Available online at www.sciencedirect.com

SCIENCE DIRECT®

Genomics 87 (2006) 57 – 67

GENOMICS

www.elsevier.com/locate/ygeno

Expansion of GAA trinucleotide repeats in mammals

Rhonda M. Clark [a,1], Sanjeev S. Bhaskar [a,1], Masaki Miyahara [a], Gillian L. Dalgliesh [a], Sanjay I. Bidichandani [a,b,*]

[a] *Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA*
[b] *Department of Pediatrics, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA*

Received 19 May 2005; accepted 7 September 2005
Available online 28 November 2005

Abstract

We have previously shown that GAA trinucleotide repeats have undergone significant expansion in the human genome. Here we present the analysis of the length distribution of all 10 nonredundant trinucleotide repeat motifs in 20 complete eukaryotic genomes (6 mammalian, 2 nonmammalian vertebrates, 4 arthropods, 4 fungi, and 1 each of nematode, amoebozoa, alveolate, and plant), which showed that the abundance of large expansions of GAA trinucleotide repeats is specific to mammals. Analysis of human–chimpanzee–gorilla orthologs revealed that loci with large expansions are species-specific and have occurred after divergence from the common ancestor. PCR analysis of human controls revealed large expansions at multiple human $(GAA)_{30+}$ loci; nine loci showed expanded alleles containing >65 triplets, analogous to disease-causing expansions in Friedreich ataxia, including two that are in introns of genes of unknown function. The abundance of long GAA trinucleotide repeat tracts in mammalian genomes represents a significant mutation potential and source of interindividual variability.

© 2005 Elsevier Inc. All rights reserved.

Keywords: GAA trinucleotide repeat; Friedreich ataxia; Expansion; Premutation; Mammal; Genome

The availability of genome sequences for many organisms has greatly aided the analysis of simple sequence or microsatellite repeats. Their distribution, which depends on the type of repeat, displays species-specific and genomic region-specific differences. In mammalian intergenic regions di-, tetra-, and pentanucleotide repeats are more frequent than trinucleotide repeats, and tri- and hexanucleotide repeats are more abundant in protein-coding exons [1]. Exonic repeats are under selective pressure at the level of protein function and codon usage, resulting in significant interspecies variability. Surprising differences are also seen in intraspecies distribution of microsatellite repeats in introns versus intergenic regions [1].

Trinucleotides constitute a special class of microsatellite repeat because of their ability to undergo significant expansion during intergenerational transmission. At least 15 inherited neurodegenerative diseases are known to be caused by abnormally large trinucleotide repeat expansions within coding or noncoding sequences of genes [2,3]. So far, expansion of only three trinucleotide repeats has been shown to cause human disease: CAG · CTG, CGG · CCG, and GAA · TTC. Expansion of these sequences occurs in a length- and sequence-dependent manner. Uninterrupted repeat tracts that attain a certain threshold length (premutation alleles) may become unstable and undergo large expansions upon intergenerational transmission (mutant alleles). The threshold length for the initiation of genetic instability differs for the various sequences and genomic loci, but is usually in the 30–40 triplet-repeat range.

Friedreich ataxia is an autosomal recessive disease that results when one inherits an *FXN* gene carrying large GAA trinucleotide repeat expansions from either parent [4]. The GAA trinucleotide repeat sequence at the *FXN* locus is polymorphic: normal alleles contain <30 triplets, and disease-causing expanded (E) alleles have 66–1700 triplets [4–7]. Among Indo-Europeans, 0.65% of chromosomes have E alleles [8] and they are maintained in the population via asymptomatic heterozygous carriers and occasional hyperexpansion of intermediate-sized, premutation (PM) alleles containing 30–65 uninterrupted repeats [5–7,9]. The GAA trinucleotide

[*] Corresponding author. Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA. Fax: +1 405 271 3910.
E-mail address: Sanjay-Bidichandani@ouhsc.edu (S.I. Bidichandani).
[1] These authors contributed equally to this work.

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.09.006

repeat sequence in the *FXN* gene has expanded in recent human evolution [10–12]. It originated at the center of an *Alu* element that inserted in an early primate genome, and while all nonhuman primates examined have <5 triplets, most humans have 8–11 triplets (short normal alleles). The stepwise transition to long normal (12–30 triplets), PM, and E alleles has occurred only in Indo-Europeans, and consequently, only Indo-Europeans and populations that have significant European genetic admixture develop Friedreich ataxia [11,13]. Friedreich ataxia is so far the only disease that is caused by abnormal expansion of a GAA repeat sequence.

We have previously shown by analysis of the entire human genome sequence that GAA trinucleotide repeats have undergone significant expansion compared with the other nine trinucleotide repeat motifs [14]. Here we report that the relative expandability of GAA trinucleotide repeats is seen mainly in mammalian genomes. Analysis of primate genomes revealed that several loci have expanded to premutation lengths in recent primate evolution, after divergence of the great apes. Analysis of allelic variability in human controls revealed the presence of expanded GAA trinucleotide repeat alleles that are analogous to *FXN* gene mutations, i.e., alleles containing >65 triplets, at multiple loci, including two novel intragenic expansions. The abundance of long GAA trinucleotide repeats in mammalian genomes represents an important source of genetic variability via germ-line mutation. Furthermore, given the ability of long GAA trinucleotide repeats to adopt non-B DNA structures [15–21], to interfere with DNA replication [16,21,22] and gene transcription [15,16,18,19], to mediate position effect variegation [23], and to cause flanking mutagenesis [24], these sequences are likely to have important implications for mammalian genomic structure and function.

Results

Expansion of GAA trinucleotide repeats is seen mainly in mammalian genomes

The frequency and size distribution of all trinucleotide repeats were analyzed in the following 20 complete eukaryotic genomes (see Methods): 6 mammalian, *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog), and *Monodelphis domestica* (opossum); 2 other vertebrates, *Gallus gallus* (chicken) and *Danio rerio* (zebrafish); four arthropods, *Drosophila melanogaster* (fly), *D. yakuba* (fly), *Apis mellifera* (honey bee), *Anopheles gambiae* (mosquito); a nematode, *Caenorhabditis elegans;* four fungal, *Saccharomyces cerevisiae, Schizosaccharomyces pombe, Candida albicans, Candida galbrata;* an amoebozoan, *Dictyostelium discoideum;* a plant, *Arabidopsis thaliana;* and an alveolate, *Plasmodium falciparum.* A search algorithm was used to detect all 10 nonredundant trinucleotide repeats: GAA · TTC, TAA · TTA, CAA · TTG, GGA · TCC, CAG · CTG, GTA · TAC, GAC · GTC, GAT · ATC, CGG · CCG, and GTG · CAC (referred to as GAA, TAA, CAA, GGA, CAG, GTA, GAC, GAT, CGG, and GTG, respectively). All independent repeat tracts consisting of ≥2 uninterrupted trinucleotide repeats were thus identified (Fig. 1).

Long repeat tracts of the GAA trinucleotide repeat motif were identified in the 6 mammalian genomes in comparison with all nine other trinucleotide repeats (Figs. 1A–1F). This expansion was not so marked in the 14 nonmammalian genomes, including the 2 nonmammalian vertebrates (Figs. 1G–1T). Analysis of the range of repeat length indicated that the upper bounds of the 6 mammalian genomes and that of *Anopheles* were considerably higher than the other 13 genomes (Fig. 2A). To correct for the large size difference among the 20 genomes, we analyzed the relative length of $(GAA)_{6+}$ repeat tracts as a proportion of all $(NAA)_{6+}$ trinucleotide repeats (GAA, TAA, CAA), as an indicator of the expandability of GAA trinucleotide repeats. $(NAA)_{6+}$ sequences were selected because we have previously shown that repeats of this length are unlikely to have arisen simply by chance in large genomes [14]. This distinguished the 6 mammalian genomes from the 14 other genomes (Fig. 2B). Among mammals, rodent genomes showed exceptionally long GAA trinucleotide repeats. It is interesting to note that in 5 of the 6 mammalian genomes there is evidence of bimodality in the size distribution, with an apparent overrepresentation of repeat tracts containing 14–30 triplets (seen as a "hump" in Figs. 1A–1E). *M. domestica* stands out from the rest of the mammals for having exceptionally long GAA repeat tracts (Fig. 1F).

In the nonmammalian genomes TAA, CAA, and GTA repeats showed maximal expansion. Some nonmammalian genomes showed very long A-rich triplets; *Anopheles* showed long TAA and GAA repeats, *Dictyostelium* had long TAA and CAA repeats, while *S. pombe* had long CAA repeats.

The other two trinucleotide motifs associated with human disease showed far fewer long repeat tracts. The CGG trinucleotide repeat motif showed the least expansion of all motifs, representing the shortest of all triplets in 12 of the 20 genomes, and as many as 5 genomes lacked even a single $(CGG)_{6+}$ tract. The CAG trinucleotide repeat, which causes the most expansion-associated diseases, did not show very long tracts in any genome. However, our analysis did not differentiate between polyglutamine-encoding versus noncoding CAG repeats.

Recent expansion of GAA trinucleotide repeats following evolutionary divergence of primates

To determine the timing of the evolutionary expansion to lengths analogous to premutations at the *FXN* locus (30–65 triplets), we identified and compared orthologs of human $(GAA)_{30+}$ sequences in the chimpanzee genome and, conversely, the orthologs of chimpanzee $(GAA)_{25+}$ sequences in the human genome. Of the previously identified 29 human $(GAA)_{30+}$ sequences [14], 25 were confirmed by PCR and sequencing (see supplementary table for primer sequences). Orthologous chimpanzee (n = 22) loci were amplified by PCR using human-specific primers and sequenced. Three additional loci (of the ones that were not amenable to PCR) were identified by homology searches of the publicly available chimpanzee genome sequence, for a total of 25 orthologs of the 25 human loci. Conversely, orthologous loci in the human genome (n = 43) of all 56 of the chimpanzee $(GAA)_{25+}$ sequences were identified by sequence homology searches. In several cases, loci containing long GAA trinucleotide repeats in the human genome did not contain long repeats at the corresponding loci in chimpanzee, and vice versa. In 79% (19/24) of cases the orthologous chimpanzee loci of the human $(GAA)_{30+}$ sequences contained <13 repeats (Fig. 3A). The median lengths of human and chimpanzee loci were 32 and 7 triplets, respectively ($p < 0.001$). Likewise, 51% (22/43) of human loci that corresponded to the chimpanzee loci with $(GAA)_{25+}$ had <13 repeats (Fig. 3B). The median lengths of human and chimpanzee loci were 13 and 26 triplets, respectively ($p < 0.001$).

Despite the significant difference between human and chimpanzee orthologs, these data do not tell us if there was a lineage-specific expansion or contraction from the common ancestor. We reasoned that determining the repeat lengths at the corresponding loci in the gorilla genome would provide useful clues as to the lineage-specific changes since the common ancestor of the great apes. We therefore identified 21 orthologous gorilla loci by PCR and sequencing using human-specific primers for the 25 human $(GAA)_{30+}$ sequences (only ~0.3% of the gorilla genome sequence was publicly available at the time of this study and homology searches did not reveal any additional loci). We were thus able to analyze 21 trispecies, human–chimpanzee–gorilla orthologs. We found flanking sequence homology of the nine gorilla orthologs that did not map within *Alu* elements (all those that mapped in *Alu* elements were in identical *Alu* subclasses across all three species) and 93–100% and 91–99% sequence identity with the corresponding human and chimpanzee loci. In 11 of the 21 trispecies orthologs, both chimpanzee and gorilla orthologs contained <13 repeats at the loci where the human genome contains >30 repeats (asterisks in Fig. 3A). These data indicate that these loci have expanded exclusively in the human lineage, following the divergence of the great apes.

Another caveat of comparing human versus chimpanzee orthologs is the uncertainty that the same sequence actually existed in the common ancestor, especially when dealing with mutable sequences like repeats (notwithstanding the high degree of flanking homology). However, given the preferential association of GAA trinucleotide repeats with *Alu* elements [14], we decided to compare the lengths of only those that mapped within old *Alu* elements (*Alu* J, *Alu* S, free monomers), i.e., those that existed in the primate genome before the divergence of the great apes. Indeed, 13 of the 25 human $(GAA)_{30+}$ sequences and 25 of the 56 chimpanzee $(GAA)_{25+}$ sequences for which interspecies orthologs were identified mapped within *Alu* elements, and in almost every case this involved identical *Alu* subfamilies (Tables 1 and 2; gorilla results not shown). Only one of the 39 orthologous pairs mapped in an *Alu* Y element, which was excluded from this analysis. Comparison of repeat lengths showed that the interspecies divergence in the length of GAA trinucleotide sequences is significant; while the median length of all human $(GAA)_{30+}$ sequences was 32 triplets, orthologous regions in chimpanzee and gorilla showed median repeat lengths of 7 and 8, respectively ($p < 0.001$ in both cases). Comparison of all chimpanzee $(GAA)_{25+}$ sequences to the orthologous human loci showed median GAA tract lengths of 26 and 7, respectively ($p < 0.001$). Altogether, these data support the notion that expansion to premutation length seems to have occurred in a locus-specific and lineage-specific manner, following the divergence of the great apes.

It is interesting in this regard that the GAA repeat at the *FXN* locus has expanded to premutation length in humans (involving specific populations [11–13]) but not in chimpanzee [10,12]. Note that of all the genomes analyzed in this study, only the chimpanzee has a GAA repeat at the orthologous locus akin to the human *FXN* gene. Indeed, the chimpanzee orthologous GAA repeat sequence is also located in an *Alu* Sx element, similar to the human *FXN* gene, and the upstream and downstream intron sequences flanking the *Alu* element (1000 bp on either side) together show 98.4% identity. Furthermore, we had previously identified two other GAA repeat tracts in the human genome similar to the *FXN* locus in their location at the center of *Alu* elements (8q13 and 10q24) [14]. The repeat tract at 10q24 showed an allelic distribution similar to that of the human *FXN* locus, including the presence of premutation length alleles. The chimpanzee ortholog of this sequence revealed a short GAA repeat tract at the center of an *Alu* element, with 98.5% identity of the flanking sequence (1000 bp on each side of the *Alu* element), indicating that the transition to premutation length at 10q24 may also be specific to the human lineage.

Large expansions of potential premutation GAA repeat sequences at multiple loci in the human genome

To determine the mutability of the potential premutation GAA trinucleotide repeats, i.e., if they behave analogous to premutations at the *FXN* locus, we analyzed the allelic distribution at all 25 confirmed $(GAA)_{30+}$ sequences in the human genome. Initial analysis of approximately 20 chromosomes derived from unrelated, normal human controls (CEPH polymorphism discovery panel) showed that in almost all cases the repeat lengths were ≥30 triplets (Fig. 4). We selected 11 of these loci for additional analysis of polymorphic variability and for evidence of large expansions. For each of the 11 loci, repeat lengths were estimated by PCR amplification of 40–70 unrelated chromosomes derived from normal human controls (Fig. 4; right). Analysis of all 25 loci revealed 9 loci at which allele lengths exceeded 65 triplets, i.e., analogous to disease-causing mutations at the *FXN* locus. Allele lengths even exceeded 100 triplets at 5 of these loci. We sequenced three to eight alleles (of various sizes) for each of these loci to confirm the presence of GAA trinucleotide repeats; in some of the cases we observed interrupted sequences with more than one polymorphic GAA tract in the same PCR product.

Interestingly, the 19 human $(GAA)_{30+}$ loci that showed the largest interspecies length difference between human and chimpanzee genomes (Fig. 3A) also showed a significantly higher prevalence of expanded alleles (frequency of alleles with >65 triplets) compared with the 6 loci that showed

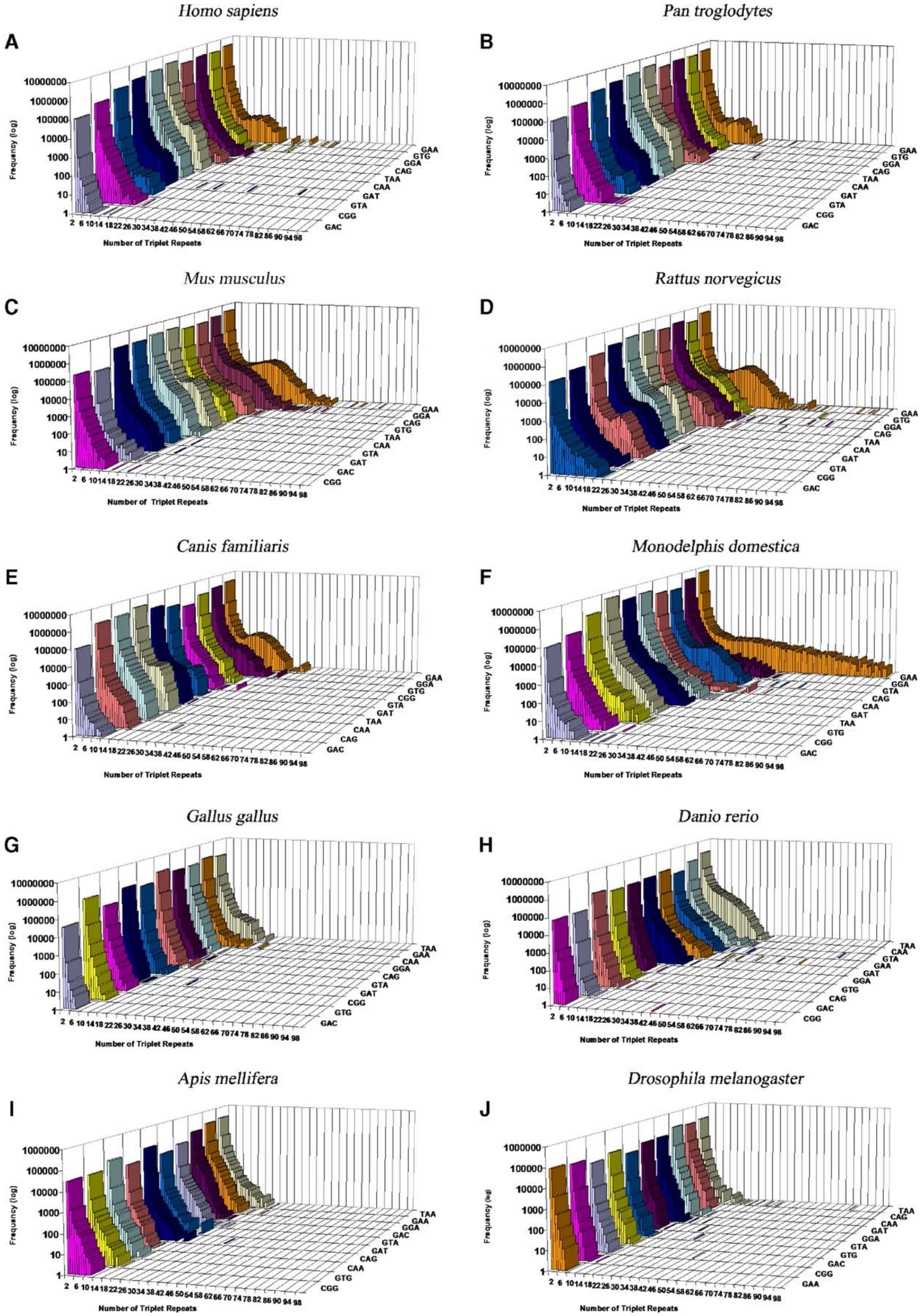

Fig. 1. Length and frequency distribution of the 10 nonredundant trinucleotide repeats in 20 complete eukaryotic genomes showing expansion of GAA repeats in mammalian genomes. Repeat length is plotted on the *x* axis (in triplets) and the frequency (log distribution) of each length is shown on the *y* axis. The repeat motifs with the highest prevalence of long repeat tracts are at the far end on the *z* axis. The same colors are used to depict triplet motifs in all 20 plots (GAA is shown in orange).

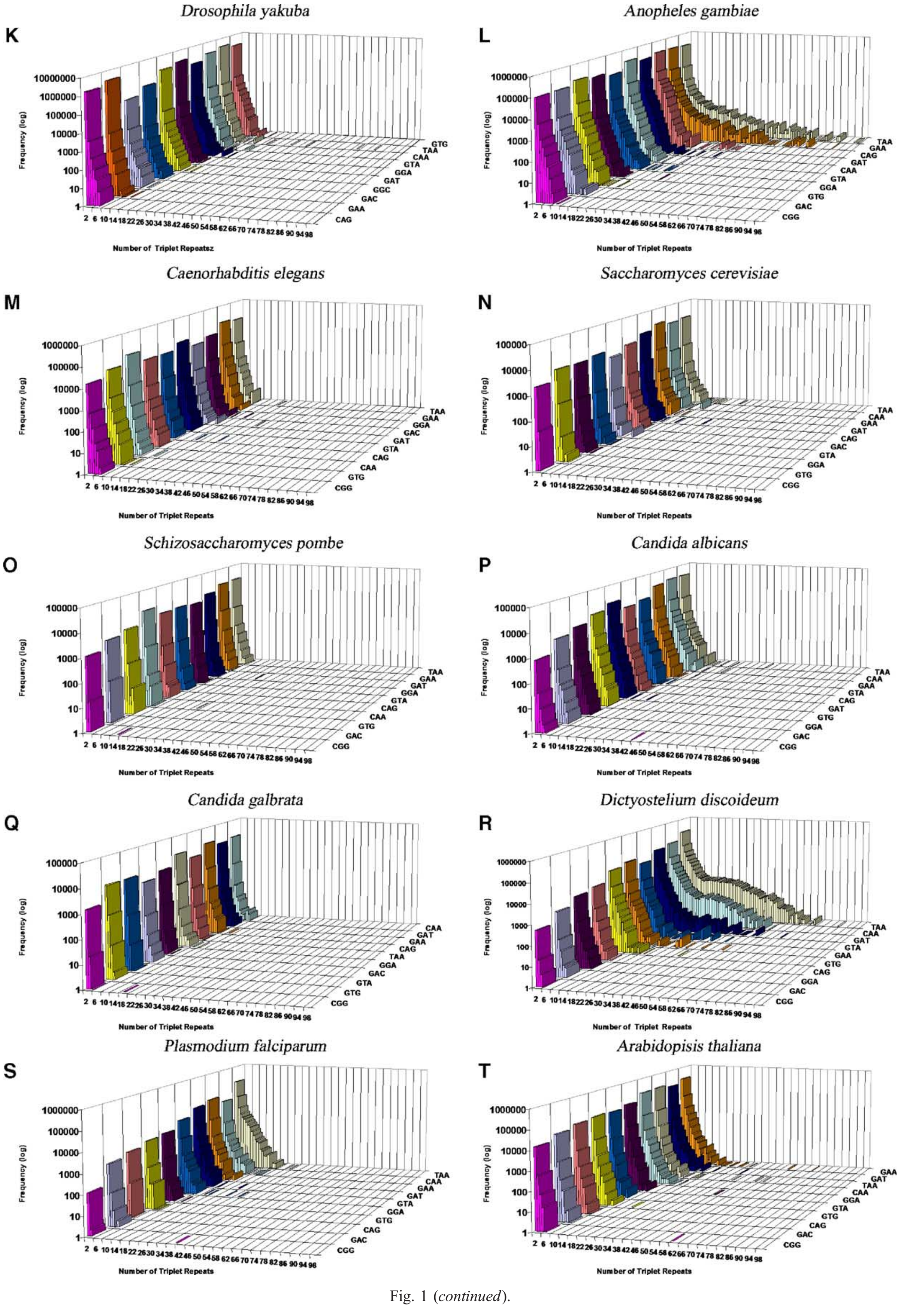

Fig. 1 (*continued*).

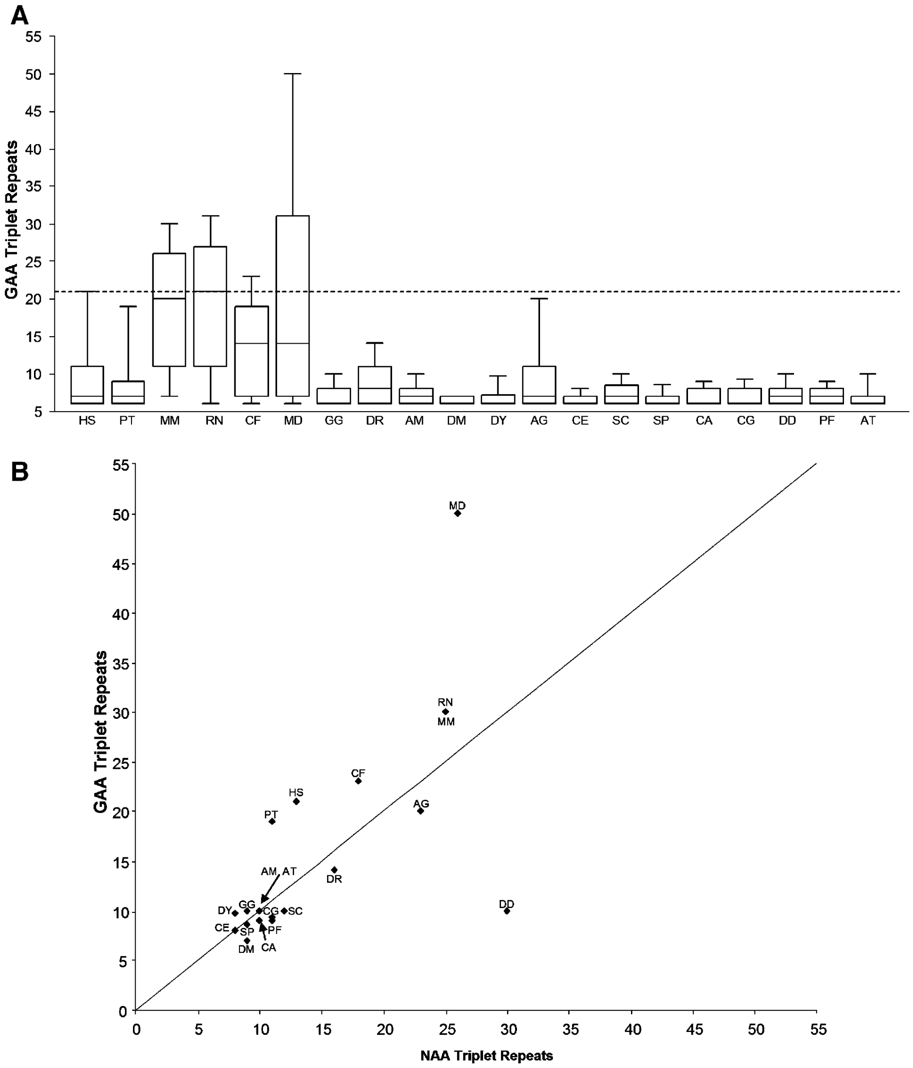

Fig. 2. GAA trinucleotide repeats have undergone significant expansion in mammalian genomes. All species are indicated by the initials of their biological names. (A) Box plots showing the length distribution of all $(GAA)_{6+}$ repeats for each species. Boxes are bound by the 25th and 75th percentiles (contained lines, when present, indicate 50th percentiles) and stems indicate 10th and 90th percentiles. A horizontal dotted line is plotted at the level of the 90th percentile for the human genome. (B) Relative expandability of GAA trinucleotide repeats as a proportion of all A-rich triplet repeats (GAA, TAA, CAA) for each species. This was calculated by measuring the proportion of the 90th percentile of $(GAA)_{6+}$ repeats versus that of all $(NAA)_{6+}$ repeats (see text). All six mammalian species (HS, PT, MM, RN, CF, MD) as a group mapped above the line drawn at $r = 1$. *Note:* The results for RN and MM overlap, as do those for AM and AT.

minimal difference (14.2% versus 8.4%, $\chi^2 = 4.66$, $p < 0.05$). This suggests that the evolutionarily recent transition to premutation length at these loci in the human lineage may have predisposed these loci to undergo frequent expansions.

Two of the loci exhibiting alleles with >100 GAA triplets are located within introns of predicted transcriptional units of unknown function (Fig. 5). One, located at 3q13, is in intron 3 of the *PLCXD2* gene and is widely expressed, including in the human brain (EST profile viewer). The other, located at 4q31.1, is in intron 1 of the *RNF150* gene, a ubiquitously expressed ring finger protein (EST profile viewer). It is noteworthy that for both genes we identified individuals who are homozygous for expanded GAA repeat alleles, i.e., similar to FRDA patients (asterisks in Fig. 5). Both of these GAA sequences are conserved in the chimpanzee genome, with flanking intron sequence identity of 99% (1000 bp upstream and downstream). However, transition to premutation length is seen only in human, with both chimpanzee orthologs showing <10 triplets (Fig. 3A).

Discussion

We along with others have previously noted that G/A-rich trinucleotide repeats are especially abundant in the human genome [1,14,25–27]. Our present data indicate that long GAA trinucleotide repeat tracts are specifically seen in mammalian genomes. Although the mechanism for this taxon-specific expansion remains unclear, it is likely that this may be because of their preferential association with the poly(A) tails of SINEs. However, the similar association of all NAA trinucleotide repeats with SINEs precludes a simple explanation given that CAA and TAA repeats, despite being far more prevalent, have not similarly expanded in mammals. Indeed, TAA repeats have expanded significantly in nonmam-

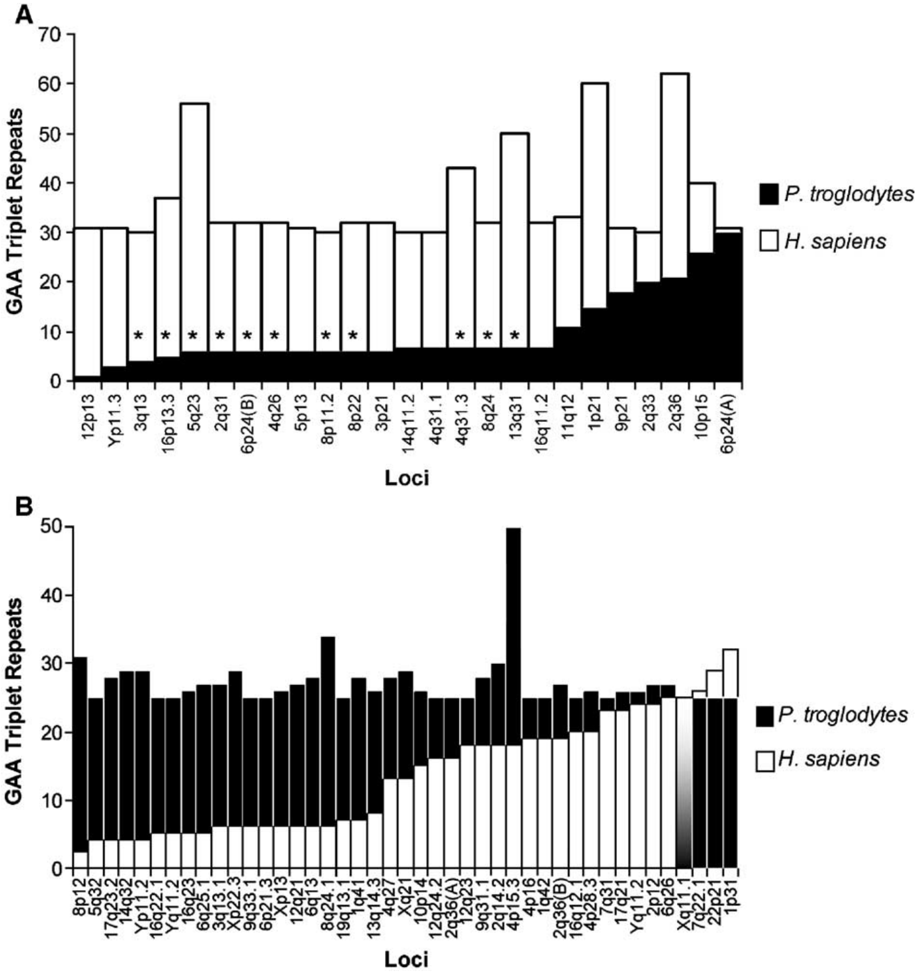

Fig. 3. Comparison of orthologous loci containing potential premutation GAA trinucleotide repeats in primates. In all cases chromosome map locations are for the human genome and the orthologs are arranged with increasing length from left to right. (A) Comparison of the length of repeat tracts at loci in the chimpanzee (*P. troglodytes*) genome that are orthologous to human genomic loci containing $(GAA)_{30+}$ sequences. Among chimpanzee orthologs of human $(GAA)_{30+}$ sequences, 19 of 24 contained <13 triplets. At 11 of these loci both chimpanzee and gorilla orthologs contained <13 triplets despite 30+triplets in the human genome (indicated by asterisks). (B) Comparison of the length of repeat tracts at loci in the human genome that are orthologous to chimpanzee genomic loci containing $(GAA)_{25+}$ sequences. Among the human orthologs, 22 of 43 loci contained <13 repeats where the chimpanzee genome contained 25+triplets. The gray bar indicates the same length in human and chimpanzee genomes.

malian vertebrates, arthropods, and *Dictyostelium.* The *Anopheles* genome is the only nonmammalian genome that showed significant expansion of GAA trinucleotide repeats. This may be because of the variety of non-LTR transposons, especially the Outcast and RTE clades that have GAA repeats in their 3′ tails, which also include the Ag-JAMMIN-1 family that is especially abundant [28]. An intriguing possibility could involve horizontal transfer from mammals given that *Anopheles* is an obligate mammalian parasite. The *Dictyostelium* genome has very long TAA repeats and short GAA repeats. This may be because its genome is 78% A/T, which in some intergenic regions even exceeds 95% [29]. Our data are consistent with those of Töth et al. [1] in that there was no correlation between the abundance and the expandability of repeats and in that rodent genomes have especially long repeats. Despite the reported higher density of GAA trinucleotide repeats in *A. thaliana, C. elegans*, and *S. cerevisiae* compared with human chromosomes 21 and 22 [30], we did not detect large expansions similar to the mammalian genomes.

Our results show that at least two of the three subclasses of mammals have relatively large GAA expansions: *M. domestica* represents the marsupials (Metatheria) and the rest represent placental mammals (Eutheria). Expanded GAA repeat tracts were seen in three separate orders of placental mammals: Primates (*H. sapiens* and *P. troglodytes*), Rodentia (*M. musculus* and *R. norvegicus*), and Carnivora (*C. familiaris*). Only partial genomic sequence was available from the platypus (*Ornithorhynchidus anatinus*), a representative of the third branch of mammals, the monotremes (Protheria), which also showed that GAA trinucleotides have undergone significant expansion (data not shown). Together, these data indicate that while expansions to premutation length have occurred even after divergence of the great apes, the relative expandability of GAA trinucleotides in comparison to all other triplet repeats is mammalian-specific and most likely dates back to early mammalian evolution.

The disparity between the locations of long (potential premutation) GAA trinucleotide repeats between the human

Table 1
Chimpanzee orthologs of human $(GAA)_{30+}$ loci that map within old *Alu* elements[a]

Locus[b]	*Homo sapiens* $(GAA)_{30+}$	Number of repeats	*Pan troglodytes*	Number of repeats
1p21	AluSp	60	AluSp	15
2q31	AluJb	32	AluJb	6
4q26	AluJb	32	AluJb	6
4q31.1	AluSx	30	AluSx	7
4q31.3	AluSp	43	AluSp	7
8p11.2	AluSq	30	AluSq	6
8p22	AluSx	32	AluSx	6
10p15	AluJb	40	AluJo/FRAM	26
11q12	FRAM	33	FRAM	8
14q11.2	AluSg	30	AluSg	7
16p13.3	AluSq	30	AluSq	5
16q11.2	AluSx	32	AluSx	7
Xp11.2	AluSp	50	AluSp	7
Xq24	AluSp	36	AluSp	7
Median[c]		32		7

[a] Including those loci in Fig. 3A that also map in *Alu* J, *Alu* S, and monomeric units (FRAM, FLAM).

[b] Human genomic locations.

[c] $p < 0.001$, Mann–Whitney *U* test.

and the chimpanzee genome is intriguing. There is a high level of sequence conservation (>95%) between the two species, similar overall size distribution of GAA repeats, and similarity in the mapping within tails of *Alu* elements [14]. Our data also indicate that the species-specific expansions occurred after the split from the common ancestor. All of these observations indicate that factors other than the sequence per se (or its ability to adopt non-B DNA structures) may play a role in the expansion process. Although apparently stochastic, we cannot rule out genomic locus-specific factors within the human and chimpanzee chromosomes. It should be noted that we used the $(GAA)_{25+}$ threshold length for the chimpanzee genome since there were fewer $(GAA)_{30+}$ sequences compared with humans. Although this may not be ideal for the study of hyperexpansions of premutations per se, it nevertheless serves as a useful way to analyze the evolutionary expansion of GAA repeats from short tracts to those that are potentially of premutation length.

We detected considerable allelic variability at all $(GAA)_{30+}$ sequences in the human genome. At least nine loci showed alleles with >65 triplets and at five of these loci repeat lengths of >100 triplets were detected. These expansions are analogous to disease-causing mutations at the *FXN* locus and show that, while there are locus-specific differences in the amount of variability, large expansions occur at multiple genomic locations. It should be noted that we have tested only the $(GAA)_{30+}$ sequences as per the published human genome sequence and that similar expansions may also occur at other locations, for example at any of the 297 $(GAA)_{20+}$ sequences.

The occurrence of expanded GAA repeats in apparently normal individuals at several loci is a potentially important source of functional genomic variation. Indeed, two of the loci showing $(GAA)_{100+}$ alleles were found within transcriptional units of unknown function. Both of these transcripts are widely expressed, including in the human brain, as has been noted for disease genes associated with trinucleotide repeat expansions. We and others have shown that the location of expanded GAA trinucleotide repeat tracts within a transcriptional unit results in transcriptional interference [15,16,19]. Therefore, it is likely that the $(GAA)_{100+}$ alleles of *RNF150* and *PLCXD2* would affect their transcript levels, with potential phenotypic consequences in individuals homozygous for such alleles (see asterisks in Fig. 5). However, it is not possible to comment on any phenotypic effects in the individuals we identified since they were derived from the anonymous CEPH polymorphism discovery panel.

Long GAA repeat alleles are known to be unstable in somatic cells [24,31,32] and following germ-line transmission [33,34]. Such a frequency of long GAA alleles among humans therefore represents an important source of interindividual and intergenerational variability. We were unable to test for similar variability among other mammalian species, but some differences in the experimentally derived sequence of chimpanzee loci versus the published chimpanzee genome sequence were noted (data not shown).

Moreover, long GAA repeat tracts are known to affect several aspects of genome function. For example, long GAA

Table 2
Human orthologs of chimpanzee $(GAA)_{25+}$ loci that map within old *Alu* elements[a]

Locus[b]	*Homo sapiens*	Number of repeats	*Pan troglodytes* $(GAA)_{25+}$	Number of repeats
1q41	AluSx	7	AluSx	28
2q36(A)	AluSx	16	AluSx	25
3q13.1	AluSx	6	AluSx	27
4p15.3	AluSq	18	AluSq	50
5q32	AluSg	4	AluSg	25
6q13	AluJb	6	AluJb	28
6q25.1	AluJo	5	AluJo	27
6q26	AluJo	25	AluJo	27
7q22.1	AluJo	26	AluJo	25
8p12	AluSx	2	AluSx	31
9q31.1	AluSx	18	AluSx	28
9q33.1	AluSx	6	AluSx	25
10p14	AluSp	15	AluSp	26
12q24.2	FLAM_C and FRAM	16	*Alu* J[c]	25
14q32	AluJo	4	AluJo	29
16q12.1	AluSq	20	AluSq	25
16q22.1	AluSx	5	AluSx	25
17q21	AluJb	23	AluJb	26
17q23.2	AluSg	4	AluSg	28
19q13.1	AluSx	7	AluSx	25
22p21	AluSx	29	AluSx	25
Xp13	AluSx	6	AluSx	26
Xp22.3	AluJo	6	AluJo	29
Xq11.1	AluSx	25	AluSx	25
7q31	AluY	23	AluY	25
Median[d]		7		26

[a] Including those loci in Fig. 3B that also map in *Alu* J, *Alu* S, and monomeric units (FRAM, FLAM).

[b] Human genomic locations.

[c] Partial *Alu* J element.

[d] $p < 0.001$, Mann–Whitney *U* test.

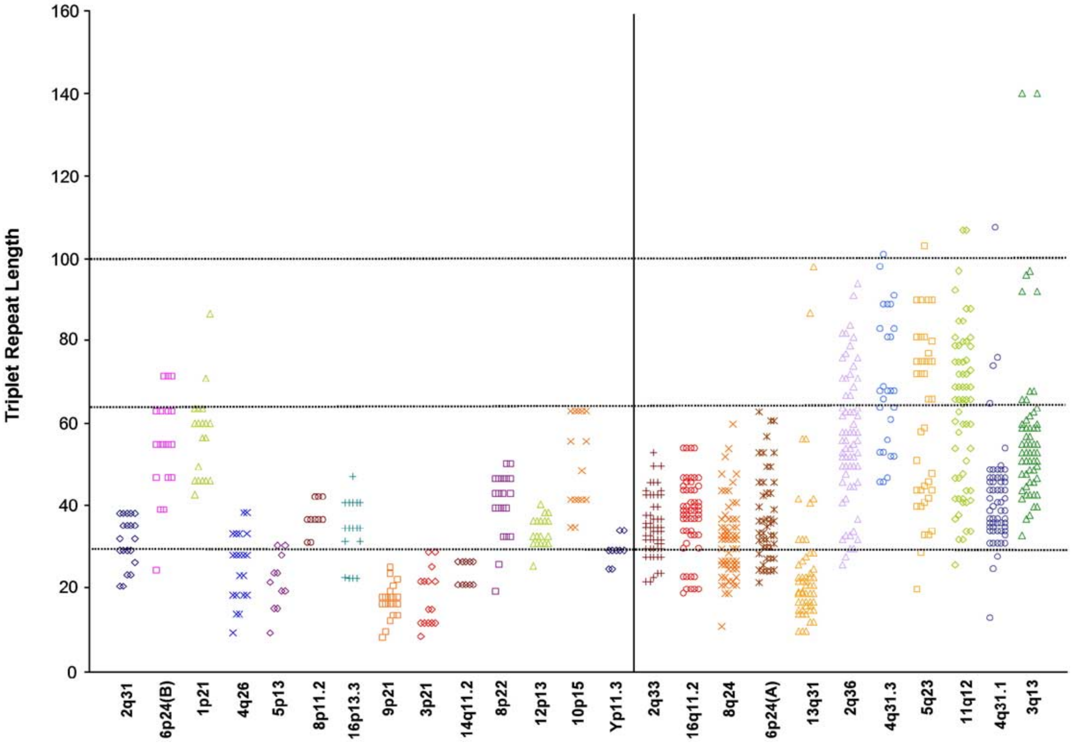

Fig. 4. Polymorphic variability of all 25 confirmed $(GAA)_{30+}$ sequences in the human genome showing large expansions at several loci. Each symbol in the graph represents a single allele, sizes of which are indicated (in triplets) on the *y* axis. Horizontal dotted lines are drawn at the levels of 30, 65, and 100 triplets. The 11 loci to the right of the vertical line were analyzed in a larger number of chromosomes (see text) and they are arranged approximately with increasing variability from left to right. As expected, almost all loci showed at least some alleles with >30 triplets. Nine loci showed at least some alleles containing >65 triplets and 5 showed alleles with >100 triplets. Loci are indicated by their chromosome map locations (A and B symbols are used arbitrarily to differentiate distinct loci with the same cytogenetic map location). Primer sequences used to amplify all 25 loci are in Supplementary Table 1.

repeats can adopt non-B DNA structures [15–21,35], interfere with gene transcription [15,16,18,19] and DNA replication [16,21,22], and mediate gene silencing via position effect variegation [23] and are associated with flanking mutagenesis [24]. Therefore, the abundance of long and polymorphic GAA repeat tracts in mammals may have important implications for local genomic structure and function.

A noteworthy caveat of our study is that "complete" genome sequences are not necessarily complete, especially for repeat sequences. Therefore, it is likely that at least some of the differences observed among the various genomes analyzed could stem from variable efficiencies of sequencing repeat-containing regions. However, our observation of long GAA tracts in all mammalian genomes mitigates some of this

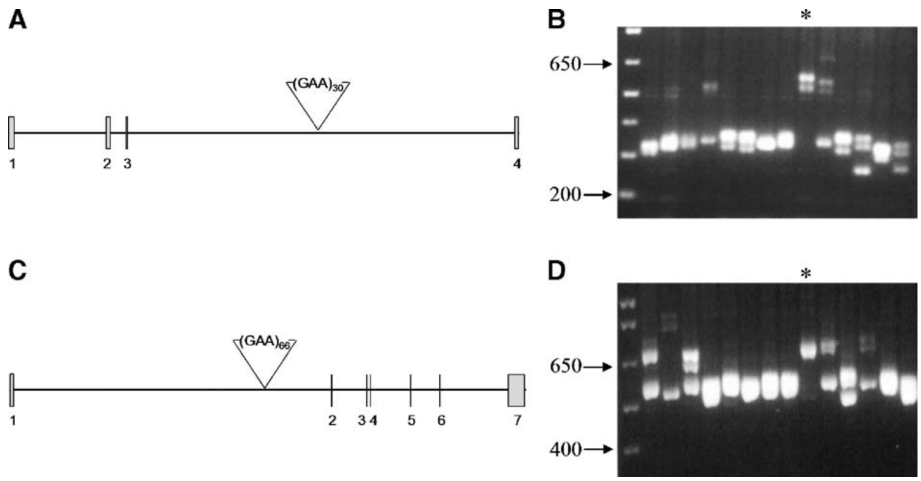

Fig. 5. Loci with $(GAA)_{100+}$ alleles that are located within introns of transcriptional units. Diagrammatic representation of the transcriptional units [(A) *PLCXD2* located at 3q13, (C) gene for ring finger protein *RNF150* located at 4q31.1] showing the relative positions of the GAA sequence. (B, D) Representative gels showing results of PCR amplification of human controls (derived from the CEPH polymorphism discovery panel). Asterisks indicate individuals who are homozygous for expanded alleles.

concern and supports the notion that expanded GAA repeats are a unique property of mammalian genomes.

In conclusion, we identified long GAA trinucleotide repeats specifically in mammalian genomes. Several loci with GAA trinucleotide repeats that have expanded to potentially premutagenic size in primates seem to have expanded in a lineage-specific manner after the split from the common ancestor, and these represent an important source of genomic sequence variability among primates. Several loci in the human genome exhibit significant polymorphic variability, including the presence of expanded GAA repeat alleles analogous to *FXN* gene mutations.

Methods

Genome sequence accession

Most of the complete genome sequences were downloaded from Ensembl (ftp://ftp.ensembl.org/pub/), NCBI (ftp://ftp.ncbi.nih.gov/genomes/), and UCSC (ftp://hgdownload.cse.ucsc.edu/apache/htdocs/goldenPath/currentGenomes/). The following genomic versions/builds were analyzed (versions, dates, and sizes of genomes scanned in the present study are indicated in parentheses): (A) from Ensembl, *P. troglodytes* (build 1v1; 11/03; 2.73 Gb), *A. gambiae* (v2b; 278 Mb), and *C. elegans* (23.116a.1; 100 Mb); (B) from NCBI, *H. sapiens* (v34a, 07/03 for all, except v30, 08/02 for data in Fig. 3A and Table 1; 2.84 Gb), *M. musculus* (v32; 10/03; 2.47 Gb), *R. norvegicus* (v2.1; 01/03; 2.66 Gb), *A. mellifera* (Amel_1.2; 197 Mb), *P. falciparum* (3D7; 23 Mb), *A. thaliana* (1v1; 115 Mb), *C. galbrata* (CBS138; 12 Mb), *S. cerevisiae* (10/03; 12 Mb), and *S. pombe* (07/04; 14 Mb); (C) from UCSC, *C. familiaris* (v1.0; 07/04; 2.4 Gb), *G. gallus* (02/04; 1.05 Gb), *D. rerio* (2v3; 11/03; 1.46 Gb), *D. melanogaster* (release 3; 117 Mb), and *D. yakuba* (release 1; 04/04; 167 Mb); (D) *C. albicans* (16 Mb) was obtained from the Stanford Genome Technology Center (http://www-sequence.stanford.edu/group/candida; courtesy of NIDR and the Burroughs Wellcome Fund); (E) *D. discoideum* (34 Mb) was downloaded from dictyBase (dictybase.org; courtesy of The Wellcome Trust Sanger Institute, Baylor College of Medicine, University of Cologne, Department of Genome Analysis in Jena of the Institute of Molecular Biotechnology and Institut Pasteur); (F) from Broad Institute, *M. domestica* (monDom1; 3.5 Gb).

Sequence analysis

A custom program in C was used to identify all 10 nonredundant triplet motifs as described under Results. All potential premutations [$(GAA)_{30+}$ in human and $(GAA)_{25+}$ in chimpanzee] along with flanking sequence were thus identified. These were extracted and used manually to find orthologous sequences using BLAT (UCSC) and BLAST (NCBI). Both search methods had the RepeatMasker filter enabled to prevent spurious matches. The extracted sequences included 300–1000 bp flanking sequence, depending on how repetitive the regions were. When multiple hits were obtained, the highest scoring hits were used for further analysis. Orthologous pairs were determined as the best reciprocal hits. Sequence homology/identity was estimated using the Local Alignment tool of Michigan Technological University (http://www.genome.cs.mtu.edu/align/align.html). ClustalW multiple sequence alignment (http://www.ebi.ac.uk/clustalw/) and Boxshade (http://www.ch.embnet.org/software/BOX_form.html) allowed for visual discernment of homology. RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) was used to detect *Alu* subfamilies (RepeatMasker v3.0.2; RepBase 04/2004). Genomic locations of *Alu* elements were also obtained from UCSC (ftp://hgdownload.cse.ucsc.edu/apache/htdocs/goldenPath/).

PCR analysis of GAA trinucleotide repeats

Flanking primers were designed to amplify all human $(GAA)_{30+}$ sequences such that they would not overlap with *Alu* sequences (see Supplementary Table 1 for all confirmed loci). The presence of repeat tracts was confirmed by direct sequencing of PCR products. Orthologous chimpanzee and gorilla loci were amplified with human-specific primers, using annealing temperatures below that used for human genomic targets. Lymphoblastoid DNA from control human subjects (CEPH polymorphism discovery panel), genomic DNA from chimpanzee (*P. troglodytes;* NGO6939) and gorilla (*Gorilla gorilla*; NG05251) were obtained from Coriell Cell Repository.

Acknowledgments

This research was supported by grants from the NIH (NS047596), MDA, and NSF-EPSCoR to S.I.B. G.L.D. is a recipient of a postdoctoral fellowship from the American Heart Association (Heartland Affiliate).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2005.09.006.

References

[1] G. Töth, Z. Gáspári, J. Jurka, Microsatellites in different eukaryotic genomes: survey and analysis, Genome Res. 10 (2000) 967–981.

[2] C.J. Cummings, H.Y. Zoghbi, Fourteen and counting: unraveling trinucleotide repeat diseases, Hum. Mol. Genet. 9 (2000) 909–916.

[3] R.P. Bowater, R.D. Wells, The intrinsically unstable life of DNA triplet repeats associated with human hereditary disorders, Prog. Nucleic Acid Res. Mol. Biol. 66 (2001) 159–202.

[4] V. Campuzano, et al., Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion, Science 271 (1996) 1423–1427.

[5] M. Cossée, et al., Evolution of the Friedreich's ataxia trinucleotide repeat expansion: founder effect and premutations, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 7452–7457.

[6] L. Montermini, et al., The Friedreich ataxia GAA triplet repeat: premutation and normal alleles, Hum. Mol. Genet. 6 (1997) 1261–1266.

[7] C. Epplen, et al., Differential stability of the (GAA)n tract in the Friedreich ataxia (STM7) gene, Hum. Genet. 99 (1997) 834–836.

[8] A. Dürr, et al., Clinical and genetic abnormalities in patients with Friedreich's ataxia, N. Engl. J. Med. 335 (1996) 1169–1175.

[9] M.B. Delatycki, et al., Sperm DNA analysis in a Friedreich ataxia premutation carrier suggests both meiotic and mitotic expansion in the FRDA gene, J. Med. Genet. 35 (1998) 713–716.

[10] P. González-Cabo, et al., Incipient GAA repeats in the primate Friedreich ataxia homologous genes, Mol. Biol. Evol. 16 (1999) 880–883.

[11] M. Labuda, et al., Unique origin and specific ethnic distribution of the Friedreich ataxia GAA expansion, Neurology 54 (2000) 2322–2324.

[12] C.M. Justice, et al., Phylogenetic analysis of the Friedreich ataxia GAA trinucleotide repeat, J. Mol. Evol. 52 (2001) 232–238.

[13] M. Gómez, R.M. Clark, S.K. Nath, S. Bhatti, R. Sharma, E. Alonzo, A. Rasmussen, S.I. Bidichandani, Genetic admixture of European *FRDA* genes is the cause of Friedreich ataxia in the Mexican population, Genomics 84 (2004) 779–784.

[14] R.M. Clark, G.L. Dalgliesh, D. Endres, M. Gómez, J. Taylor, S.I. Bidichandani, Expansion of GAA triplet repeats in the human genome: unique origin of the *FRDA* mutation at the center of an Alu, Genomics 83 (2004) 373–383.

[15] S.I. Bidichandani, T. Ashizawa, P.I. Patel, The GAA triplet-repeat expansion in Friedreich ataxia interferes with transcription and may be associated with an unusual DNA structure, Am. J. Hum. Genet. 62 (1998) 111–121.

[16] K. Ohshima, L. Montermini, R.D. Wells, M. Pandolfo, Inhibitory effects of expanded GAA·TTC triplet repeats from intron I of the Friedreich ataxia gene on transcription and replication in vivo, J. Biol. Chem. 273 (1998) 14588–14595.

[17] R.D. Wells, D.A. Collier, J.C. Hanvey, M. Shimizu, F. Wohlrab, The chemistry and biology of unusual DNA structures adopted by oligopurine·oligopyrimidine sequences, FASEB J. 2 (1998) 2939–2949.

[18] N. Sakamoto, K. Ohshima, L. Montermini, M. Pandolfo, R.D. Wells, Sticky DNA, a self-associated complex formed at long GAA*TTC repeats in intron 1 of the frataxin gene, inhibits transcription, J. Biol. Chem. 276 (2001) 27171–27177.

[19] E. Grabczyk, K. Usdin, The GAA*TTC triplet repeat expanded in Friedreich's ataxia impedes transcription elongation by T7 RNA polymerase in a length and supercoil dependent manner, Nucleic Acids Res. 28 (2000) 2815–2822.

[20] V.N. Potaman, et al., Length-dependent structure formation in Friedreich ataxia (GAA)n*(TTC)n repeats at neutral pH, Nucleic Acids Res. 32 (2004) 1224–1231.

[21] A.M. Gacy, et al., GAA instability in Friedreich's ataxia shares a common, DNA-directed and intraallelic mechanism with other trinucleotide diseases, Mol. Cell 1 (1998) 583–593.

[22] L.M. Pollard, et al., Replication mediated instability of the GAA triplet repeat mutation in Friedreich ataxia, Nucleic Acids Res. 32 (2004) 5962–5971.

[23] A. Saveliev, C. Everett, T. Sharpe, Z. Webster, R. Festenstein, DNA triplet repeats mediate heterochromatin-protein-1-sensitive variegated gene silencing, Nature 422 (2004) 909–913.

[24] S.I. Bidichandani, et al., Somatic sequence variation at the Friedreich ataxia locus includes complete contraction of the expanded GAA triplet repeat, significant length variation in serially passaged lymphoblasts and enhanced mutagenesis in the flanking sequence, Hum. Mol. Genet. 8 (1999) 2425–2436.

[25] P. Astolfi, D. Bellizzi, V. Sgaramella, Frequency and coverage of trinucleotide repeats in eukaryotes, Gene 317 (2003) 117–125.

[26] S. Subramanian, R.K. Mishra, L. Singh, Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions, Genome Biol. 4 (2003) R13.1–R13.9.

[27] S. Subramanian, et al., Triplet repeats in human genome: distribution and their association with genes and other genomic regions, Bioinformatics 19 (2003) 549–552.

[28] J. Biedler, Z. Tu, Retrotransposons in the African malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity, Mol. Biol. Evol. 20 (2003) 1811–1825.

[29] G. Glöckner, et al., The complex repeats of *Dictyostelium discoideum*, Genome Res. 11 (2001) 585–594.

[30] M.V. Katti, P.K. Ranjekar, V.S. Gupta, Differential distribution of simple sequence repeats in eukaryotic genome sequences, Mol. Biol. Evol. 18 (2001) 1161–1167.

[31] R. Sharma, S. Bhatti, M. Gümez, R.M. Clark, C. Murray, T. Ashizawa, S.I. Bidichandani, The GAA triplet-repeat sequence in Friedreich ataxia shows a high level of somatic instability in vivo, with a significant predilection for large contractions, Hum. Mol. Genet. 11 (2002) 2175–2187.

[32] R. Sharma, M. Gómez, I. De Biase, T. Ashizawa, S.I. Bidichandani, Friedreich ataxia in carriers of somatically unstable borderline GAA repeat alleles, Ann. Neurol. 56 (2004) 898–901.

[33] E. Monrós, et al., Phenotype correlation and intergenerational dynamics of the Friedreich ataxia GAA trinucleotide repeat, Am. J. Hum. Genet. 61 (1997) 101–110.

[34] G. De Michele, et al., Parental gender, age at birth and expansion length influence GAA repeat intergenerational instability in the X25 gene: pedigree studies and analysis of sperm from patients with Friedreich's ataxia, Hum. Mol. Genet. 7 (1998) 1901–1906.

[35] N. Sakamoto, et al., Sticky DNA: self-association properties of long GAA·TTC repeats in R·R·Y triplex structures from Friedreich's ataxia, Mol. Cell 3 (1999) 465–475.