 10^{-5}$, Fisher's Exact Test). Two pathogenetic poly-Qs were confirmed as polymorphic by manual curation using dbSNP136 (*light gray*) and conjoined with polymorphic poly-Qs categorized by automated classification using dbSNP125 (*dark gray*). 13 genes containing both of polymorphic and monomorphic poly-Qs are counted as polymorphic, which includes 2 pathogenetic genes

gene but is a general phenomenon of human genes, which has been partially known (Rado-Trilla and Alba 2012; Ramazzotti et al. 2012). Thus, the two classes of SARs represented by poly-Q and poly-P stretches are associated with each other to optimize common biological functions, such as binding. This association could be another adaptation to escape hyper-expansion of STRs.

Molecular function and biological processes involving poly-Q stretches

The observed relation between molecular function (Table 2; Fig. 6a) and biological processes (Table 2) of GO terms for poly-Q-containing genes is straightforward. These are related to the regulation of transcription and biosynthesis, which is consistent with previous studies suggesting an overrepresentation in functional categories including developmental processes, signaling, and gene regulation (Huntley and Clark 2007; Kozlowski et al. 2010; Legendre et al. 2007; Vinces et al. 2009), as well as enzymatic or biosynthetic function (Siwach et al. 2006). These essentially require the molecular function of 'binding to molecules' (GO-III, Table 2). Generally, poly-Q stretches are observed as intrinsically disordered (ID) regions that are flexible domains suitable for binding (Gojobori and Ueda 2011; Rees et al. 2012; Takahashi et al. 2009). This accounts for SAR tracts playing a role as spacers between functional domains (Siwach et al. 2006).

Table 2 Overrepresented GO terms in genes containing poly-Q stretches with difference in rank between polymorphic and monomorphic classes

ΔRank	Molecular Function (Top 11)	Biological Process (Top 20)
(64,∞)	-	●**Cell death** ●**Death** ○**Regulation of cell development** ○**Regulation of neuron differentiation**
(32,64]	-	-
(16,32]	-	-
(8,16]	●RNA polymerase II transcription factor activity ○RNA binding ○Zinc ion binding	-
(4,8]	-	○Positive regulation of transcription, DNA-dependent
(2,4]	○Sequence-specific DNA binding	●Transcription ○Regulation of transcription, DNA-dependent ○Positive regulation of macromolecule biosynthetic process ○Positive regulation of RNA metabolic process ○Positive regulation of cellular biosynthetic process ○Positive regulation of macromolecule metabolic process
(0,2]	●Transcription factor binding ●Transcription cofactor activity ●Transcription activator activity ○Transcription factor activity ○Transcription repressor activity	●Regulation of transcription ●Regulation of RNA metabolic process ○Positive regulation of transcription ○Positive regulation of gene expression ○Positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process ○Positive regulation of nitrogen compound metabolic process ○Positive regulation of biosynthetic process ○Positive regulation of transcription from RNA polymerase II promoter
0	●Transcription regulator activity ●DNA binding	●Regulation of transcription from RNA polymerase II promoter

Higher ranked terms in polymorphic class are indicated by darker circles. Bold indicates terms enriched only in polymorphic class

Table 3 Juxtaposing SARs observed and deviation from random distribution

SAR Combination (1st & 2nd)	#Pairs	Proportion in all pairs	#1st SAR	#2nd SAR	Probability[a]
Q & P	13	0.342	278	810	2.66×10^{-15}
E & D	6	0.158	866	114	2.04×10^{-8}
S & G	4	0.105	561	489	0.001
E & A	3	0.079	866	650	0.092
E & Q	2	0.053	866	278	0.037
F & V	1	0.026	28	21	1.57×10^{-6}
S & T	1	0.026	561	58	0.005
P & T	1	0.026	810	58	0.009
A & H	1	0.026	650	79	0.011
A & R	1	0.026	650	128	0.027
K & E	1	0.026	256	866	0.147
A & G	1	0.026	650	489	0.254
S & A	1	0.026	561	650	0.307
P & L	1	0.026	810	605	0.445
E & P	1	0.026	866	810	0.641
Total	38	1.000	4984[b]	4984[b]	–

[a] Probability of neighboring two different SARs more than observed assuming binominal distribution of observed total number (i.e., 38 times) trials

[b] Total number of SARs detected SARs in the human genome

Molecular basis of harmful hyper-expansion

The characteristic cellular process in poly-Q protein toxicity is forming amyloid-like aggregates (Burke et al. 2013; Takahashi et al. 2010). Hyper-expanded poly-Q stretches are known to cross-liked to other proteins such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH), which form aggregates (Guzhova et al. 2011). Moreover, longer poly-Q repeats in human THAP11 are associated with more intracellular aggregations (Yin et al. 2014). This leads to conformational change in a poly-Q length-dependent manner (Vachharajani et al. 2012).

In addition to protein toxicity, pathological mechanisms of repeat expansion diseases have also been studied from the toxicity of expanded CAG/CUG repeats in RNA (RNA toxicity). Those include forming RNA foci that trap important proteins required for cellular process, forming double-stranded RNA that causes silencing, and binding directly to splicing factors (Fiszer and Krzyzosiak 2013). Of note, these RNA toxicities are known to lead to neurodegeneration (Mohan et al. 2014). Although multiple pathological mechanisms of triplet repeat expansion disease exist, once hyper-expansion occurs and is transcribed, which causes serious damage even if the transcript is not translated. Either way, the presence of long repeats increases the risk of incurring harmful hyper-expansion effects.

Repeat expansion diseases and repeat length polymorphism

We have shown that the 4984 SARs contained 48 polymorphic poly-Q SARs in 39 genes, in total (Fig. 6b, Table S5). It is known that all of the nine poly-Q SARs causing expansion diseases are polymorphic. Although our methods predicted seven out of the nine as polymorphic poly-Q STRs, the other two poly-Qs were also polymorphic according to dbSNP Build 136 or later (Table S5 and S6). Adding these two genes into the 39 polymorphic gene set, the fact that the 41 genes with polymorphic poly-Q SARs out of 198 poly-Q SARs (20.7 %) contained the nine genes associated with repeat expansion disease is significantly unlikely if repeat expansion diseases develop randomly from both polymorphic and monomorphic poly-Q stretches ($P < 10^{-5}$, Fisher's Exact Test, Fig. 6b). This suggests that length polymorphism is associated with repeat expansion diseases.

Our GO-III analysis showed that the 41 genes containing the polymorphic poly-Q SARs were enriched for transcription regulatory activity (39.0 %, Table 2) as the most frequent molecular function. This GO term, transcription regulatory activity, is ranked first in both genes with polymorphic and monomorphic poly-Q SARs (Table 2). Other GO terms of molecular function are also ranked high in both of the polymorphic and monomorphic gene sets. However, we have shown that four out of the top twenty biological processes are enriched in genes containing polymorphic but not monomorphic poly-Q SARs (bold in Table 2). These four are related to apoptosis and regulation of neuronal development. In fact, we observed that at least 17 of the 41 (≥41.5 %) genes coding polymorphic poly-Q stretches are annotated to be related to neuronal function in the UniProt (The UniProt Consortium 2014) (Table S5). This may account for why triplet repeat expansion diseases are limited to neurodegenerative and neuromuscular diseases. Since length polymorphism tends to be accompanied by long repeats (Gemayel et al. 2010; Ogasawara et al. 2005), the association of the harmful effects of hyper-expansion with neural function can be attributed to their long and polymorphic nature (i.e., hyper-expansions tend to occur at long repeats).

Polymorphic CAG/CAA repeats are known to be subjected to significantly less selective constraints than monomorphic repeats in various mouse strains (Ogasawara et al. 2005). Generally, CAG repeats causing expansion diseases have become longer during the evolution of primates (Andrés et al. 2004; Mularoni et al. 2010). Moreover, our comprehensive search showed that these disease-causing poly-Q SARs are longer than other poly-Qs, even in comparison to polymorphic poly-Q SARs (Fig. 4). To transcribe long CAG repeats, an extra molecular system involving transcription elongation factor SPT4 is required (Liu et al. 2012). Considering such an additional cost and high risk of hyper-expansion for long CAG repeats, selection pressure would limit the lengths of poly-Q SARs to shorter ranges than those observed unless long poly-Q SARs have some functional benefit. In fact, ranges of the number of repeats for normal and pathological are adjoining each other (Table S6). This implies the existence of a molecular basis that is advantageous for long poly-Q SARs, which maintain the balance in evolutionary forces between benefit of long repeats at the protein level and cost at the DNA level.

Association between the disease-causing repeats and the existence of benefits from long and/or highly polymorphic poly-Q repeats indicates molecular pathways shared by genes involving the expansion diseases. Several models of common molecular pathomechanistic networks among repeat expansion diseases have been proposed (Chan 2014; Labbadia and Morimoto 2013; Okazawa 2003; Shao and Diamond 2007; Tsoi and Chan 2014). Key players in such common networks are proteins that bind to poly-Q stretches and are associated with multiple neurodegenerative diseases. For example, polyglutamine-binding protein 1 (PQBP1) interacts with the poly-Q stretches found in huntingtin and androgen receptor which are hyper-expanded in Huntington's disease and spinal and bulbar muscular atrophy, respectively (Waragai et al. 1999). Moreover, the binding affinity of PQBP1 positively correlates with the length

of the poly-Q stretches of ATAXIN-1 whose hyper-expansion is associated with spinocerebellar ataxia 1 (Okazawa et al. 2002). PQBP1 is known as a regulatory factor of transcription and alternative splicing (AS) by binding to various molecules at WW and CTD domains as well as to poly-Qs at the poly-Q binding domains (Mizuguchi et al. 2014). Moreover, the depletion of PQBP1 in mouse neurons alters the AS pattern in mRNA enrolling neurite growth and neuron projection, such as *NCAM1*, resulting in a reduction of dendrite growth, which contributes to developmental stage differences in AS and causes PQBP1-related neurological disorders (Wang et al. 2013). This indicates a critical role for PQBP1 in the development of complex brains by regulating AS with tissue specificity through association with molecules within the splicing complex. Therefore, PQBP1 is one of the most likely candidates of key players intervening among long poly-Q repeats, the regulation of alternative splicing, and neural development.

Evolutionary mechanism of STRs with biological function

Two aspects (neutral and non-neutral) of the mechanisms of existence or persistence of STRs in the genome have been studied (King 2012; Takezaki and Nei 2009). First, the existence of functionally neutral STRs has been attributed to a balance between the generation/expansion of repeats by replication slippage and their extinction/contraction by genetic drift, replication slippage, and nucleotide substitutions leading to the interruption and degradation of repeats [i.e., 'life cycle' model of non-functional STR (Buschiazzo and Gemmell 2006)]. Second, the length variations of STRs in some genomic locations have been demonstrated to affect phenotypes. Consequently, they undergo some form of selection. Our survey of genomic locations using polymorphic STR makers of the H-GOLD/GDBS data suggests that a much smaller number of STRs exist in exonic than in inter-genic regions (Fig. S1a). This suggests a much stronger selection on STRs in exonic regions than that in inter-genic regions. Some of STRs located in regulatory regions have been known as "evolutionary tuning knobs" that quantitatively regulate gene expression depending on their number of repeats (Gebhardt et al. 1999; King 2012; Trifonov 1989). Moreover, effects of STR variation on phenotypes are not restricted to quantitative fine tuning but also influence dynamic changes in development (Birge et al. 2010; Fondon and Garner 2004; Galant and Carroll 2002) and behavior (Hammock and Young 2005) by alternation of mRNA splicing (Lorenz et al. 2007) and amino acid translation (Chang et al. 2001; Erwin et al. 2006; Weiser et al. 1989), which are called "evolutionary switches." These evolutionary switches bring about not only selection of individuals carrying advantageous alleles but also a shift to heterogeneous populations with various alleles that is advantageous for organisms in rapid adaptation to variable environments [e.g., counter strategy to host immune system in *Haemophilus influenza*, (Erwin et al. 2006) and diversifying sexual selection in the stalk-eyed fly, *Teleopsis dalmanni* (Birge et al. 2010; Cotton et al. 2014)]. Furthermore, according to the hypothesis focusing on rapid evolution of the human brain (social brain hypothesis) (Dunbar 1998), highly variable STRs leading to neurodegenerative disease may work as an evolutionary switch, by contributing to the increase of the diversities of populations in sensory, motor and cognitive abilities, which are a definitive feature of the human population with highly structured society (Fondon et al. 2008; Nithianantharajah and Hannan 2007).

Diversity and complexity of brain: AS role in human evolution

Because long repeat apt to be variable in repeat length, the selection pressure for extension of poly-Q repeats in genes functioning in neuronal development can foster diversity in such genes through repeat length variation. The synonymous substitutions within the STRs coding poly-Q to prevent hyper-expansion can also diversify repeat length variation of poly-Q stretches without increasing the risk of long CAG repeats. An increase in the diversity of genes with poly-Q stretches in humans may diversify the consequence of PQBP1 binding to poly-Q stretches. Accordingly, diversification in AS regulation by PQBP1 increases the diversity of transcript isoforms in the human brain.

The human brain displays an extreme diversity of transcript isoforms brought by AS (Zaghlool et al. 2014). This indicates that AS regulation in brain development is an essential element for human evolution. The most significant feature of human evolution is highly structured society, which is realized by accelerated brain evolution, and requires transcript diversification in brain. Because the genomic difference between humans and chimpanzees is about 1 %, the vast transcriptome diversity is largely attributed to the evolution of alternative splicing (Molnár et al. 2014; Zaghlool et al. 2014). Ultimately, this is considered to lead to the formation of large and organized human society and has enabled survival under various environments of human beings.

The comparative study on the disease causing repeats among humans and apes emphasized a significantly wide range of length variation rather than the slightly longer average of human alleles (Andrés et al. 2004). This is now well-understood by our study. The diversity of AS-regulating factors may result in diversification in brain development, which results in the diversification of sensory, motor and cognitive abilities in the human population. The vast diversity of poly-Q repeats in genes involving

neurodevelopmental regulation pose a risk of bearing long alleles which are enough to result in hyper-expansion. Various toxicities of RNAs and proteins caused by hyper-expansion represent cell-type-specific differences in their impact (Shiraishi et al. 2014; Shiwaku and Okazawa 2015). Our hypothesis will be supported if AS regulations by poly-Q binding molecules such as PQBP1 are observed in cells which are vulnerable against hyper-expansion in the developmental stage.

In this paper, we offered insight into the evolutionary principles in length polymorphism of STRs in the human genome. We showed that the most enigmatic STRs, encoding poly-Q stretches, can be understood by distinguishing length polymorphic poly-Q repeats from monomorphic ones. We offered the verifiable hypothesis of a molecular relationship between poly-Q and neurodegenerative disease (Fig. S6). This study may indicate a new clue for the evolutionary switch hypothesis of STRs/SARs in humans that maintains brain function diversity in human populations. Further characterization of the relationship between the diversity of poly-Q repeats, AS regulation in neurodevelopment, and human-characterized evolution will be expected.

Methods

Distribution of simple tandem repeats (STRs) in genome using H-GOLD/GDBS

We analyzed simple tandem repeat (STR) data obtained from the H-GOLD/GDBS project (Tamiya et al. 2005) (http://dbarchive.biosciencedbc.jp/jp/gdbs/download.html) with respect to genomic position in human reference sequences and diversity in 125 Japanese individuals. We eliminated non-informative data (i.e., no heterozygosity value, unclear description in allele number and IDs) from the STR diversity information. We also eliminated marker regions (regions inspected by a pooled PCR procedure) containing more than one STR regions from the STR diversity information.

Analysis of diversity properties of STRs between genomic regions using H-GOLD/GDBS

We calculated the correlation coefficients among a number of repeating units, number of alleles, and heterozygosity, then compared them among H-GOLD/GDBS STRs in the following four regions; the whole genomic region, exonic regions, coding sequence (CDS) regions, and untranslated regions (UTRs) (Fig. S1b). Because H-GOLD/GDBS contained a small number of di-nucleotide STRs in the CDS and trinucleotide STRs in the UTR, correlations were not calculated in these two categories.

Transcript database search for STRs and SARs within exonic region

In addition to the foregoing genome-wide analysis, we separately extracted whole exonic STRs in human transcript sequences (All Human Genes 2, AHG2) in three databases including the H-Invitational Database release 3.8 (H-InvDB 3.8) (Yamasaki et al. 2008), Ensembl database (Flicek et al. 2012), and RefSeq database (Pruitt et al. 2002). The total number of human transcripts was 224,613. We mapped these human transcripts on to the human genome reference sequences (NCBI build 35.1, hg17) and identified 37,616 human gene clusters (H-InvDB gene clusters, HIXs). A representative transcript sequence was uniquely selected from transcripts constituting each cluster (representative H-InvDB transcript, representative HITs). We distinguished transcripts with protein evidence from others (validated HITs), using criteria corresponding to the Category I–III of H-InvDB (Imanishi et al. 2004; Yamasaki et al. 2008) (http://h-invitational.jp/hinv/help/contents2/proteins.html). Briefly, predicted amino acid sequences are identical (Category I) or similar (Category II) to sequences of known proteins, or they contain an InterPro domain (Category III). Using in-house programs, we identified di-, tri-, tetra-, and penta-nucleotide STRs in the AHG2 with five or longer numbers of repeating units. We translated all of these AHG2 transcript sequences into amino acid sequences, and then single amino acid repeats (SARs) containing five or longer repeats were identified in the same way. We filtered STRs and SARs to obtain reliable and non-redundant data sets of STRs/SARs located on representative HITs with protein evidence as aforementioned.

Prediction of variable STRs

We detected length polymorphism of STRs and SARs in the AHG2 database by combining the following three methods; (1) detecting repeat-length differences in alignments of published transcript sequences, (2) matching to a polymorphic STR marker database (H-GOLD/GDBS), and (3) matching to position and sequence information of deletion/insertion polymorphism (DIP) in a public database, dbSNP (build 125).

1. Detecting repeat-length differences in alignments of published transcript sequences

   H-InvDB is an annotated human gene database based on all transcript sequences submitted to the public database for nucleic acid sequence (i.e., DDBJ/EMBL/GenBank, The International Nucleotide Sequence Databases Collaboration, INSDC). Therefore, H-InvDB includes alignment data with multiple

transcript sequences if multiple transcript sequences are available for the gene in the INSDC transcriptome. To annotate STR length polymorphism, we inspected the alignments of these transcripts whether numbers of repeating units of STRs were variable or not. The results of this method are available at a satellite database of H-InvDB, VarySysDB (Shimada et al. 2009).

2. Matching to a polymorphic STR marker database (H-GOLD/GDBS)
   Because H-GOLD/GDBS contains validating polymorphic STRs, we assumed that STRs in H-Inv transcripts were polymorphic when they were also found in H-GOLD/GDBS.

3. Matching to deletion/insertion polymorphism (DIP) in a public database (dbSNP)

   The dbSNP database is a central repository for both SNPs and DIPs by NCBI in collaboration with the National Human Genome Research Institute. We collected DIPs from the dbSNP and incorporated them into the AHG2 database according to their genome position. We examined the allele description of these DIPs in the dbSNP. If these DIPs were consistent with the AHG2 in position and allele, we identified the sites as variable AHG2 STRs.

According to method 1 that used in polymorphic STR detection, we also identified polymorphic SARs using alignment of amino acid sequences deduced from AHG2 transcript sequences. The coordinates determined by H-InvDB were used to select STRs located in CDS regions.

Statistical analysis of number of repeating units

The distributions of the numbers of repeating units of STRs and SARs found in human exonic regions were counted non-redundantly in genome positions, and were depicted in boxplots using the R-package (R Core Team 2012). We extracted STRs with five or longer repeating units. However, the values of the number of repeating units decreased by one repeat (i.e., four or longer), when we counted the number of repeating codons if the codon-frame did not start from the first nucleotide. The distributions of the poly-Q and poly-P repeats were compared with those of other repeats, using the two–tailed Mann–Whitney-Wilcoxon test with the 'wilcox.exact' in the R package.

SNP density

To estimate the density of synonymous and nonsynonymous SNPs in the repeat regions, we estimated the number of potential sites of synonymous (Syn), nonsynonymous (Nsy) and nonsense (Ter) mutations, respectively, by dividing potential nucleotide sites of each amino acid contained in the repeat regions of each representative HIT (mRNA) sequence data, using the method of Nei and Gojobori (1986) and taking into account the transition/transversion rate ratio (Suzuki 2011) that has been estimated to be 4 in mammals (Jiang and Zhao 2006; Rosenberg et al. 2003; Zhang et al. 2007).

GO analyses

(GO-I) Distribution of InterPro annotation for domain consisting poly-P or poly-Q repeats

We searched GO terms linked to protein domains (or motifs) of poly-P or poly-Q repeats in the InterPro database by keywords, such as "proline", "glutamine", "proline rich", or "glutamine rich" via the InterPro web page (Mulder et al. 2002). Among the search results, we selected entries containing a description of proline-rich or glutamine-rich regions in the section of "GO Term annotation" and "InterPro annotation" in their web pages. Then, we extracted these GO terms and identifiers of InterPro, and classified and summarized them according to the categories that were uniquely defined using in-house programs as a part of the functional annotation of H-InvDB.

(GO-II) Distribution of InterPro annotation for genes containing poly-P or poly-Q repeats

In H-InvDB annotations, all of HITs have established links to InterPro IDs by performing InterProScan (Quevillon et al. 2005). Using this annotation set, we extracted the molecular function in GO terms of representative HITs containing STRs or SARs. Then, we classified these GO terms according to GO categories as mentioned in the GO-I. We conducted this classification process for four gene classes, including genes containing polymorphic STRs, monomorphic STRs, polymorphic SARs, and monomorphic SARs.

(GO-III) Gene set enrichment analysis for genes containing poly-P or poly-Q repeats

We divided the poly-Q stretches into 48 polymorphic and 230 monomorphic for the state of their length polymorphism. Then, 41 genes containing at least one polymorphic poly-Q stretch (polymorphic gene set) and 157 genes containing monomorphic poly-Qs only (monomorphic gene set, Fig. 6b) were subjected to gene set enrichment analyses using the DAVID database (Huang

et al. 2009a, b). Because polymorphic poly-P repeats were observed in only 19 counts, we did not divide poly-P containing genes and subjected them to the same analyses as a gene set. First, we analyzed gene-enrichment (proportion of genes falling under a certain annotation category) in each of the selected gene sets compared to that in all human genes, which was evaluated using *p* value of modified Fisher's exact test. Then, we compared the results between polymorphic and monomorphic gene sets (Table 2). In the DAVID database, the total number of human genes with the "GO Molecular Function" was 12,983, which included 32 out of 41 polymorphic and 126 out of 157 monomorphic genes, respectively. The result for the "GO Biological Process" was 13,528, which included 34 out of 41 polymorphic and 120 out of 157 monomorphic genes, respectively. Since these gene enrichment analyses are multiple comparison, the obtained *p* values were corrected using Bonferroni (Armstrong 2014; Dunn 1961) and Benjamini-Hochberg (Benjamini and Hochberg 1995) methods (Tables S1–S4).

Analysis for juxtaposing SARs

We counted the number of cases in which two different SARs were adjacent. The probability distribution of the number of cases observed in the total cases was obtained based on binominal distribution for each SAR combination, by considering the observed SAR frequencies.

Acknowledgments We are grateful to Hidetoshi Inoko for support to use H-GOLD/GDBS data, Yasuyuki Fujii, Katsuhiko Murakami, Yoshiharu Sato and Jun-ichi Takeda for providing gene structure and annotation data, Ryuzo Matsumoto and Yosuke Hayakawa for useful suggestion on computer programming, and other former member of the H-Invitational 2 consortium, Genome Information Integration Project (GIIP), the Integrated Database and Systems Biology Team of BIRC, AIST for their helpful support. This research was financially supported by the Ministry of Economy, Trade and Industry of Japan (METI) and the Japan Biological Informatics Consortium (JBIC). Also, this work is partly supported by the Grants-in-Aid for Scientific Research (C) to MKS (JSPS Grant Numbers 24510271 and 21510205), and the Saito Gratitude Foundation to MKS.

Compliance with ethical standards

Conflict of interest All authors declare no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent This article does not contain any studies with human participants.

Data availability Updated data on STRs and SARs within known human transcriptome sequences will be continuously provided in the VarySysDB database (http://h-invitational.jp/varygene/home.htm). The original data of STRs and SARs in human exonic region are available at the web site of the first author, MKS (http://www.fujita-hu.ac.jp/~mshimada/sub/STR_SAR_Data.html) and a web-based data sharing system provided by the research map (http://researchmap.jp/shimada-mk/%E8%B3%87%E6%96%99%E5%85%AC%E9%96%8B/).

References

Alba MM, Guigo R (2004) Comparative analysis of amino acid repeats in rodents and humans. Genome Res 14:549–554. doi:10.1101/gr.1925704

Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, Chiaromonte F, Makova KD (2013) Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. Genome Biol Evol 5:606–620. doi:10.1093/gbe/evs116

Andrés AM, Soldevila M, Lao O, Volpini V, Saitou N, Jacobs HT, Hayasaka I, Calafell F, Bertranpetit J (2004) Comparative genetics of functional trinucleotide tandem repeats in humans and apes. J Mol Evol 59:329–339. doi:10.1007/s00239-004-2628-5

Armstrong RA (2014) When to use the Bonferroni correction. Ophthalmic Physiol Opt 34:502–508. doi:10.1111/opo.12131

Astolfi P, Bellizzi D, Sgaramella V (2003) Frequency and coverage of trinucleotide repeats in eukaryotes. Gene 317:117–125. doi:10.1016/S0378-1119(03)00659-0

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc 57:289–300

Bhattacharyya A, Thakur AK, Chellgren VM, Thiagarajan G, Williams AD, Chellgren BW, Creamer TP, Wetzel R (2006) Oligoproline effects on polyglutamine conformation and aggregation. J Mol Biol 355:524–535. doi:10.1016/j.jmb.2005.10.053

Birge L, Pitts M, Baker R, Wilkinson G (2010) Length polymorphism and head shape association among genes with polyglutamine repeats in the stalk-eyed fly, *Teleopsis dalmanni*. BMC Evol Biol 10:227. doi:10.1186/1471-2148-10-227

Burke KA, Kauffman KJ, Umbaugh CS, Frey SL, Legleiter J (2013) The interaction of polyglutamine peptides with lipid membranes is regulated by flanking sequences associated with huntingtin. J Biol Chem 288:14993–15005. doi:10.1074/jbc.M112.446237

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 28:1040–1050. doi:10.1002/bies.20470

Chan HYE (2014) RNA-mediated pathogenic mechanisms in polyglutamine diseases and amyotrophic lateral sclerosis. Front Cell Neurosci 8:431. doi:10.3389/fncel.2014.00431

Chang DK, Metzgar D, Wills C, Boland CR (2001) Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. Genome Res 11:1145–1146. doi:10.1101/gr.186301

Choudhry S, Mukerji M, Srivastava AK, Jain S, Brahmachari SK (2001) CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. Hum Mol Genet 10:2437–2446. doi:10.1093/hmg/10.21.2437

Core Team R (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Cotton AJ, Foldvari M, Cotton S, Pomiankowski A (2014) Male eyespan size is associated with meiotic drive in wild stalk-eyed flies (*Teleopsis dalmanni*). Heredity 112:363–369

Darnell G, Orgel JP, Pahl R, Meredith SC (2007) Flanking polyproline sequences inhibit beta-sheet structure in polyglutamine segments by inducing PPII-like helix structure. J Mol Biol 374:688–704. doi:10.1016/j.jmb.2007.09.023

Deka R, Guangyun S, Smelser D, Zhong Y, Kimmel M, Chakraborty R (1999a) Rate and directionality of mutations and effects of allele size constraints at anonymous, gene-associated, and disease-causing trinucleotide loci. Mol Biol Evol 16:1166–1177

Deka R, Guangyun S, Wiest J, Smelser D, Chunhua S, Zhong Y, Chakraborty R (1999b) Patterns of instability of expanded CAG repeats at the ERDA1 locus in general populations. Am J Hum Genet 65:192–198. doi:10.1086/302453

Dunbar RI (1998) The social brain hypothesis. Brain 9:178–190

Dunn OJ (1961) Multiple comparisons among means. J Am Stat Assoc 56:52–64. doi:10.1080/01621459.1961.10482090

Elhaik E, Landan G, Graur D, Can GC (2009) Content at third-codon positions be used as a proxy for isochore composition? Mol Biol Evol 26:1829–1833. doi:10.1093/molbev/msp100

Erwin AL, Bonthuis PJ, Geelhood JL, Nelson KL, McCrea KW, Gilsdorf JR, Smith AL (2006) Heterogeneity in tandem octanucleotides within *Haemophilus influenzae* lipopolysaccharide biosynthetic gene *losA* affects serum resistance. Infect Immun 74:3408–3414. doi:10.1128/IAI.01540-05

Faux N (2012) Single amino acid and trinucleotide repeats: function and evolution. In: Hannan AJ (ed) Tandem repeat polymorphisms: genetic plasticity, neural diversity and disease, vol 769. Landes Bioscience and Springer Science + Business Media, New York, pp 26–40

Fiszer A, Krzyzosiak W (2013) RNA toxicity in polyglutamine disorders: concepts, models, and progress of research. J Mol Med 91:683–691. doi:10.1007/s00109-013-1016-2

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S et al (2012) Ensembl 2012. Nucleic Acids Res 40:D84–D90. doi:10.1093/nar/gkr991

Fondon JW III, Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci USA 101:18058–18063. doi:10.1073/pnas.0408118101

Fondon JW III, Hammock EAD, Hannan AJ, King DG (2008) Simple sequence repeats: genetic modulators of brain function and behavior. Trends Neurosci 31:328–334. doi:10.1016/j.tins.2008.03.006

Fukuda K, Ichiyanagi K, Yamada Y, Go Y, Udono T, Wada S, Maeda T, Soejima H, Saitou N, Ito T et al (2013) Regional DNA methylation differences between humans and chimpanzees are associated with genetic changes, transcriptional divergence and disease genes. J Hum Genet 58:446–454. doi:10.1038/jhg.2013.55

Galant R, Carroll SB (2002) Evolution of a transcriptional repression domain in an insect Hox protein. Nature 415:910–913. doi:10.1038/nature717

Gebhardt F, Zanker KS, Brandt B (1999) Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. J Biol Chem 274:13176–13180. doi:10.1074/jbc.274.19.13176

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44:445–477. doi:10.1146/annurev-genet-072610-155046

Gojobori J, Ueda S (2011) Elevated evolutionary rate in genes with homopolymeric amino acid repeats constituting nondisordered structure. Mol Biol Evol 28:543–550. doi:10.1093/molbev/msq225

Grabczyk E, Mancuso M, Sammarco MC (2007) A persistent RNA. DNA hybrid formed by transcription of the Friedreich ataxia triplet repeat in live bacteria, and by T7 RNAP in vitro. Nucleic Acids Res 35:5351–5359. doi:10.1093/nar/gkm589

Guo W-J, Ling J, Li P (2009) Consensus features of microsatellite distribution: microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes. Genomics 93:323–331. doi:10.1016/j.ygeno.2008.12.009

Guzhova IV, Lazarev VF, Kaznacheeva AV, Ippolitova MV, Muronetz VI, Kinev AV, Margulis BA (2011) Novel mechanism of Hsp70 chaperone-mediated prevention of polyglutamine aggregates in a cellular model of Huntington disease. Hum Mol Genet 20:3953–3963. doi:10.1093/hmg/ddr314

Haasl RJ, Payseur BA (2013) Microsatellites as targets of natural selection. Mol Biol Evol 30:285–298. doi:10.1093/molbev/mss247

Hammock EAD, Young LJ (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308:1630–1634. doi:10.1126/science.1111427

Huang DW, Sherman BT, Lempicki RA (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37:1–13. doi:10.1093/nar/gkn923

Huang DW, Sherman BT, Lempicki RA (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols 4:44–57. doi:10.1038/nprot.2008.211

Hui J, Hung LH, Heiner M, Schreiner S, Neumüller N, Reither G, Haas SA, Bindereif A (2005) Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J 24:1988–1998. doi:10.1038/sj.emboj.7600677

Huntley MA, Clark AG (2007) Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. Mol Biol Evol 24:2598–2609. doi:10.1093/molbev/msm129

Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M et al (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2:e162. doi:10.1371/journal.pbio.0020256

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. doi:10.1038/35057062

Jiang C, Zhao Z (2006) Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. Genomics 88:527–534. doi:10.1016/j.ygeno.2006.06.003

Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22:253–259. doi:10.1016/j.tig.2006.03.005

King DG (2012) Evolution of simple sequence repeats as mutable sites. In: Hannan AJ (ed) Tandem repeat polymorphisms: genetic plasticity, neural diversity and disease, vol 769. Landes Bioscience and Springer Science + Business Media, New York, pp 10–25

Kozlowski P, de Mezer M, Krzyzosiak WJ (2010) Trinucleotide repeats in human genome and exome. Nucleic Acids Res 38:4027–4039. doi:10.1093/nar/gkq127

Kurosaki T, Ninokata A, Wang L, Ueda S (2006) Evolutionary scenario for acquisition of CAG repeats in human SCA1 gene. Gene 373:23–27. doi:10.1016/j.gene.2005.12.017

Labbadia J, Morimoto RI (2013) Huntington's disease: underlying molecular mechanisms and emerging concepts. Trends Biochem Sci 38:378–385. doi:10.1016/j.tibs.2013.05.003

Laffita-Mesa JM, Velazquez-Perez LC, Santos Falcon N, Cruz-Marino T, Gonzalez Zaldivar Y, Vazquez Mojena Y, Almaguer-Gotay D, Almaguer Mederos LE, Rodriguez Labrada R (2012) Unexpanded and intermediate CAG polymorphisms at the SCA2 locus (ATXN2) in the Cuban population: evidence about the origin of expanded SCA2 alleles. Eur J Hum Genet 20:41–49. doi:10.1038/ejhg.2011.154

Legendre M, Pochet N, Pak T, Verstrepen KJ (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res 17:1787–1796. doi:10.1101/gr.6554007

Lin Y, Wilson JH (2012) Nucleotide excision repair, mismatch repair, and R-loops modulate convergent transcription-induced cell

death and repeat instability. PLoS One 7:e46807. doi:10.1371/journal.pone.0046807
Lin Y, Leng M, Wan M, Wilson JH (2010) Convergent transcription through a long CAG tract destabilizes repeats and induces apoptosis. Mol Cell Biol 30:4435–4451. doi:10.1128/mcb.00332-10
Lin W, Lin Y, Wilson J (2014) Convergent transcription through microsatellite repeat tracts induces cell death. Mol Biol Rep 41:5627–5634. doi:10.1007/s11033-014-3432-y
Liu C-R, Chang C-R, Chern Y, Wang T-H, Hsieh W-C, Shen W-C, Chang C-Y, Chu IC, Deng N, Cohen SN et al (2012) Spt4 is selectively required for transcription of extended trinucleotide repeats. Cell 148:690–701. doi:10.1016/j.cell.2011.12.032
Lo Sardo V, Zuccato C, Gaudenzi G, Vitali B, Ramos C, Tartari M, Myre MA, Walker JA, Pistocchi A, Conti L et al (2012) An evolutionary recent neuroepithelial cell adhesion function of huntingtin implicates ADAM10-Ncadherin. Nat Neurosci 15:713–721. doi:10.1038/nn.3080
Lorenz M, Hewing B, Hui J, Zepp A, Baumann G, Bindereif A, Stangl V, Stangl K (2007) Alternative splicing in intron 13 of the human eNOS gene: a potential mechanism for regulating eNOS activity. FASEB J 21:1556–1564. doi:10.1096/fj.06-7434com
McIvor EI, Polak U, Napierala M (2010) New insights into repeat instability: role of RNA-DNA hybrids. RNA Biol 7:551–558. doi:10.4161/rna.7.5.12745
Mishra R, Jayaraman M, Roland BP, Landrum E, Fullam T, Kodali R, Thakur AK, Arduini I, Wetzel R (2012) Inhibiting the nucleation of amyloid structure in a huntingtin fragment by targeting α-helix-rich oligomeric intermediates. J Mol Biol 415:900–917. doi:10.1016/j.jmb.2011.12.011
Mizuguchi M, Obita T, Serita T, Kojima R, Nabeshima Y, Okazawa H (2014) Mutations in the PQBP1 gene prevent its interaction with the spliceosomal protein U5-15kD. Nat Commun 5:3822. doi:10.1038/ncomms4822
Mohan A, Goodwin M, Swanson MS (2014) RNA-protein interactions in unstable microsatellite diseases. Brain Res 1584:3–14. doi:10.1016/j.brainres.2014.03.039
Molnár Z, Kaas JH, de Carlos JA, Hevner RF, Lein E, Němec P (2014) Evolution and development of the mammalian cerebral cortex. Brain Behav Evol 83:126–139. doi:10.1159/000357753
Mühlau M, Winkelmann J, Rujescu D, Giegling I, Koutsouleris N, Gaser C, Arsic M, Weindl A, Reiser M, Meisenzahl EM (2012) Variation within the Huntington's disease gene influences normal brain structure. PLoS One 7:e29809. doi:10.1371/journal.pone.0029809
Mularoni L, Veitia RA, Alba MM (2007) Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. Genomics 89:316–325. doi:10.1016/j.ygeno.2006.11.011
Mularoni L, Ledda A, Toll-Riera M, Albà MM (2010) Natural selection drives the accumulation of amino acid tandem repeats in human proteins. Genome Res 20:745–754. doi:10.1101/gr.101261.109
Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P et al (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. Brief Bioinform 3:225–235. doi:10.1093/bib/3.3.225
Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426
Nithianantharajah J, Hannan AJ (2007) Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. Bioessays 29:525–535. doi:10.1002/bies.20589
Ogasawara M, Imanishi T, Moriwaki K, Gaudieri S, Tsuda H, Hashimoto H, Shiroishi T, Gojobori T, Koide T (2005) Length variation of CAG/CAA triplet repeats in 50 genes among 16 inbred mouse strains. Gene 349:107–119. doi:10.1016/j.gene.2004.11.050
Okazawa H (2003) Polyglutamine diseases: a transcription disorder? Cell Mol Life Sci 60:1427–1439. doi:10.1007/s00018-003-3013-z
Okazawa H, Rich T, Chang A, Lin X, Waragai M, Kajikawa M, Enokido Y, Komuro A, Kato S, Shibata M et al (2002) Interaction between mutant ataxin-1 and PQBP-1 affects transcription and cell death. Neuron 34:701–713. doi:10.1016/S0896-6273(02)00697-9
Paulson HL (2000) Toward an understanding of polyglutamine neurodegeneration. Brain Pathol 10:293–299. doi:10.1111/j.1750-3639.2000.tb00263.x
Perutz MF, Johnson T, Suzuki M, Finch JT (1994) Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. Proc Natl Acad Sci USA 91:5355–5358
Pruitt K, Brown G, Tatusova T, Maglott D (2002) The reference sequence (RefSeq) database. The NCBI handbook. National Center for Biotechnology Information, U.S. National Library of Medicine. http://www.ncbi.nlm.nih.gov/books/NBK21091/. Accessed 30 Jun 2015
Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120. doi:10.1093/nar/gki442
Rado-Trilla N, Alba M (2012) Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evol Biol 12:155. doi:10.1186/1471-2148-12-155
Ramazzotti M, Monsellier E, Kamoun C, Degl'Innocenti D, Melki R (2012) Polyglutamine repeats are associated to specific sequence biases that are conserved among eukaryotes. PLoS One 7:e30824. doi:10.1371/journal.pone.0030824
Rees M, Gorba C, de Chiara C, Bui TT, Garcia-Maya M, Drake AF, Okazawa H, Pastore A, Svergun D, Chen YW (2012) Solution model of the intrinsically disordered polyglutamine tract-binding protein-1. Biophys J 102:1608–1616. doi:10.1016/j.bpj.2012.02.047
Richard G-F, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72:686–727. doi:10.1128/mmbr.00011-08
Rosenberg MS, Subramanian S, Kumar S (2003) Patterns of transitional mutation biases within and among mammalian genomes. Mol Biol Evol 20:988–993. doi:10.1093/molbev/msg113
Salinas-Rios V, Belotserkovskii BP, Hanawalt PC (2011) DNA slip-outs cause RNA polymerase II arrest in vitro: potential implications for genetic instability. Nucleic Acids Res 39:7444–7454. doi:10.1093/nar/gkr429
Shao J, Diamond MI (2007) Polyglutamine diseases: emerging concepts in pathogenesis and therapy. Hum Mol Genet 16:R115–R123. doi:10.1093/hmg/ddm213
Shimada MK, Matsumoto R, Hayakawa Y, Sanbonmatsu R, Gough C, Yamaguchi-Kabata Y, Yamasaki C, Imanishi T, Gojobori T (2009) VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. Nucleic Acids Res 37:D810–D815. doi:10.1093/nar/gkn798
Shiraishi R, Tamura T, Sone M, Okazawa H (2014) Systematic analysis of fly models with multiple drivers reveals different effects of Ataxin-1 and huntingtin in neuron subtype-specific expression. PLoS One 9:e116567. doi:10.1371/journal.pone.0116567
Shiwaku H, Okazawa H (2015) Impaired DNA damage repair as a common feature of neurodegenerative diseases and psychiatric disorders. Curr Mol Med 15:119–128. doi:10.2174/1566524015666150303002556
Shriver MD, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134:983–993

Siwach P, Pophaly SD, Ganesh S (2006) Genomic and evolutionary insights into genes encoding proteins with single amino acid repeats. Mol Biol Evol 23:1357–1369. doi:10.1093/molbev/msk022

Siwach P, Sengupta S, Parihar R, Ganesh S (2011) Proline repeats, in *cis*- and *trans*-positions, confer protection against the toxicity of misfolded proteins in a mammalian cellular model. Neurosci Res 70:435–441. doi:10.1016/j.neures.2011.05.001

Sobczak K, Krzyzosiak WJ (2004) Patterns of CAG repeat interruptions in SCA1 and SCA2 genes in relation to repeat instability. Hum Mutat 24:236–247. doi:10.1002/humu.20075

Sobczak K, Michlewski G, de Mezer M, Kierzek E, Krol J, Olejniczak M, Kierzek R, Krzyzosiak WJ (2010) Structural diversity of triplet repeat RNAs. J Biol Chem 285:12755–12764. doi:10.1074/jbc.M109.078790

Suzuki Y (2011) Overestimation of nonsynonymous/synonymous rate ratio by reverse-translation of aligned amino acid sequences. Genes Genet Syst 86:123–129

Takahashi M, Mizuguchi M, Shinoda H, Aizawa T, Demura M, Okazawa H, Kawano K (2009) Polyglutamine tract binding protein-1 is an intrinsically unstructured protein. Biochim Biophys Acta Proteins Proteom 1794:936–943. doi:10.1016/j.bbapap.2009.03.001

Takahashi T, Katada S, Onodera O (2010) Polyglutamine diseases: where does toxicity come from? what is toxicity? where are we going? J Mol Cell Biol 2:180–191. doi:10.1093/jmcb/mjq005

Takezaki N, Nei M (2009) Genomic drift and evolution of microsatellite DNAs in human populations. Mol Biol Evol 26:1835–1840. doi:10.1093/molbev/msp091

Tamiya G, Shinya M, Imanishi T, Ikuta T, Makino S, Okamoto K, Furugaki K, Matsumoto T, Mano S, Ando S et al (2005) Whole genome association study of rheumatoid arthritis using 27039 microsatellites. Hum Mol Genet 14:2305–2321. doi:10.1093/hmg/ddi234

Tartari M, Gissi C, Lo Sardo V, Zuccato C, Picardi E, Pesole G, Cattaneo E (2008) Phylogenetic comparison of huntingtin homologues reveals the appearance of a primitive polyQ in sea urchin. Mol Biol Evol 25:330–338. doi:10.1093/molbev/msm258

Tatarinova T, Elhaik E, Pellegrini M (2013) Cross-species analysis of genic GC3 content and DNA methylation patterns. Genome Biol Evol 5:1443–1456. doi:10.1093/gbe/evt103

The UniProt Consortium (2014) Activities at the universal protein resource (UniProt). Nucleic Acids Res 42:D191–D198. doi:10.1093/nar/gkt1140

Tompa P (2003) Intrinsically unstructured proteins evolve by repeat expansion. Bioessays 25:847–855. doi:10.1002/bies.10324

Trifonov EN (1989) The multiple codes of nucleotide sequences. Bull Math Biol 51:417–432

Tsoi H, Chan HYE (2014) Roles of the nucleolus in the CAG RNA-mediated toxicity. Biochim Biophys Acta Mol Basis Dis 1842:779–784. doi:10.1016/j.bbadis.2013.11.015

Vachharajani SN, Chaudhary RK, Prasad S, Roy I (2012) Length of polyglutamine tract affects secondary and tertiary structures of huntingtin protein. Int J Biol Macromol 51:920–925. doi:10.1016/j.ijbiomac.2012.07.022

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ (2009) Unstable tandem repeats in promoters confer transcriptional evolvability. Science 324:1213–1216. doi:10.1126/science.1170097

Wang Q, Moore MJ, Adelmant G, Marto JA, Silver PA (2013) PQBP1, a factor linked to intellectual disability, affects alternative splicing associated with neurite outgrowth. Genes Dev 27:615–626. doi:10.1101/gad.212308.112

Waragai M, Lammers C-H, Takeuchi S, Imafuku I, Udagawa Y, Kanazawa I, Kawabata M, Mouradian MM, Okazawa H (1999) PQBP-1, a novel polyglutamine tract-binding protein, inhibits transcription activation by Brn-2 and affects cell survival. Hum Mol Genet 8:977–987. doi:10.1093/hmg/8.6.977

Weiser JN, Love JM, Moxon ER (1989) The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. Cell 59:657–665. doi:10.1016/0092-8674(89)90011-1

Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, J-i Takeda, Taniya T, Sakate R, Kikugawa S, Shimada M et al (2008) The H-invitational database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Nucleic Acids Res 36:D793–D799. doi:10.1093/nar/gkm999

Yin R-H, Li Y, Yang F, Zhan Y-Q, Yu M, Ge C-H, Xu W-X, Tang L-J, Wang X-H, Chen B et al (2014) Expansion of the polyQ repeats in THAP11 forms intranuclear aggregation and causes cell G0/G1 arrest. Cell Biol Int 38:757–767. doi:10.1002/cbin.10255

Zaghlool A, Ameur A, Cavelier L, Feuk L (2014) Splicing in the Human Brain. In: Robert H, Shannon M (eds) International review of neurobiology, vol 116., Academic PressWaltham, MA, pp 95–125

Zhang W, Bouffard GG, Wallace SS, Bond JP (2007) Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. J Mol Evol 65:207–214. doi:10.1007/s00239-007-9000-5

Zhang W, Zeng F, Liu Y, Zhao Y, Lv H, Niu L, Teng M, Li X (2013) Crystal structures and RNA-binding properties of the RNA recognition motifs of heterogeneous nuclear ribonucleoprotein L: insights into its roles in alternative-splicing regulation. J Biol Chem 288:22636–22649. doi:10.1074/jbc.M113.463901