

Comparative Genetics of Trinucleotide Repeats in the Human and Ape Genomes

Loris Mularoni, *Fundació Institut Municipal d'Investigació Mèdica, Barcelona, Spain*

Macarena Toll-Riera, *Fundació Institut Municipal d'Investigació Mèdica, Barcelona, Spain*

M Mar Albà, *Fundació Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Barcelona, Spain*

The genome of humans and other apes shows important differences in trinucleotide repeat sequences. This has implications for the evolution of protein function and susceptibility to repeat expansion disease.

Introduction

Eukaryotic genomes are rich in repetitive sequences, including retroposons, minisatellites and microsatellites (Lander *et al.*, 2001; Tautz and Renz, 1984). Microsatellites are tandem iterations of 1–6 bp units and account for about 3% of the human genomic sequence (Lander *et al.*, 2001). When repeat unit is of size 3, they are known as trinucleotide repeats. The term simple sequence repeat (SSR) refers to both tandem and nontandem, interrupted, repetitions. Microsatellites show high mutation rates, which have been estimated to be of several orders of magnitude than those of nonrepetitive sequences (Ellegren, 2000). The high repeat number variability typically exhibited by microsatellites has resulted in a wide use of these sequences in genetic mapping, forensic testing and population studies. Expansion or contraction of microsatellites is assumed to occur primarily by replication slippage, a process by which misalignment of repeats between deoxyribonucleic acid (DNA) strands results in the gain or loss of one or a few repeat units in the nascent strand (Levinson and Gutman, 1987). Long uninterrupted tandem repeat tracts generally show a greater tendency to expansion, as well as higher polymorphism levels, than those containing interruptions (Alba *et al.*, 1999; Mularoni *et al.*, 2006;

O'Dushlaine *et al.*, 2005; Primmer and Ellegren, 1998; Wren *et al.*, 2000). **See also:** [Microsatellite Instability](#)

Trinucleotide repeats have attracted much interest since the early 1990s, when a number of neurological diseases associated with repeat expansion mutations were discovered (Brown and Brown, 2004; Gatchel and Zoghbi, 2005). Among microsatellites, trinucleotide repeats are the most common microsatellite type in protein-coding sequences, while mononucleotide and dinucleotide repeats predominate in noncoding sequences (Toth *et al.*, 2000). This can be explained by the fact that expansion of trinucleotide repeats within coding sequences does not disrupt the reading frame, and the resulting stretches of repetitive amino acids – also called homopeptides – can often be accommodated in functional proteins. The analysis of trinucleotide repeat distributions in primate coding sequences has shown that a subset of trinucleotide motifs, including CGG, CAG and GAA, are particularly prone to suffer expansions (Borstnik and Pumpernik, 2002). **See also:** [Trinucleotide Repeat Expansions: Disorders](#); [Trinucleotide Repeat Expansions: Mechanisms and Disease Associations](#)

The comparative analysis of microsatellite length in orthologous loci from related species is of key importance to understand the evolutionary dynamics of these sequences and to detect relevant interspecific differences. Early comparisons on different repeat loci, based on the amplification of preselected human microsatellites in chimpanzee, pointed to a significant trend for the repeats to be longer in humans than in other apes (Rubinsztein *et al.*, 1995a). This finding raised the possibility that there was directionality in the rate of microsatellite evolution, with a specific tendency towards expansion in the human lineage (Rubinsztein *et al.*, 1995b). However, studies based on more recent data have indicated that this trend shows important variations depending on the loci and size of the repeat unit considered (Webster *et al.*, 2002), or that the trend is, overall, smaller than previously

Advanced article

Article Contents

- Introduction
- Microsatellites in Human and Chimpanzee
- Lineage-specific Amino Acid Repeat Size Changes
- Interspecific Differences in Disease-associated Amino Acid Repeats
- Glutamine Repeat Polymorphism and Disease
- Summary

Online posting date: 30th April 2008

ELS subject area: Evolution and Diversity of Life

How to cite:

Mularoni, Loris; Toll-Riera, Macarena; and, Albà, M Mar (April 2008) Comparative Genetics of Trinucleotide Repeats in the Human and Ape Genomes. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.

DOI: 10.1002/9780470015902.a0020844

thought (Rubinsztein *et al.*, 1995b; Vowles and Amos, 2006). The recent availability of the chimpanzee draft genome sequence (Consortium TCSaA, 2005) permits a re-evaluation of these differences and, maybe more importantly, the identification of a large number of potentially adaptive lineage-specific repeat expansions.

Microsatellites in Human and Chimpanzee

Divergence of the human and chimpanzee lineages took place around 5–7 Mya (Glazko and Nei, 2003; Kumar *et al.*, 2005; Leakey *et al.*, 1998). Of note, much of the variation found between the two genomes is due to short deletions and insertions (Britten, 2002), many of which fall into microsatellite sequences (Messer and Arndt, 2007). Interestingly, among microsatellite mutations, there is a predominance of repeat unit expansions over contractions (Messer and Arndt, 2007). The characteristics of microsatellites sequences in primates have been compared to those of other mammals (Alba and Guigo, 2004; Toth *et al.*, 2000). In the compilation by Toth *et al.* (2000), mononucleotide repeats were the most abundant microsatellite type in primate genomic sequences, whereas dinucleotide repeats were the most frequent type in rodents. Among primate trinucleotide repeats, AAC and AAT (and their permutations) were the most common triplets in introns and intergenic regions, whereas AGC and CCG (and their permutations) were the most common triplets in exons. However, examination of different in-frame codon repeat types, of size 5 or longer, in the complete human and rodent genomes showed that CAG (resulting in glutamine repeats) and GAG (resulting in glutamic acid repeats) were the most abundant trinucleotide repeats in both lineages (Alba and Guigo, 2004).

The direct comparison of different species can be best achieved by examining orthologous loci. The first human and chimpanzee comparative studies were based on already-known human microsatellites (Rubinsztein *et al.*, 1995a, 1995b), and therefore suffered from the problem of biased selections of loci (Ellegren *et al.*, 1995; Forbes *et al.*, 1995). The loci under study had been originally selected because of their high polymorphism, and thus high mutation rate, in humans, which was likely to lead to erroneous conclusions about the differences in microsatellite length between the two species. With the availability of longer chimpanzee genomic sequences the ‘ascertainment bias’ problem could be reexamined. Webster *et al.* (2002) performed an analysis of 2467 microsatellite loci derived from alignments of 5.1 Mb human and chimpanzee orthologous genomic sequences. They found a significant tendency for human dinucleotide repeats to be longer in human than in chimpanzee, whereas mononucleotide repeats showed the opposite trend. However, there was an excess of trinucleotide repeats that were longer in human, although in this case the difference was not significant. However, Cooper

et al. (1998) and later Vowles and Amos (2006) tried to quantify the effect of ascertainment bias by using microsatellites selected either from the human genome or from the chimpanzee genome. These studies concluded that, after correcting for ascertainment bias, human microsatellites were still longer than their chimpanzee counterparts. In the specific case of GAA trinucleotide repeats, analysis of human, chimpanzee and gorilla orthologous sequences indicated that large expansions of GAA took place, in a locus- and lineage-specific manner, following the divergence of the great apes (Clark *et al.*, 2006).

Lineage-specific Amino Acid Repeat Size Changes

Trinucleotide repeats are abundant in eukaryotic coding sequences. In the open reading frame, trinucleotide (codon) repeats arranged in tandem, or interrupted by synonymous mutations, are translated as single amino acid repeats. As mentioned earlier, uncontrolled expansion of these tracts can give rise to disease. Disease-associated repeats, however, represent only a minor fraction of the existing amino acid repeats (Alba *et al.*, 2007; Green and Wang, 1994). Indeed, as many as 15–20% of the human proteins have been reported to contain tandem repeats of size 5 amino acids or longer, and among them there is a significant overrepresentation of transcription factors and developmental proteins (Alba and Guigo, 2004; Karlin *et al.*, 2002). An important and controversial question is how many of these repeats contribute to protein function. While many amino acid tandem repeats may be neutral and simply tolerated in proteins, there is a growing number of studies that point to functional or structural roles, mainly related to the modulation of protein–protein interactions (Shimohata *et al.*, 2005) or gene transcriptional activity (Gerber *et al.*, 1994). As these sequences can suffer important size modifications in a relatively short time, they are prime candidates to drive changes in protein networks (Hancock and Simon, 2005) and to contribute to adaptive processes (Caburet *et al.*, 2005; Kashi and King, 2006).

Comparison of amino acid repeat length in orthologous proteins from related species has been previously undertaken in mammals (Alba and Guigo, 2004) and in *Drosophila* (Huntley and Clark, 2007). For the purpose of this review, we compiled a dataset of 19 319 human and chimpanzee 1:1 orthologous complementary DNA (cDNA) and protein sequences from Ensembl versus 44 (Hubbard *et al.*, 2005) and identified 3320 orthologous pairs that contained at least one amino acid tandem repeat, of size 5 or longer, in at least one of the two species. The number of amino acid tandem repeats was very similar in both species: 4914 repeats in human and 4853 repeats in chimpanzee (no significant differences by a χ^2 test). For repeats of size 8 or longer, there were 898 repeats in human and 878 repeats in chimpanzee (again the differences were no significant).

Examination of the types of amino acids forming the repeats showed a very similar distribution in the two species. Therefore, we merged the repeats in both species to obtain a descriptive view of the most common amino acid repeat types in primates (Table 1). On the corresponding coding sequences, we calculated the codon homogeneity (CH) as the fraction of the repeat-coding sequence occupied by the longest pure trinucleotide repeat (Mularoni *et al.*, 2007). Long stretches of pure codon repeats, with high CH values, are likely to be slippage products. The average CH, and the percentage of repeats encoded by completely perfect codon runs, was highest for glutamine repeats (polyglutamine) and lowest for proline repeats (polyproline), similar to a previous study based on human and mouse genes (Mularoni *et al.*, 2007). In particular, we determined that 29.1% of primate glutamine repeats were encoded by pure CAG repeats. Other frequently found repeats in pure codon tracts were AGA and GGC (and some of their permutations).

Some of the sequence differences that can be now observed between human and chimpanzee orthologous proteins may be the result of lineage-specific adaptations (reviewed in Kehrer-Sawatzki and Cooper, 2007). Differences in amino acid repeat size between human and chimpanzee have been detected in several cancer genes, which are characterized by otherwise highly conserved sequences (Puente *et al.*, 2006). We systematically search for differences in the compilation of 5064 human and chimpanzee orthologous repeats. We found 864 repeats that varied repeat number between the two species (17%). It seems possible that some of these differences will have an effect on the function of the protein. For example, *in vitro* studies have shown correlation between the number of proline or glutamine repeats and transcriptional activation (Gerber *et al.*, 1994). Furthermore, variations in amino acid repeat length in a number of developmental proteins have been related to morphological change in different species or animal breeds (Anan *et al.*, 2007; Fondon and Garner, 2004; Galant and Carroll, 2002). We analysed human and chimpanzee repeat size divergence in more detail. Of special interest were sequences that showed very large repeat size differences, and we selected those with conservation of at most 50% of the

perfect repeat tract (i.e. 6–3, 7–3, 8–4, etc.). Interestingly, we found that, among them, there was an excess of repeats that were longer in human than in chimpanzee (96 out of 142, χ^2 test: 2.7×10^{-5}). By amino acid repeat type, significant differences were detected for proline and alanine repeats ($p < 0.01$), with a larger number of repeats longer in human than the opposite case (12 versus 2 and 22 versus 5, respectively). Table 2 shows a list of repeats of size 10 or longer in at least one species and repeat size difference of at least 50%. The list includes several transcription factors, such as a number of zinc finger proteins, and disease-associated proteins (e.g. HOXA13, atrophin 1).

Interspecific Differences in Disease-associated Amino Acid Repeats

Of special interest are trinucleotide repeats whose uncontrolled expansion, often reaching a very large number of repeat units, causes human disease. There are two main types of disease-associated repeats in coding sequences: those that can give rise to abnormally long polyglutamine tracts, associated with neurodegenerative disorders (Gatchel and Zoghbi, 2005 and references therein) and those that can result in abnormally long polyalanine tracts, associated with developmental diseases (Amiel *et al.*, 2004; Brown and Brown, 2004 and references therein). The first class always involves the triplet CAG, and is composed of at least nine genes: huntingtin in Huntington disease (HD), atrophin 1 (ATN1) in dentatorubral-pallidoluysian atrophy (DRPLA) or Haw River syndrome, androgen receptor (AR) in spinal and bulbar muscular atrophy, P/Q-type $\alpha 1A$ calcium channel subunit (CACNA1A) in spinocerebellar ataxia 6, TATA-binding protein (TBP) in spinocerebellar ataxia 17, ataxin 1 (ATXN1) in spinocerebellar ataxia 1, ataxin 2 (ATXN2) in spinocerebellar ataxia 2, ataxin 3 (ATXN3) in spinocerebellar ataxia 3 or Machado–Joseph disease and ataxin 7 in spinocerebellar ataxia 7. These genes are involved in DNA-dependent regulation of transcription and/or neurogenesis processes. The disorders share similar clinical features like selective neuronal death.

Table 1 Amino acid tandem repeats in human and chimpanzee orthologous proteins

AA	<i>N</i>	Avg size	Avg (median) CH	Avg longest codon run	CH = 1 (%)	Codon CH = 1
E	1711	6.55	0.56/0.50	3.51	13.6	GAG(197), GAA(35)
P	1571	6.35	0.41/0.40	2.54	3.7	CCG(29), CCT(29), CCA(3)
A	1250	6.66	0.49/0.40	3.15	7.9	GCG(35), GCC(36), GCT(18), GCA(9)
S	1191	6.65	0.45/0.40	2.88	9.7	AGC(87), TCC(25), TCG(2), TCA(1)
G	984	6.43	0.51/0.41	3.20	10.3	GGC(95), GGA(4), GGT(2)
L	1134	5.74	0.54/0.50	3.15	13.5	CTG(140), CTC(11), TTG(2)
Q	546	8.46	0.67/0.60	5.21	29.1	CAG(159)
K	433	5.80	0.52/0.40	3.01	13.2	AAG(42), AAA(15)

N, number of occurrences; Avg size, average length of amino acid tandem repeat; Avg longest codon run, average size of the longest pure codon run and CH, codon homogeneity.

Table 2 Amino acid tandem repeats with extreme size variation between human and chimpanzee. Repeat size difference was of at least 50% of the repeat tract, length of the longest repeat was at least 10

AA	Hs protein ID	Pt protein ID	Hs rep. length	Pt rep. length	Codon(s)	CH	Description
Longer in human							
A	ENSP00000313362	ENSPTRP0000001361	14	1	GCA/GCG	0.14	Myc-associated zinc finger protein
A	ENSP00000354703	ENSPTRP0000004905	17	8	GCA/GCC/GCT	0.12	Zinc finger protein 358
A	ENSP00000265081	ENSPTRP0000002917	12	6	GCA/GCC/GCT/ GCG	0.08	DNA mismatch repair protein Msh3
A	ENSP00000222753	ENSPTRP0000004809	14	7	GCG	0.21	Homoeobox protein Hox-A13
E	ENSP00000362441	ENSPTRP0000004735	15	5	GAG	0.33	Transcriptional regulator ATRX
E	ENSP00000252825	ENSPTRP0000001936	12	0	GAG	0.67	Sarcoplasmic reticulum histidine-rich calcium-binding protein precursor
E	ENSP00000264448	ENSPTRP0000002070	16	7	GAG	0.75	Alstrom syndrome protein 1
G	ENSP00000282111	ENSPTRP0000002080	10	3	GGC	0.70	Transcription factor 7-like 1
G	ENSP00000268489	ENSPTRP0000001424	14	7	GGC	0.57	Alpha-fetoprotein enhancer-binding protein
G	ENSP00000307342	ENSPTRP0000002887	12	6	GGC	0.58	Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 1
K	ENSP00000221282	ENSPTRP0000001853	16	8	AAA	1.00	Zinc finger protein 302
P	ENSP00000320948	ENSPTRP0000005251	10	5	CCG	0.30	Proline-rich membrane anchor 1 precursor
Q	ENSP00000349076	ENSPTRP0000004977	19	7	CAG	0.79	Atrophin-1 (DRPLA)
L	ENSP00000301873	ENSPTRP0000000669	10	4	CTG	1.00	Latent-transforming growth factor beta-binding protein 3 precursor
Longer in chimpanzee							
Q	ENSP00000296452	ENSPTRP0000002576	5	10	CAG	0.90	Protein bassoon
Q	ENSP00000352463	ENSPTRP0000002489	5	10	CAG	0.40	Transcription factor 20
Q	ENSP00000332163	ENSPTRP0000000225	4	10	CAG	1.00	Small proline-rich protein 4
E	ENSP00000263574	ENSPTRP0000000766	7	15	GAG	0.20	Amyloid-like protein 2 precursor
E	ENSP00000313731	ENSPTRP0000001580	6	13	GAG	0.77	Phospholipase C delta 3
S	ENSP00000302655	ENSPTRP0000004507	6	12	AGC	0.67	Uncharacterized protein C5orf25

Protein identifier (ID) and description were extracted from Ensembl; Hs rep. length and Pt rep. length, length of the human and chimpanzee repeats, respectively; Codon(s), codon(s) in the longest pure codon run and CH, codon homogeneity.

Table 3 Characteristics of disease-associated amino acid repeats in human and chimpanzee. Poly A and poly Q containing human proteins associated with disease and chimpanzee orthologues

Disease (gene)	Human peptide ID	Chimp peptide ID	Path. AA	Other AA	Hs rep. len.	Pt rep. len.	Path. rep. len.
Spinocerebellar ataxia 17 (TBP)	ENSP00000230354	ENSPTRP0000003214	Q		38	32	47–63
Spinocerebellar ataxia 2 (ATXN2)	NP_002964.2	XP_509374.2	Q		23	26	> 34
Spinal and bulbar muscular atrophy (AR)	NP_000035.2	NP_001009012	Q	QPAG	23	22	40–55
Huntington disease (HD)	ENSP00000347184	ENSPTRP0000002731	Q	EP	21	15	> 35
Dentatorubral pallidolusian atrophy (ATN1)	ENSP00000349076	ENSPTRP0000004977	Q	HPS	19	7	49–88
Spinocerebellar ataxia 1 (ATXN1)	NP_000323.2	XP_001170170.1	Q		12 + 14	10 + 10	39–81
Spinocerebellar ataxia 6 (CACNA1A)	ENSP00000353362	ENSPTRP0000004249	Q		13	11	20–29
Spinocerebellar ataxia 3 (ATXN3)	ENSP00000352872	ENSPTRP0000004400	Q		10	14	> 55
Spinocerebellar ataxia 7 (ATXN7)	ENSP00000295900	ENSPTRP0000002603	Q	APS	10	8	37–306
Congenital central hypoventilation syndrome (PHOX2B)	ENSP00000371160	ENSPTRP0000002755	A	AG	20	20	25–29
Cleidocranial dysplasia (RUNX2)	ENSP00000352514	ENSPTRP0000000061	A	Q	17	17	27
Synpolydactyly (HOXD13)	ENSP00000249505	ENSPTRP0000002164	A	AS	15	15	22–29
Holoprosencephaly (ZIC2)	ENSP00000365514	ENSPTRP0000001020	A	AGH	15	15	25
Blepharophimosis (FOXL2)	ENSP00000333188	ENSPTRP0000002661	A	GPR	14	14	22–24
Hand-foot-genital syndrome (HOXA13)	ENSP00000222753	ENSPTRP0000004809	A	A	14	7	24–26
Oculopharyngeal muscular dystrophy (PABPN1)	ENSP00000216727	ENSPTRP0000001045	A	E	10	10	12–17

Protein identifier (ID) was from Ensembl (ENS) or Refseq (NP, XP); Path. AA, amino acid in the disease-associated repeat; Other AA, other repeat tracts, of size 5 or longer, found in the human protein; Hs rep. len. and Pt rep. len., length of the repeats in the human and chimp wild-type proteins, respectively; Path. rep. len., length of pathogenic repeats, extracted from Gatchel and Zoghbi (2005) and Brown and Brown (2004).

Abnormal expansion of polyalanine repeats, the other class of pathogenic repeats, has been associated to congenital malformations, skeletal dysplasia and nervous system anomalies (Amiel *et al.*, 2004). Proteins with pathogenic tracts form inclusions that lead to cell death. Some of the expanded alanine repeats may have originated by unequal allelic homologous recombination rather than triplet slippage, as indicated by specific codon arrangement patterns in the mutated alleles, and higher stability of the mutated alleles over generations (Warren, 1997; Amiel *et al.*, 2004). Known genes where these mutations can occur are HOXD13 in synpolydactyly, PHOX2B in congenital central hypoventilation syndrome, RUNX2 in cleidocranial dysplasia, ZIC2 in holopresencephaly, FOXL2 in blepharophimosis, HOXA13 in hand-foot-genital syndrome, PABPN1 in oculopharyngeal muscular dystrophy, SOX3 in mental retardation and ARX in Partington syndrome. These genes are transcription factors, with the exception of PABPN1, involved in ribonucleic acid (RNA) processing. Although the function of alanine repeats is poorly understood, a role in transcriptional repression has been proposed (Lanz *et al.*, 1995).

It was early noticed that repeats associated with neurological disease showed differences in length between humans and other apes. In Rubinsztein *et al.* (1995c), analysis of CAG repeat length in a number of disease-associated genes – ATXN1, ATXN3 and AR – revealed shorter repeat lengths in some of the nonhuman primates with respect to humans. Another evolutionary study on the AR

gene indicated a trend towards CAG expansion in the higher primate species (Choong *et al.*, 1998). Subsequent studies (Andres *et al.*, 2003b; Andres *et al.*, 2004), revealed larger CAG repeat sizes in additional disease loci, such as SCA8. Using data from the human and chimpanzee genes available at Ensembl, we compiled wild-type repeat tract sizes for disease-associated human proteins and the corresponding chimpanzee orthologues (Table 3). All disease-associated glutamine repeats were of different size in the two species. In contrast, only one alanine repeat, in HOXA13, showed different repeat length. Most of the repeats were longer in humans than in chimpanzee (TBP, AR, HD, ATN-1, ATXN1 and CACNA1A), but some showed the opposite trend (ATXN2, ATXN3 and ATXN7). In all disease-associated glutamine repeats the predominant codon was CAG, and the average CH was 0.84. In alanine repeats the most abundant codon was GCG, but the trinucleotide repeat tracts tended to be much more interrupted, with average CH = 0.3. This observation is consistent with the idea that some disease-associated alanine repeats expand by unequal crossover instead of trinucleotide slippage (Warren, 1997).

Figure 1 shows some illustrative human and chimpanzee protein alignments, including the expandable repeat and repeat-surrounding regions. In general, more amino acid sequence variability was observed in both the repeat itself and repeat-surrounding regions in proteins containing glutamine repeats than in those containing alanine repeats. This trend was consistent with previous observations, using

PolyA disease-associated repeat alignments



PolyQ disease-associated repeat alignments

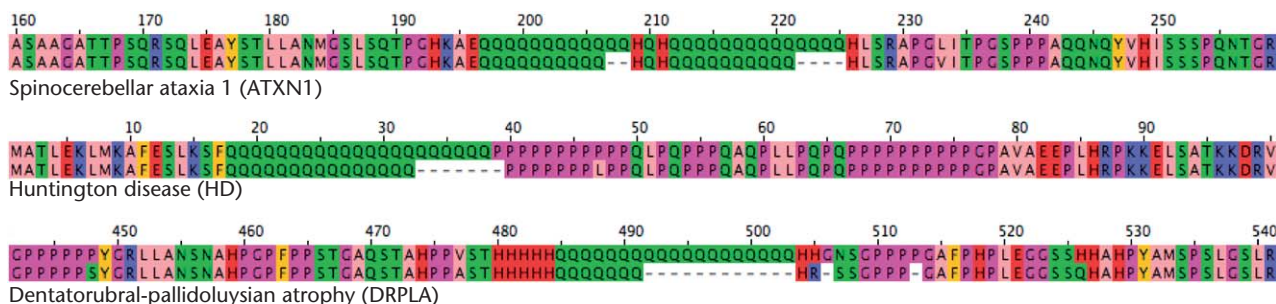


Figure 1 Alignments of protein regions containing human disease-associated repeats. The sequence at the top corresponds to the human sequence and the one at the bottom to the orthologous chimpanzee sequence. See Table 3 for more details.

human and mouse orthologues, of lower evolutionary rates in sequences surrounding well-conserved repeats than in sequences surrounding highly variable repeats (Hancock *et al.*, 2001; Mularoni *et al.*, 2007). To assess whether this trend could be generalized in the primate dataset, we defined a group of proteins with highly conserved repeats (HC; size ≥ 8 , no size difference between the two species, 617 repeats in 527 proteins) and a group of proteins with highly variable repeats (HV; size in at least one species ≥ 8 , size difference $\geq 25\%$, 168 repeats in 157 proteins). Interestingly, differences in evolutionary rates (discarding repetitive regions) were also apparent at this short phylogenetic distance: average nonsynonymous nucleotide substitution rate (K_a) was 0.0057 for proteins that contained only HV repeats, whereas it was 0.0031 for proteins that contained only HC repeats ($p < 0.001$, Kolmogorov–Smirnov test). Similar significant differences were found for the nonsynonymous to synonymous nucleotide substitution rate ratio (K_a/K_s). Therefore, the study showed that highly variable repeats, such as those

involved in glutamine expansion diseases, were generally embedded in rapidly evolving protein sequences.

Glutamine Repeat Polymorphism and Disease

A well-documented trait of pathogenic glutamine repeats is their high polymorphism within human populations (Rubinsztein *et al.*, 1995c; Jodice *et al.*, 1997; Hsing *et al.*, 2000; Andres *et al.*, 2003a; Hong *et al.*, 2006; Rozanska *et al.*, 2007). This is not surprising, if we consider that these repeats are encoded by long pure CAG triplets, and repeat purity is, in general, linked to increased polymorphism levels (Wren *et al.*, 2000; O'Dushlaine *et al.*, 2005; Mularoni *et al.*, 2006). A comparison of CAG repeat loci in four major human populations, including eight pathogenic and two non-pathogenic repeats, indicated that population differences account for a very small part of the total variation (Andres

Table 4 CAG repeat allele distribution in disease-associated loci.

Repeat length distributions were extracted from Andres *et al.* (2004), except for AR from Hong *et al.* (2006)

Gene	Disease	Species	CAG repeat size range
ATXN2	Spinocerebellar ataxia 2	Human	15–33
		Chimpanzee	22–27
		Gorilla	16–18
		Orangutan	16–17
AR	Spinal and bulbar muscular atrophy	Human	8–35
		Chimpanzee	14–26
		Gorilla	7–9
		Orangutan	12
DRPLA	Dentatorubral-pallidolysian atrophy	Human	6–35
		Chimpanzee	11–17
		Gorilla	9
		Orangutan	15
ATXN1	Spinocerebellar ataxia 1	Human	14–38
		Chimpanzee	20–26
		Gorilla	20–22
		Orangutan	20–25
CACNA1A	Spinocerebellar ataxia 6	Human	4–19
		Chimpanzee	9–13
		Gorilla	8–10
		Orangutan	11–13
ATXN3	Spinocerebellar ataxia 3	Human	14–40
		Chimpanzee	14–20
		Gorilla	8–11
		Orangutan	24–25
SCA8	Spinocerebellar ataxia 8	Human	15–42
		Chimpanzee	12–21
		Gorilla	10–21
		Orangutan	7
PPP2R2B	Spinocerebellar ataxia 12	Human	6–45
		Chimpanzee	7–19
		Gorilla	8–9
		Orangutan	8–11

et al., 2003a). Repeat length distribution varies greatly depending on the loci, suggesting that repeat evolutionary dynamics are, to a large extent, loci-dependent (Andres *et al.*, 2003a; Butland *et al.*, 2007). In Butland *et al.* (2007) they characterized the human allele length distribution of 64 glutamine-encoding CAG tracts in a set of individuals of mixed ethnicity, including nine loci involved in disease. They found that the best predictors of known disease repetitive tracts were a long uninterrupted CAG tract and high repeat length variance in the normal population.

As shown before, differences in CAG repeat size between human and chimpanzee, in disease-associated loci, is in general, relatively small. However, CAG expansion diseases are only found in humans, which suggest there are lineage-specific differences in the susceptibility to suffer large expansions in these loci. In Andres *et al.*, (2004) they compared CAG repeat polymorphism in human and apes, using nine loci that comprised the disease-associated genes ATXN1/SCA1, ATXN2/SCA2, ATXN3/SCA3, CACNA1A/SCA6, SCA8 (noncoding), PPP2R2B/SCA12 (noncoding) and DRPLA, and the nondisease genes KCNN3 (potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3) and NCOA3 (nuclear receptor coactivator 3). Twenty chimpanzees, 13 gorillas and 4–6 orangutans were typed for the number of repeats at these loci. They found that, while humans presented slightly larger alleles than the rest of species in most loci, the most relevant difference was in intraspecific repeat size diversity levels (Table 4). The variation in the number of repeats showed a decreasing trend from humans to orangutans. In contrast, repeat size variance in the two nonexpanding loci was similar in the four species. Overall, the data suggested a link between high repeat size variance and increased species-specific susceptibility to disease. In Hong *et al.* (2006) they compared androgen receptor CAG (glutamine-encoding) and GGN (glycine-encoding) repeat polymorphism in humans and apes. In healthy humans, it had been reported that repeat size ranged from 8 to 35 for the CAG repeat and from 10 to 30 for the CGN repeat (Hsing *et al.*, 2000). Diversity in the chimpanzees was lower, with size ranges from 14 to 26 for CAG and from 14 to 22 for GGN. Gorillas were less polymorphic than chimpanzees in both types of repeats. Orangutans, agile gibbons and siamangs exhibited monomorphic alleles for CAG, although great diversity was observed for GGN repeats in siamangs. In chimpanzees, as in humans, alleles with long CAG and short GGN were particularly frequent, pointing to a combined functional activity of the two regions.

Summary

Trinucleotide repeats can suffer expansions or contractions of the repetitive tract by replication slippage. The uncontrolled expansion of these tracts within genes has been associated with a number of neurodegenerative and developmental human diseases. Interestingly, there are important differences between human and other apes in the

characteristics of CAG repeats associated with neurodegenerative disorders, with a trend towards longer and more diverse repeat tracts in humans. At a genome-wide scale, however, the tendency for longer trinucleotide repeats in the human genome with respect to the chimpanzee genome is rather minor and heterogeneous. In coding sequences, trinucleotide repeats generate single amino acid repeats or homopeptides. These structures have been involved in the modulation of protein–protein interactions and transcriptional activation. Comparison of orthologous human and chimpanzee amino acids repeats (of size 5 or longer) reveals that approximately 17% (864 repeats) show some difference in length between the two species. In addition, there seem to be an excess of amino acid repeats that are much longer in humans than the opposite case. The high variation observed in these sequences provides a substrate for possible evolutionary adaptations in the human and ape lineages.

References

- Alba MM and Guigo R (2004) Comparative analysis of amino acid repeats in rodents and humans. *Genome Research* **14**(4): 549–554.
- Alba MM, Santibanez-Koref MF and Hancock JM (1999) Conservation of polyglutamine tract size between mice and humans depends on codon interruption. *Molecular Biology and Evolution* **16**(11): 1641–1644.
- Alba MM, Tompa P and Veitia RA (2007) Amino acid repeats and the structure and evolution of proteins. In: Volff J-N (ed.) *Gene and Protein Evolution* pp. 119–130. Basel: Karger.
- Amiel J, Trochet D, Clement-Ziza M, Munnich A and Lyonnet S (2004) Polyalanine expansions in human. *Human Molecular Genetics* **13**(Spec. no. 2): R235–R243.
- Anan K, Yoshida N, Kataoka Y *et al.* (2007) Morphological change caused by loss of the taxon-specific polyalanine tract in Hoxd-13. *Molecular Biology and Evolution* **24**(1): 281–287.
- Andres AM, Lao O, Soldevila M, Calafell F and Bertranpetit J (2003a) Dynamics of CAG repeat loci revealed by the analysis of their variability. *Human Mutation* **21**(1): 61–70.
- Andres AM, Soldevila M, Lao O *et al.* (2004) Comparative genetics of functional trinucleotide tandem repeats in humans and apes. *Journal of Molecular Evolution* **59**(3): 329–339.
- Andres AM, Soldevila M, Saitou N *et al.* (2003b) Understanding the dynamics of Spinocerebellar Ataxia 8 (SCA8) locus through a comparative genetic approach in humans and apes. *Neuroscience Letters* **336**(3): 143–146.
- Borstnik B and Pumpernik D (2002) Tandem repeats in protein coding regions of primate genes. *Genome Research* **12**(6): 909–915.
- Britten RJ (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proceedings of the National Academy of Sciences of the USA* **99**(21): 13633–13635.
- Brown LY and Brown SA (2004) Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends in Genetics* **20**(1): 51–58.

- Butland SL, Devon RS, Huang Y *et al.* (2007) CAG-encoded polyglutamine length polymorphism in the human genome. *BMC Genomics* **8**: 126.
- Caburet S, Cocquet J, Vaiman D and Veitia RA (2005) Coding repeats and evolutionary "agility". *BioEssays* **27**(6): 581–587.
- Choong CS, Kempainen JA and Wilson EM (1998) Evolution of the primate androgen receptor: a structural basis for disease. *Journal of Molecular Evolution* **47**(3): 334–342.
- Clark RM, Bhaskar SS, Miyahara M, Dalgliesh GL and Bidichandani SI (2006) Expansion of GAA trinucleotide repeats in mammals. *Genomics* **87**(1): 57–67.
- Consortium TCSaA (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055): 69–87.
- Cooper G, Rubinsztein DC and Amos W (1998) Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. *Human Molecular Genetics* **7**(9): 1425–1429.
- Ellegren H (2000) Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genetics* **24**(4): 400–402.
- Ellegren H, Primmer CR and Sheldon BC (1995) Microsatellite 'evolution': directionality or bias? *Nature Genetics* **11**(4): 360–362.
- Fondon JW 3rd and Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the USA* **101**(52): 18058–18063.
- Forbes SH, Hogg JT, Buchanan FC, Crawford AM and Allendorf FW (1995) Microsatellite evolution in congeneric mammals: domestic and bighorn sheep. *Molecular Biology and Evolution* **12**(6): 1106–1113.
- Galant R and Carroll SB (2002) Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* **415**(6874): 910–913.
- Gatchel JR and Zoghbi HY (2005) Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews. Genetics* **6**(10): 743–755.
- Gerber HP, Seipel K, Georgiev O *et al.* (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**(5148): 808–811.
- Glazko GV and Nei M (2003) Estimation of divergence times for major lineages of primate species. *Molecular Biology and Evolution* **20**(3): 424–434.
- Green H and Wang N (1994) Codon reiteration and the evolution of proteins. *Proceedings of the National Academy of Sciences of the USA* **91**(10): 4298–4302.
- Hancock JM and Simon M (2005) Simple sequence repeats in proteins and their significance for network evolution. *Gene* **345**(1): 113–118.
- Hancock JM, Worthey EA and 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. *Molecular Biology and Evolution* **18**(6): 1014–1023.
- Hong KW, Hibino E, Takenaka O *et al.* (2006) Comparison of androgen receptor CAG and GGN repeat length polymorphism in humans and apes. *Primates* **47**(3): 248–254.
- Hsing AW, Gao YT, Wu G *et al.* (2000) Polymorphic CAG and GGN repeat lengths in the androgen receptor gene and prostate cancer risk: a population-based case-control study in China. *Cancer Research* **60**(18): 5111–5116.
- Hubbard T, Andrews D, Caccamo M *et al.* (2005) Ensembl 2005. *Nucleic Acids Research* **33**(Database issue): D447–D453.
- Huntley MA and Clark AG (2007) Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Molecular Biology and Evolution* **24**(12): 2598–2609.
- Jodice C, Giovannone B, Calabresi V *et al.* (1997) Population variation analysis at nine loci containing expressed trinucleotide repeats. *Annals of Human Genetics* **61**(Pt 5): 425–438.
- Karlin S, Brocchieri L, Bergman A, Mrazek J and Gentles AJ (2002) Amino acid runs in eukaryotic proteomes and disease associations. *Proceedings of the National Academy of Sciences of the USA* **99**(1): 333–338.
- Kashi Y and King DG (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* **22**(5): 253–259.
- Kehrer-Sawatzki H and Cooper DN (2007) Understanding the recent evolution of the human genome: insights from human–chimpanzee genome comparisons. *Human Mutation* **28**(2): 99–130.
- Kumar S, Filipski A, Swarna V, Walker A and Hedges SB (2005) Placing confidence limits on the molecular age of the human–chimpanzee divergence. *Proceedings of the National Academy of Sciences of the USA* **102**(52): 18842–18847.
- Lander ES, Linton LM, Birren B *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860–921.
- Lanz RB, Wieland S, Hug M and Rusconi S (1995) A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. *Nucleic Acids Research* **23**(1): 138–145.
- Leakey MG, Feibel CS, McDougall I, Ward C and Walker A (1998) New specimens and confirmation of an early age for *Australopithecus anamensis*. *Nature* **393**(6680): 62–66.
- Levinson G and Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* **4**(3): 203–221.
- Messer PW and Arndt PF (2007) The majority of recent short DNA insertions in the human genome are tandem duplications. *Molecular Biology and Evolution* **24**(5): 1190–1197.
- Mularoni L, Guigo R and Alba MM (2006) Mutation patterns of amino acid tandem repeats in the human proteome. *Genome Biology* **7**(4): R33.
- Mularoni L, Veitia RA and Alba MM (2007) Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* **89**(3): 316–325.
- O'Dushlaine CT, Edwards RJ, Park SD and Shields DC (2005) Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biology* **6**(8): R69.
- Primmer CR and Ellegren H (1998) Patterns of molecular evolution in avian microsatellites. *Molecular Biology and Evolution* **15**(8): 997–1008.
- Puente XS, Velasco G, Gutierrez-Fernandez A *et al.* (2006) Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC Genomics* **7**(1): 15.
- Rozanska M, Sobczak K, Jasinska A *et al.* (2007) CAG and CTG repeat polymorphism in exons of human genes shows distinct features at the expandable loci. *Human Mutation* **28**(5): 451–458.

- Rubinsztein DC, Amos W, Leggo J *et al.* (1995a) Microsatellite evolution – evidence for directionality and variation in rate between species. *Nature Genetics* **10**(3): 337–343.
- Rubinsztein DC, Leggo J and Amos W (1995b) Microsatellites evolve more rapidly in humans than in chimpanzees. *Genomics* **30**(3): 610–612.
- Rubinsztein DC, Leggo J, Coetzee GA *et al.* (1995c) Sequence variation and size ranges of CAG repeats in the Machado–Joseph disease, spinocerebellar ataxia type 1 and androgen receptor genes. *Human Molecular Genetics* **4**(9): 1585–1590.
- Shimohata M, Shimohata T, Igarashi S, Naruse S and Tsuji S (2005) Interference of CREB-dependent transcriptional activation by expanded polyglutamine stretches – augmentation of transcriptional activation as a potential therapeutic strategy for polyglutamine diseases. *Journal of Neurochemistry* **93**(3): 654–663.
- Tautz D and Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research* **12**(10): 4127–4138.
- Toth G, Gaspari Z and Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research* **10**(7): 967–981.
- Vowles EJ and Amos W (2006) Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. *Molecular Biology and Evolution* **23**(3): 598–607.
- Warren ST (1997) Polyalanine expansion in synpolydactyly might result from unequal crossing-over of HOXD13. *Science* **275**(5298): 408–409.
- Webster MT, Smith NG and Ellegren H (2002) Microsatellite evolution inferred from human–chimpanzee genomic sequence alignments. *Proceedings of the National Academy of Sciences of the USA* **99**(13): 8748–8753.
- Wren JD, Forgacs E, Fondon JW III *et al.* (2000) Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *American Journal of Human Genetics* **67**(2): 345–356.

Further Reading

- Fry M and Usdin K (eds) (2006) *Human Nucleotide Expansion Disorders (Nucleic Acids and Molecular Biology)*. Berlin: Springer.
- Goldstein DB and Schlotterer C (eds) (1999) *Microsatellites: Evolution and Applications*. Oxford, UK: Oxford University Press.
- Li W-H (2006) *Molecular Evolution*. Sunderland: Sinauer Associates Inc.
- Rubinsztein D (1998) *Analysis of Triplet Repeat Disorders (Human Molecular Genetics)*. London, UK: Garland Science.
- Yang Z (2006) *Computational Molecular Evolution*. Oxford, UK: Oxford University Press.