Tandem repeats in human disorders: mechanisms and evolution

Pratibha Siwach1, Subramaniam Ganesh1

1Department of Biological Sciences and Bioengineering, Indian Institute of Technology, Kanpur, India

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. TRS expansions and their biology in relation to disorders
   3.1. Functional significance of TRS in non-coding regions of genes
      3.1.1. Disease associated TRS in the 5'-flanking regions of genes
      3.1.2. Disease associated TRS in the intronic regions of genes
      3.1.3. Disease associated TRS in the 3'-untranslated regions of genes
      3.1.4. Disease associated TRS in a non-coding RNA
   3.2. Functional significance of TRS in the coding regions of genes
      3.2.1. TRS in the coding region: polyglutamine disorders
         3.2.1.1. The role of PolyQ aggregates in neuropathology
         3.2.1.2. The role of PolyQ expansion in transcriptional dysregulation
      3.2.2. TRS in the coding region: polyalanine disorders
      3.2.3. TRS in the coding region disorders: other amino acid repeats
4. Factors modulating TRS instability
   4.1. Replication and TRS expansion
   4.2. Post-replication repair and TRS expansion
   4.3. Recombination and TRS instability
   4.4. TRS in the coding region and codon bias
   4.5. Parent-of-origin effects in TRS expansion
   4.6. Role of evolution in TRS expansion
5. Perspective and concluding remarks
6. Acknowledgements
7. References

1. ABSTRACT

One of the most compelling reasons for the study of repetitive DNA sequence in the human genome has been the instability of simple repeat sequences associating with a growing and an interesting group of disorders affecting the neurological, neuromuscular or developmental processes. As a result, the molecular processes that underlie this unique form of mutation and the pathological pathways that lead to the disorders are being uncovered rapidly and are being intensively investigated. Genes with expanded repeats exhibit either loss-of-function or gain-of-function effect at the protein and/or RNA level. In this review, we aim to provide an overview of the recent advances in molecular pathology of disorders associated with heritable changes in the length of the repeat sequences, and examine how dynamism in these repeats is regulated.

2. INTRODUCTION

The three billion base pairs of our genome consist of a combination of four nucleotide bases in a random but functionally conserved arrangement. In spite of this randomness, many nucleotide sequences occur repeatedly to constitute a tandemly repeated signature motif. Recent analysis of the human genome sequence advocates that roughly 95% of genome is made up of tandem repeat sequences (TRS) (1-2). The TRS consists of consecutive and head-to-tail copies of sequence and posses varied levels of compositional complexity ranging from a few bases to more than hundred or so. For example, surveys based on database analysis showed that A/T repeats are maximal among mononucleotides, CA/GT as dinucleotide repeats, and that the GC rich repeats are overrepresented in tri-, and tetra-nucleotides (2-3). TRS are spread more or less evenly
Tandem repeats in human disorders

throughout the entire human genome, as they are found to be located in the coding, non-coding and regulatory regions of the genes (3-4).

Numerous evidences demonstrate that TRS are significantly important in genomic organization because of their effect on various cellular activities like chromatin organization, regulation of gene expression, replication, recombination, mismatch repair systems, and monitoring of cell cycle (5). TRS are dynamic in nature as they are quite prone to undergo repeat length variations (6). Although the mechanisms that generate and maintain TRS in genomes are poorly understood, available evidences suggest that, due to the reiterated nucleotide sequences, processes coupled with replication and recombination could lead to their expansion and/or contraction (7-8). Perhaps the most compelling reason for studying TRS is their involvement in plethoras of human diseases. Though length variation (expansion or contraction) in TRS largely remain neutral because of their location in nongenic regions, there are sufficient examples of their involvement in disturbing gene functions and regulations (4). Particularly, TRS lying in or close to gene sequences are more prone to perturb normal gene function by altering their activity. Indeed, TRS are known to be associated both with heritable (germ-line mutations) and non-heritable (somatic mutations) disorders in humans (4). More often TRS manifest “dynamism” in heritable disorder wherein the probability of repeat expansion is directly proportional to number of repeating units within the TRS (7, 9-11). In these disorders, therefore, symptoms appear at an early age when the expanded TRS is passed on to the next generation (7, 10-11). In this review, we focus on progress made in understanding the significance of TRS in human disorders. We discuss the molecular mechanisms of genome instability associated with TRS with the hope of providing insights into the mechanisms and the evolution of this novel group of disorders.

3. TRS EXPANSION AND THEIR BIOLOGY IN RELATION TO DISORDER

3.1. Functional significance of TRS in non-coding regions of genes

There is growing number of evidences that suggest critical roles for TRS in gene control and regulation. For example, removal of CA repeats normally present in the 3'-untranslated region (UTR) of BCL2 gene led to stabilization of BCL2 mRNA abundance (4), and removal of CAG repeat stretch in the 5'-UTR of the calmodulin-1 gene reduced its expression by 45% (12). Repeat tracts present in the 3'-UTR perhaps influence various aspect of mRNA metabolism by serving as binding sites for RNA binding proteins (13). Curiously, a majority of transcripts having GATA repeats in the 3'-UTR were found to encode membrane proteins, indicating that such repeats could serve as signals for translational regulation of genes involved in membrane functions (14). Intronic TRS are also known to regulate gene expression. For example, a tetranucleotide repeat located in the first intron of the gene encoding tyrosine hydroxylase in human is shown to function as a transcriptional enhancer, suggesting that TRS in introns might have a direct influence on gene expression (15-16). Similar to TRS present in the intergenic regions, the TRS lying in the genic region also show a higher mutation rate as compared to regions that do not harbor repetitive sequences (17). Considering the essential role that such repeat elements play in gene function, elongation/shortening events in TRS are expected to affect the normal gene function which eventually could lead to a disorder (Figure 1). At least 10 genetic disorders are known in humans in which the primary cause of the defect is the expansion of TRS located in the non-coding region of corresponding genes (Table 1).

3.1.1. Disease associated TRS in the 5'-flanking regions of genes

Three neurological disorders are associated with expansion of TRS located in the 5'-flanking region that lead to the silencing of corresponding gene functions (Table 1). These include progressive myoclonic epilepsy of Unverricht and Lundborg (EPM1) associated with mutations in CSTB gene, the fragile X mental retardation syndrome associated with mutations in FMR1 gene (a common form known as FMR), or the FMR2 gene (a milder syndrome known as FRAXE) (Table 1). The CSTB gene encodes cystatin B protein, a cysteine protease inhibitor, which is thought to play a role in protecting against the proteases leaking from lysosomes (18). Expansion of a dodecamer repeat in the 5'-flanking region of CSTB gene is the common mutation mechanism in EPM1 (19). This expansion is associated with a marked reduction of CSTB mRNA and subsequent loss of cystatin B activity, leading to the onset of EPM1 (20). Among the most probable explanation for the reduced transcription of CSTB gene include hypermethylation of the CpG elements, altered spacing of promoter elements, altered chromatin structure, and recruitment of transcriptional repressors on the dodecameric repeat (21). Similar observations were made for the genes FMR1 and FMR2, except that the trinucleotide repeat sequences in these genes fall within the transcribed region (5'-UTR) (Table 1). Expression of both FMR1 and FMR2 genes was shown to be downregulated by repeat expansion and methylation (22-24). FMR1 gene encodes an RNA binding protein that associates with polyribosome and regulates the process of translation (25-26). Thus, translational dysregulation of mRNAs normally associated with FMR1 protein was thought be the primary cause of FMR syndrome (27). FMR2 encodes a DNA binding nuclear protein, and could perhaps function as a transcription factor (28). Loss-of-function of corresponding genes in disorders EPM1 and FMR is the likely to be the primary cause for the resulting phenotype as mutations other than repeat expansion are also reported for CSTB and FMR1 genes (29-31). TRS in FMR1 and FMR2 genes represent rare folate-sensitive fragile sites (known as FRAXA and FRAXE, respectively) that are especially prone to forming gaps on metaphase chromosomes when cells are cultured in vitro under conditions that block DNA replication (32). While chromosomal breaks in FRAXA or FRAXE loci have not yet been reported in FMR patients, another fragile site, FRA11B located at 11q23.3 and harboring CCG-repeats, was shown to associate with
Figure 1. Mechanism of pathogenesis associated with expansion of non-coding TRS. Loss-of-function effect: In EPM1, SIX5, FMR and FRA1XE, then unstable TRS element is localized either the promoter or the 5'-UTR (Table 1), and expansion of this TRS leads to reduced transcription of the respective gene. The molecular mechanism of gene-silencing could be hypermethylation of the CpG elements, altered spacing of promoter elements, altered chromatin structure, and/or recruitment of transcriptional repressors. Thus, loss-of-function of corresponding genes is the likely to be the primary cause for the resulting phenotype. Gain-of-function effect: In DM1, DM2, and HDL2, the unstable TRS fall in the transcribed part of the gene (Table 1). Transcripts bearing the expanded TRS form distinct RNA foci in the nucleus, and affect the RNA splicing processes of several downstream genes. Thus, the TRS in these disorders could have a ‘trans-dominant’ effect on a select group of genes and this may explain their dominant mode of inheritance (Table 1). The intronic TRS of ATX10 gene is believed to have the same effect although yet to be supported by experimental evidences and therefore identified with a question mark in the figure. TRS, location of tandem repeat sequence, cds, coding sequence.

3.1.2. Disease associated TRS in the intronic regions of genes

Instability in TRS located in the intronic regions of genes is linked with three distinct forms of hereditary ataxias and two form of myotonic dystrophies (Table 1). The clinical phenotypes of ataxias are characterized by predominantly cerebellar symptoms such as incoordination and unsteady gait (38). A majority of cases with one of the major forms of ataxias, known as Friedreich ataxia, are due to expansion of a GAA trinucleotide repeat in intron 1 of the FXN gene (39-40). This nuclear gene encoded mitochondrial protein, named frataxin, was shown to regulate mitochondrial iron transport and respiration (41). The expanded alleles of FXN gene produce lesser levels of mature FXN mRNA than the alleles in the normal range (42). The reduction in the FXN messenger level could be due to the repeat mediated stalling of RNA polymerase (43) or due to the repeat mediated epigenetic changes in the chromatin flanking the FXN gene (44). In either case, expansion of GAA repeats would result in frataxin deficiency, leading to impaired iron metabolism, oxidative damage, and progressive iron accumulation (41). The second form of ataxia (SCA10) is associated with an expansion of a pentanucleotide (ATTCT) repeat in intron 9 of the ATXN10 gene (45). Unlike FXN, however, the expansion of ATTCT repeat did not alter the normal level of ATX10 transcripts in the affected (46). However, in mice, homozygous null mutants for ATX10 gene resulted in embryonic lethality and heterozygotes were asymptomatic (46). SCA10 is being a dominant disorder in humans, a simple loss of function or haploinsufficiency of ATXN10 is unlikely to be the cause for the disease phenotype (46). Thus the precise mechanism by which the intronic repeats exert a dominant role in the etiology of SCA10 is yet to be resolved.
Table 1. Molecular and genetic features of disease-associated unstable repeats falling in the non-coding regions of genes

<table>
<thead>
<tr>
<th>Diseases</th>
<th>Gene implicated</th>
<th>Chromosomal location of the gene</th>
<th>Affected protein</th>
<th>Location of the repeat in the gene</th>
<th>Repeat unit</th>
<th>Number of repeats: wild-type (disease-associated)</th>
<th>Mode of inheritance1</th>
<th>Parental gender bias1</th>
<th>Effect of mutation1</th>
<th>OMIM*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Progressive myotonic epimyopathy</td>
<td>CSTB</td>
<td>21q22.3</td>
<td>Cystatin B</td>
<td>Promoter</td>
<td>CCCC</td>
<td>CTG, COCG</td>
<td>&gt;12 (&lt;17)</td>
<td>AR, NR</td>
<td>LOF</td>
<td>601145</td>
</tr>
<tr>
<td>Spinal-cerebellar ataxia type-12</td>
<td>PTPRZB</td>
<td>5q13-5q32</td>
<td>Phosphatase 2</td>
<td>Regulatory subunit B</td>
<td>UAG</td>
<td>&lt;28 (&lt;60)</td>
<td>AD, NR</td>
<td>M, LOF</td>
<td>LOF</td>
<td>604325</td>
</tr>
<tr>
<td>Oxidative proteinopathy associated with DMD</td>
<td>SIX3</td>
<td>19q13.32</td>
<td>DM fasciculostabilis homeodomain protein</td>
<td>Promoter</td>
<td>CTG</td>
<td>&lt;55 (&lt;90)</td>
<td>AD, M</td>
<td>M, LOF</td>
<td>600963</td>
<td></td>
</tr>
<tr>
<td>Fragile X syndrome</td>
<td>FSMB1</td>
<td>Xq27.3</td>
<td>FMRP</td>
<td>5'-UTR</td>
<td>CGG</td>
<td>&lt;90 (&lt;200)</td>
<td>XD</td>
<td>M, LOF</td>
<td>309150</td>
<td></td>
</tr>
<tr>
<td>Fragile X; mental retardation</td>
<td>FSMB2</td>
<td>Xq28</td>
<td>FMR2 protein</td>
<td>5'-UTR</td>
<td>CGG</td>
<td>&lt;50 (&lt;300)</td>
<td>NR</td>
<td>M, LOF</td>
<td>309548</td>
<td></td>
</tr>
<tr>
<td>Jacobsen's syndrome</td>
<td>CBL2</td>
<td>11q23.3</td>
<td>Adapter for RPTK</td>
<td>5'-UTR</td>
<td>CUG</td>
<td>&lt;11 (&lt;100)</td>
<td>NR</td>
<td>P, Chromosomal deletion</td>
<td>653160</td>
<td></td>
</tr>
<tr>
<td>Spinocerebellar ataxia type 10</td>
<td>ATXN10</td>
<td>22q13.3</td>
<td>Autosomal-10</td>
<td>Intron 9</td>
<td>ATCT</td>
<td>&lt;1000 (&lt;1000)</td>
<td>AD, P</td>
<td>NR</td>
<td>603110</td>
<td></td>
</tr>
<tr>
<td>Myotonic dystrophy 2</td>
<td>ZNF9</td>
<td>3q21.3</td>
<td>RNA-binding protein</td>
<td>Intron 1</td>
<td>CUG</td>
<td>&lt;75 (&lt;1110)</td>
<td>AD</td>
<td>NR, GOF</td>
<td>116955</td>
<td></td>
</tr>
<tr>
<td>Friedrich's ataxia</td>
<td>FXN</td>
<td>9q35-21.1</td>
<td>Fmratin</td>
<td>Intron 1</td>
<td>GAA</td>
<td>&lt;90 (&lt;100)</td>
<td>AR, M</td>
<td>LOF</td>
<td>600829</td>
<td></td>
</tr>
<tr>
<td>Myotonic dystrophy 1</td>
<td>DMPK</td>
<td>19q13</td>
<td>MD protein kinase</td>
<td>5'-UTR</td>
<td>CTG</td>
<td>&lt;55 (&lt;90)</td>
<td>AD</td>
<td>M, GOF</td>
<td>603177</td>
<td></td>
</tr>
<tr>
<td>Spinal-cerebellar ataxia type-8</td>
<td>ATXN8OS</td>
<td>13q21</td>
<td>Non-coding RNA</td>
<td>Transcribed sequence</td>
<td>CCG</td>
<td>&lt;54 (&lt;127)</td>
<td>AD</td>
<td>M, GOF</td>
<td>603166</td>
<td></td>
</tr>
<tr>
<td>Huntington disease like-2</td>
<td>JPH3</td>
<td>16q24-25</td>
<td>Juxtaphilin-3</td>
<td>5'-UTR</td>
<td>CTG</td>
<td>&lt;28 (&lt;90)</td>
<td>AD, NR</td>
<td>NR</td>
<td>603260</td>
<td></td>
</tr>
</tbody>
</table>

1Abbreviations used: AR, autosomal dominant; AR, autosomal recessive; XD, X-linked dominant; XR, X-linked recessive; NR, not reported; M, maternal; P, paternal; LOF, loss-of-function; GOF, gain-of-function

Myotonic dystrophy (DM) is a multisystem disorder and also the common form of muscular dystrophy affecting the adults. One form of DM is caused by CTG repeat expansion in intron 1 of the ZNF9 gene and the resulting phenotype is referred to as DM2 (47). The ZNF9 gene encodes an RNA-binding protein with zinc finger domains and was shown to play critical roles in translational regulation (48–49). Similar to the ATTCT repeat of ATXN10 gene, the expansion of CCTG repeat affects the adults. One form of dominantly inherited ataxias, referred to as DM, results due to an expansion of a CTG repeat located in the 3'-UTR of DMPK gene encoding a serine-threonine kinase involved in signal pathways of myogenesis (51–52) (Table 1). Three mechanisms have been suggested to explain the dominant mode of DM1 inheritance. These include: (i) Haploinsufficiency of the DMPK protein because its mRNA and protein levels were reduced in muscle biopsies and cell lines derived from patients (53). (ii) Altered expression of neighboring genes, possibly through repeat-induced changes in chromatin structure (54); For example, the CTG repeat of DMPK gene also falls on the promoter region of SIX3 gene which is located immediately downstream of DMPK (55). It was indeed shown that the expansion of CTG in DMPK results in reduction in cellular levels of SIX5 transcripts (37). Curiously, SIX5 knockout mice developed ocular cataracts, similar to DM1 patients, raising a possibility that DM1 is a contiguous gene syndrome involving, among others, deficiency of DMPK and SIX5 proteins (56). (iii) Pathogenic effects of the CUG expansion in mRNA that accumulates as ribonuclear foci and disrupts mRNA splicing of cTNT, IR, CIC-1, Tau and MTMR1 genes (57–62). Curiously, expansion of intrinsic CCTG in the ZNF6 gene (involved in DM2) also leads to the formation of RNA foci and abnormal splicing of above mentioned genes (63, 61). Thus, the similarity of mechanism of mutation between DM2 and DM1 is striking, and suggest that a dominant effect of expanded TRS bearing transcripts on the RNA processing of downstream genes could underlie the genesis of some subphenotypes that are common among the two forms of DM (64).

Second example in this category is the hereditary disorder known as Huntington's disease-like 2 (HDL2). Similar to HD, HDL2 is an adult onset autosomal dominant and progressive neurological disorder clinically characterized by abnormal movements, dementia, and psychiatric syndromes. HDL2 is caused by a CTG expansion mutation located within a variably spliced exon of Juxtaphilin-3 gene (JPH3; 65). Recent studies have identified the presence of a JPH3 splice variant wherein the CTG repeat localize to the 3'-UTR of that transcripts. Transcripts bearing expanded repeats form RNA foci, very similar the foci found in DM1 and DM2, and were detected in neurons of HDL2 patients (66). This study suggests that, similar to DM pathology, RNA toxicity could underlie the pathology in HDL2.

3.1.4 Disease associated TRS in a non-coding RNA
In one form of dominantly inherited ataxias, referred to as SCA8, an expansion of CTG repeats occurs in a region that is transcribed into what is believed to be a processed non-coding RNA (67) (Table 1). Curiously, sequence that transcribes the 5’ end of this untranslated...
### Tandem repeats in human disorders

#### Table 2. Molecular and genetic features of disease-associated unstable repeats falling in the coding regions of genes

<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene implicated</th>
<th>Chromosomal location of the gene</th>
<th>Affected protein</th>
<th>Repeat unit (amino acid)</th>
<th>Repeat number (Q)</th>
<th>Mode of inheritance</th>
<th>Parental gender -base</th>
<th>Effect of mutation</th>
<th>OMIM#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinal-bulbar muscular atrophy</td>
<td>AR</td>
<td>Xq11–12</td>
<td>Androgen receptor</td>
<td>CAG (Q)</td>
<td>581-680</td>
<td>XR</td>
<td>M</td>
<td>GOF</td>
<td>313700</td>
</tr>
<tr>
<td>Huntington Disease</td>
<td>IT15</td>
<td>4q13.5</td>
<td>Huntington</td>
<td>CAG (Q)</td>
<td>581-680</td>
<td>AM</td>
<td>P</td>
<td>GOF</td>
<td>143100</td>
</tr>
<tr>
<td>Dementia-predominant spinocerebellar ataxia type 1</td>
<td>ATN1</td>
<td>12p13.13</td>
<td>Atrophin-1</td>
<td>CAG (Q)</td>
<td>231-249</td>
<td>AD</td>
<td>P</td>
<td>GOF</td>
<td>607482</td>
</tr>
<tr>
<td>Spinal-cerebellar ataxia type 2</td>
<td>ATXN2</td>
<td>6q23</td>
<td>Ataxin-1</td>
<td>CAG (Q)</td>
<td>391-481</td>
<td>AD</td>
<td>P</td>
<td>GOF</td>
<td>145536</td>
</tr>
<tr>
<td>Spinocerebellar ataxia type 3</td>
<td>ATXN3</td>
<td>1q43</td>
<td>Ataxin-3</td>
<td>CAG (Q)</td>
<td>38-688</td>
<td>AD</td>
<td>P</td>
<td>GOF</td>
<td>607040</td>
</tr>
<tr>
<td>Spinocerebellar ataxia type 4</td>
<td>ATXN7</td>
<td>3p12–13</td>
<td>Ataxin-7</td>
<td>CAG (Q)</td>
<td>171-386</td>
<td>AD</td>
<td>P</td>
<td>GOF</td>
<td>607040</td>
</tr>
<tr>
<td>Spinocerebellar ataxia type 5</td>
<td>ATXN10</td>
<td>19p13.2-p13.1</td>
<td>Voltage-dependent calcium channel</td>
<td>CAG (Q)</td>
<td>16-251</td>
<td>AD</td>
<td>P</td>
<td>GOF</td>
<td>601011</td>
</tr>
<tr>
<td>Spinocerebellar ataxia type 6</td>
<td>CAACNA1A</td>
<td>19p13.2-p13.1</td>
<td>Voltage-dependent calcium channel</td>
<td>CAG (Q)</td>
<td>42-47</td>
<td>AD</td>
<td>P</td>
<td>NR</td>
<td>600075</td>
</tr>
<tr>
<td>Infantile onset spinocerebellar ataxia</td>
<td>ATXN6</td>
<td>13q21</td>
<td>Unidentified</td>
<td>CAG (Q)</td>
<td>&gt;26 (74)</td>
<td>AD</td>
<td>NR</td>
<td>NR</td>
<td>601040</td>
</tr>
<tr>
<td>Bipolar transcriptomic prion disease</td>
<td>KIF18B</td>
<td>4p12</td>
<td>Parkin-like proteins</td>
<td>CAG (Q)</td>
<td>&gt;20 (25)</td>
<td>AD</td>
<td>NR</td>
<td>NR</td>
<td>605851</td>
</tr>
<tr>
<td>Mental retardation, Epilepsy, Paroxysmal Strokes</td>
<td>ARX</td>
<td>Xq22.1.2-p21.3</td>
<td>Atrophin-related homoclupe</td>
<td>CAG (Q)</td>
<td>&gt;122 (20)</td>
<td>AD</td>
<td>NR</td>
<td>Partial LOF</td>
<td>500382</td>
</tr>
<tr>
<td>Mental retardation with growth hormone deficiencies</td>
<td>SOX1</td>
<td>Xq27.1</td>
<td>Serum-related HMGB-box transcription factor</td>
<td>CAG (Q)</td>
<td>&gt;15 (26)</td>
<td>XR</td>
<td>NR</td>
<td>LOF</td>
<td>31430</td>
</tr>
<tr>
<td>Ciliakomotosioplasia</td>
<td>RUNX2</td>
<td>4p21</td>
<td>Retinoid-related transcription factor</td>
<td>GCA (A)</td>
<td>&gt;17 (27)</td>
<td>AD</td>
<td>NR</td>
<td>LOF</td>
<td>600021</td>
</tr>
<tr>
<td>Spondylodyplasia</td>
<td>H3X13</td>
<td>2q11.1</td>
<td>Homeobox D13 transcription factor</td>
<td>GCA (A)</td>
<td>&gt;15 (22)</td>
<td>AD</td>
<td>NR</td>
<td>LOF</td>
<td>142999</td>
</tr>
<tr>
<td>Oralepharyngeal muscular dystrophy</td>
<td>PABPN1</td>
<td>14q11.2-q13</td>
<td>Poly Adenylate-binding protein</td>
<td>GCA (A)</td>
<td>&gt;6.9</td>
<td>AD</td>
<td>NR</td>
<td>GOF</td>
<td>602279</td>
</tr>
<tr>
<td>Hand-foot-genital syndrome</td>
<td>RYX1</td>
<td>17p13.13</td>
<td>Homeobox A13 transcription factor</td>
<td>GCA (A)</td>
<td>&gt;12 (18)</td>
<td>AD</td>
<td>NR</td>
<td>LOF</td>
<td>142999</td>
</tr>
<tr>
<td>Pseudohydeplasia-pleurodiaphyseal synostosis</td>
<td>CHMP3</td>
<td>15q15.1</td>
<td>Centriolectin motax protein</td>
<td>GCA (A)</td>
<td>5-9</td>
<td>AD</td>
<td>NR</td>
<td>LOF</td>
<td>600310</td>
</tr>
<tr>
<td>Huntington disease like-2</td>
<td>JPH1</td>
<td>16q23.23</td>
<td>Juncophilins-3</td>
<td>CAG (A or L)</td>
<td>&gt;28 (50)</td>
<td>AD</td>
<td>NR</td>
<td>NR</td>
<td>605268</td>
</tr>
</tbody>
</table>

1Abbreviations used: AR, autosomal dominant; AR, autosomal recessive; XR, X-linked recessive; NR, not reported; De novo, De novo mutation in a germ cell; M, maternal; P, paternal; LOF, loss-of-function; GOF, gain-of-function #Online Mendelian Inheritance in Man (OMIM) ID (More details: http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM)

transcript (known as ATXN8OS) shares the first exon of another gene (KLH1) that is transcribed in the opposite orientation (68). The CTG/CAG repeats, however, do not fall within the coding region of KLH1 gene (68). Thus, ATXN8OS transcript appear to be an endogenous RNA that regulates the expression of sense transcript encoding a Kelch-like protein, and the expanded repeat might affect its regulatory function (68). Targeted disruption of KLH1 gene in mouse resulted in ataxic phenotype, both in homozygous and heterozygous conditions, suggesting that loss of KLH1 activity might play significant part in the SCA8 etiology (69). A recent study also suggests that the SCA8 locus might encode a protein with polyglutamine from a previously unidentified antiparallel transcript of ATXN8OS spanning the CAG repeats and transgenic expression of it results in neurological phenotype in mice (70) (Table 2). Thus, SCA8 pathogenesis might involve dysfunction at both the protein and RNA levels.

### 3.2. Functional significance of TRS in the coding regions of genes

While repeats falling in the transcribed and translated parts of the genome although makeup very insignificant proportion of total repeat content of human genome, they are of utmost functional importance. Thus any dynamism in these repeats is very likely to affect the function that they encode. Analyses of the TRS in the coding regions of genes reveal stronger selection pressure on modulating the length of the repeat. For example, TRS of three or multiples of three bases are found to be overrepresented in the coding regions as compared to mono- or di-nucleotide motifs (71). Taken together, these observations on TRS motifs indicate that the selection against frame shifts strongly influences distribution of repeat lengths (71). Amino acid repeats coded by the TRS in the translated sequences are three times more abundant in the proteome of eukaryotes than in prokaryotes (72-74). Moreover, a majority of these repeats are clustered around non-homologous proteins suggesting that the proteins containing repeats are predominantly unique to eukaryotes (72). For example amino acid repeats are predominantly found in muscle, brain and synaptic cell adhesion proteins, but are underrepresented in very basic cellular functions. Functional classification of proteins with TRS encoded amino acid repeats suggest that these repeats might facilitate protein-protein interactions and therefore required for the formation of multi-protein complexes (75-76). These observations strongly indicate that coding repeats are evolutionarily favored and have an important role in protein functions (72-76).

#### 3.2.1. TRS in the coding region: polyglutamine disorders

Polyglutamine (PolyQ) disorders are perhaps the most infamous and the most vigorously investigated disorders...
Figure 2. Mechanism of pathogenesis associated with PolyQ disorders. Protein with expanded PolyQ tract adopts abnormal conformation and accumulates as intracellular insoluble aggregates in all PolyQ disorders. These aggregates are expected to affect the cellular functions in a variety of ways. These include, recruiting molecular chaperones and components of the ubiquitin-proteasome system (UPS). This effort does remove a portion of the PolyQ proteins from the cytoplasmic pool (arrows shown with dotted lines). However, due to an imbalance between the UPS capacity and the synthesis of aggregation-prone, the aggregates continue to form, and might as well increase the cellular load of abnormal/misfolded proteins. The PolyQ aggregates are also known to affect the transcription and RNA metabolic processes by recruiting several transcriptional activators and repressors (TBP, CBP, CREB, TAFII, SP1, PQBP1) into the aggregates. Other cellular processes can also be affected by the specific interaction of proteins with the aggregates, which ultimately leading to neuronal dysfunction and death.
hypothesis because longer polyglutamine repeats provoke earlier onset of the disease symptoms, and their faster progression (104). However, the functional relationship between PolyQ aggregates and their toxic effects on neurons has been challenged by several studies, and instead argued that the PolyQ aggregates are neuronal protective in nature (105-106). For example, suppression of inclusions formation in a cell-based system resulted in an increased cell-death in a cellular model (107), and neurons of a formation in a cell-based system resulted in an increased nature (105-106). For example, suppression of inclusions argued that the PolyQ aggregates are neuronal protective in

3.2.1.2. The role of PolyQ expansion in transcriptional dysregulation

It is of interest to note that proteins bearing PolyQ tract and associated with disorders are known to be involved in transcriptional regulation (127-128). For example, androgen receptor and TATA-binding protein (involved in SBMA and SCA17 respectively; (Table 2) are transcription factors (129-130). Atrophin-1 (DRPLA), ataxin-1 (SCA1), ataxin-3 (SCA3), and ataxin-7 (SCA7) are known to function as transcriptional regulators (131-134). In addition, huntingtin (HD) is believed function as a co-repressor for transcription factors (135). It is therefore obvious to expect that PolyQ expansion in these proteins would lead to quantitative changes in the neuronal transcriptome. Gene expression studies indeed have shown that dysregulation of neuronal genes could be an early pathogenic event leading to disorders (87, 136-139). Dysregulation involves numerous genes representing diverse cellular pathways, and therefore the mechanism by which PolyQ might cause transcriptional alteration is yet to be fully resolved. Nevertheless, transcriptional regulators such as CREB-binding protein, TBP-associated factor, SP1 transcription factor, and p53 were shown to interact with expanded PolyQ-containing proteins and to localize with their aggregates (140-142). Moreover, increased acetylation of histone H3 at the promoters of a group of genes in the retinal tissue and their transcriptional down-regulation have also been demonstrated in a SCA7 mouse model (143) (Figure 2). Intriguingly, PolyQ pathology has also been linked to micro-RNAs (miRNAs) that are known to modulate the half-life and/or the translational efficiency of mRNAs (144). For example, reduction in the cellular levels of a miRNA, known as ban, correlated with enhanced toxicity to ataxin-3, and its over expression rescued the PolyQ-mediated neurotoxicity, thereby revealing novel pathways and targets for therapeutics for the treatment of PolyQ diseases (144).

3.2.2. TRS in the coding region: polyalanine disorders

Besides the PolyQ expansion associating with several neurodegenerative disorders, recent reports demonstrate the expansion of polyalanine (PolyA) stretches as a disease mechanism in predominantly developmental and congenital malformation syndromes (134) (Table 2). Intriguingly, eight of the nine genes associated with the PolyA disorders encode transcription factors (Table 2). This is not unexpected because stretches of alanine residues are known to be conserved in transcription factors, and thought to play a role in transcriptional repression (76, 146-147). PolyA stretches, however, are relatively shorter repeats and are coded by imperfect trinucleotide repeats (GCN) as against homogenous CAG repeats encoding longer PolyQ sequence (137) (Table 2). In a majority of these disorders, mutations other than PolyA expansion also result in developmental abnormalities, thus suggesting that a different disease mechanism is involved in PolyA disorders as compared to PolyQ disorders (14). Available information suggests that PolyA expansion could lead to complete, partial loss-of-function or a dominant negative gain-of-function of the protein (Table 2). For example, in synpolydactyly limb malformation syndrome, the length of PolyA tract in the HOXD13 protein clearly correlates with the severity of anomalies in both heterozygotes and homozygotes, with homozygosity resulting in more severe phenotype (151, 152). In both human and mice, null alleles of HOXD13 gene manifest less severe phenotype of synpolydactyly than HOXD13 protein with PolyA expansion (152-154). Intriguingly, mice deficient for HoxD complex (simultaneous deletion of Hoxd13, Hoxd12 and
Tandem repeats in human disorders

Hoxd11 genes develop anomalies similar to patients with PolyA expansion in HOXD13 protein suggesting a functional hierarchy among HoxD gene clusters (155). It has indeed been shown in mice models that the PolyA expanded allele of Hoxd11 interferes with the functions of Hoxd11 and Hoxd1, and is likely to have a dominant-negative effect over them (156). Based on the observations on the families with Hand–foot–genital syndrome and mice models, very similar mechanism was also proposed for PolyA expansion in HOXD13 protein, (149, 157-158). PolyA expansion in six other transcription factors (RUNX2, ZIC2, FOXL2, SOX-3, PHOX2B and ARX) thought to result in the loss of protein function because clinical symptoms of patients with PolyA expansions are not noticeably different from those with loss-of-function mutations such as deletions or frame-shift mutations (reviewed in 149) (Table 2). Thus, the dominant nature of the phenotype is likely to result from haploinsufficiency.

Amongst the PolyA expansion disorders, oculoepharyngeal muscular dystrophy (OPMD) is unique because it is a progressive, late onset dominantly inherited syndrome and thus resembles the disease mechanism of PolyQ disorders. The OPMD is also unique because the affected gene (PABPN1) does not code for a transcription factor; PABPN1 protein binds to nascent poly-adenine tails of pre-mRNAs and regulates their formation and length (159-160). PABPN1 protein with expanded PolyA tract tend to form insoluble and ubiquitin-positive aggregates in the nucleus and are toxic to the cell (161, 162). While such aggregates recruit nucleoplasmic RNA, they do not seem to affect the length of the poly-adenine tails of pre-mRNAs, suggesting a gain-of-function effect (162). Analogous to PolyQ disorders, longer lengths of PolyA tracts directly correlate with disease severity (163).

3.2.3. TRS in the coding region: disorders with other amino acid repeats

Mutations in the gene encoding cartilage oligomeric matrix protein (COMP) are associated with two forms of autosomal dominant skeletal dysplasias (pseudoachondroplasia and multiple epiphyseal dysplasia), characterized by variable short stature, joint laxity and early-onset degenerative joint pathologies (164) (Table 2). One of the known mutations is the instability in codons that code for 5 consecutive aspartic acid residues within COMP protein, making it by far the shortest disease-causing repeat expansion mutations described (165). A unique feature of this repeat is that both expansion and shortening of the repeat cause the same disease; deletion of one repeat (4 aspartate residues) cause pseudoachondroplasia and their expansion cause multiple epiphyseal dysplasia (6 aspartate residues) and pseudoachondroplasia (7 aspartate residues) (165). These mutations are likely to impart a loss-of-function effect on COMP as null mutations also result in the same phenotype (165).

The CTG triplet repeat expansion in JPH3 gene associated with Huntington's Disease-like 2, like HD, also has the potential to code for amino acid repeats (65). The CTG repeat is contained within an alternatively spliced exon that has multiple splice sites. Alternate splicing in JPH3 gene could result in transcripts having different reading frames such that the CTG may fall outside (3'UTR) or within the coding region but in two different reading frames, to code polyalanine or polyleucine (65). No study has however confirmed the translational properties of these transcripts, and therefore their involvement in the disease in yet to be established.

4. FACTORS MODULATING TRS INSTABILITY IN THE GENOME

A majority of the TRS implicated in the disorders discussed above are polymorphic with regard to repeat length but their pathological threshold varies depending on whether the TRS lie in the coding or the non-coding region, and if present in the coding region, whether it would code for glutamine or other amino acids (Table 1 and 2) (166, 167). For example, expansions are massive (>1000 repeats) for TRS in the non-coding regions whereas in the coding regions they are relatively smaller (normally <200 repeats) (Table 1 and 2). Within the group of coding TRS, repeats coding for PolyQ show larger length as against short and interrupted repeats that code for PolyA stretches (167). At least for disorder associated with CAG/CTG repeat motifs, the instability is dependent on repeat length as expansion becomes probable only when the repeat length reaches a particular threshold for respective loci (Table 1 and 2). Several studies have proposed a variety of molecular mechanisms to explain the TRS expansions associated with human disorders (reviewed in 7, 78, 168-170). These include DNA replication slippage, DNA damage repair and meiotic recombination. Although whether one or all of these processes contribute to the expansion of repeat is unknown, it is widely believed that the secondary structure that the repeat tracts might adopt could be a critical step in the initiation of expansion process (7, 166, 168, 171-173).

4.1. Replication and TRS expansion

Repeat instability can occur during replication because the extent of repeat instability is known to be dependent on the location of the replication origin (174). For example, the distance between the TRS stretches and the replication origin and the orientation of the TRS with respect to the replication origin were shown to affect the stability of the TRS in various model systems (175-178). Similarly, replication fork dynamics, repeat length, and CpG methylation of the genome closer to the TRS are also known to affect the repeat stability (179, 180). For example, CTG and CCG tracts are more prone to expansion when found in the leading strand, rather than in lagging strand of the replication fork (176, 180, 181). Interestingly, an increased degree of repeat instability was observed when TRS motifs were tested in yeast models deficient in replication proteins (182-184). Corroborating these results from yeast models, studies with human cell lines have shown that the DNA replication initiation regions fall close (<3 kb) to the CAG/CTG tracts of at least 4 different TRS loci associated with neurological disorders (185). Taken together, these studies underscore the role for origins of DNA replication in influencing the TRS length instability and favor the replication slippage model (186-187). In this model, TRS regions in the daughter strand thought to form
'loop-out' structures during DNA replication without disturbing the paring with the parent strand, eventually leading to an increase in the number of repeats in the newly formed DNA strand (172, 186).

4.2. Post-replication repair and TRS expansion

In addition to replication, repeat stability can also be influenced by defects in DNA repair mechanisms, especially in non-proliferating cells (188, 189). DNA repair process might get triggered because of the unusual DNA structure formed within the TRS and might get uncorrected by the repair mechanism (166). For example, single strands of CNG repeat tracts are inefficiently repaired during meiotic recombination in yeast model (180), and such sequence are associated with double strand breaks, leading to the repeat instability (190). An intriguing observation on the TRS instability HD comes from a mouse model; the expanded HD allele becomes very unstable in striatal tissue as compared to other regions of the brain which eventually lead to increased load of PolyQ proteins and neuronal cell death (189). Molecular mechanism of such region-specific expansion is yet to be understood.

4.3. Recombination and TRS instability

Recent studies suggest that recombination mediated gene conversion play a significant role in repeat expansion (191-193). Gene conversion is a non-reciprocal transfer of genetic information wherein there is an alteration in the acceptor allele with no change in the donor allele and no change in the flanking sequence. As expansion in one allele is generally not accompanied by alteration in the acceptor allele with no change in the donor sequence are associated with double strand breaks, leading to the repeat instability (190). An interesting observation on the TRS instability HD comes from a mouse model; the expanded HD allele becomes very unstable in striatal tissue as compared to other regions of the brain which eventually lead to increased load of PolyQ proteins and neuronal cell death (189). Molecular mechanism of such region-specific expansion is yet to be understood.

4.4. TRS in the coding region and codon bias

Anomalous DNA structure that the TRS might form is thought to be one of the prerequisite for initiation of the repeat instability in the genome (196-197). Among the various repeat tested, the CTG/CAG and CGG/CCG repeat tracts were proven to have a higher potential to form secondary DNA structure (7, 198-199). In agreement with this view, a majority of the TRS that code for PolyQ domains, for example, are GC rich (1, 76, 200-201). However, this pattern did not seem to differ between the TRS encoding polymorphic and non-polymorphic PolyQ, suggesting that some cis-acting elements (flanking sequence) might regulate the instability of TRS (172, 202, 203). A recent study also suggests that codon bias is one of the limiting factors in the motif selection for TRS (76). This study demonstrates that rare codons, despite their GC-richness, are not favored when coding amino acid repeat tracts. In this regard, it is interesting to note that clustering of rare codons within a narrow region significantly reduce the half-life of the mRNA (204 - 206). Thus, mRNA stability could be one of the factors that work against the usage of rare codons in the TRS (76).

4.5. Parent-of-origin effect in TRS expansion

Gender bias in the transmission of unstable repeats from the parent to offspring is a common feature of dynamic mutations (Table 1 and 2). For example, the transmission of TRS through males was less stable than that through females for genes involved in DRPLA (207), HD (208), and SCA1 (209). However, it is the female sex in the case of FRDA (210) and SCA7 (211). The underlying cause for this sex-dependent transmission bias is thought to be difference in the meiotic process in the germ cell or may be determined at the time of embryonic development itself (211, 212). Alternatively, sex bias in the recombination process could be the reasons for the TRS instability. A recent study shows that the sex-specific recombination rates for ~5 mega base genomic region spanning the six genes associated with TRS expansion and disorder strongly correlate with the parental gender that positively influences the repeat instability (76). Since the recombination rate is known to be non-homogeneous in the genome (213) and that there are regions in the genome where recombination rate is particularly high in women and particularly low in men or vice versa (214), the regional differences in the recombination rate might influence the stability of TRS that they harbor (76).

4.6. Role of evolution in TRS instability

Length of TRS and its distribution within a gene sequence is nonrandom and strongly biased, and does have some selection constraints. Amino acid repeats coded by TRS are abundant in the human proteome but only a few of them are known to play any functional role (75-76, 215-217). However, a significant proportion of such genes are linked to OMIM and Morbid databases suggesting their potential association with genetic disorders (76, 218-220). Conservation of the non-coding TRS in gene orthologous is rather low; however, TRS in the coding region does show conservation at least among mammals, but not in their size or variability in the population, suggesting differential selection for TRS bearing genes (76, 201, 217, 221). It is of interest to note that, TRS encoded larger stretches of amino acid repeats are far more frequent in orphan proteins as compared to familial proteins, suggesting that the two forms of repeat coding genes are subjected to differential selection constraints (76, 201). Based on these observations, Siwach et al (76) proposed that the recently evolved solitary genes might acquire longer TRS stretches because of the weaker constraints placed on them. However, expansion beyond a "threshold" could become pathogenic and therefore would not be fixed in the population (76). Corroborating this view, pathogenic genes bearing longer TRS do represent orphan groups, and do not have orthologues in invertebrates (76).
Tandem repeats in human disorders

5. PERSPECTIVE AND CONCLUDING REMARKS

While on one hand the genome tries to copy itself with utmost fidelity, other mechanisms like replication, recombination and the evolution of sex, on the other hand, attempts to increase variations in the genome. Variations make species more robust and adaptive to changes but considerable number of individuals in the population will have to suffer in order to increase the option of selection. Among the different types of variations known in the genome, TRS provide a wide range of genotypes (>10) at a given locus as against single nucleotide polymorphism that offer only two genotypes. Thus instability in TRS falling within the gene sequence has greater potential to increase the phenotypic variability in the population and therefore likely to provide increased fitness under natural selection (222, 223). For example, length variations in TRS falling in the coding regions of the Alx-4 and Runx-2 genes were shown to quantitatively associate with evolution of limb and skull morphology in dogs (224). Similarly, TRS variations in the Avpr1a gene generate diversity in socio-behavioral traits in voles by modulating the expression level of the gene (225). Thus, the quantitative effects of the repeat-length variations in the TRS appear to function as “tuning knobs” for gene regulation and may have been favored by evolution as a source of phenotypic variability (222, 223).

6. ACKNOWLEDGMENTS

Work on the tandem repeats in our laboratory was supported by a research grant from the Ministry of Human Resource Development, Government of India, to SG. PS was supported by a research fellowship from the Council of Scientific and Industrial Research, Government of India.

7. REFERENCES

Tandem repeats in human disorders


32. G. R. Sutherland: Rare fragile sites. *Cytogenet Genome Res* 100(1-4), 77-84 (2003)


40. M. B. Delatycki, M. Knight, M. Koenig, M. Cossee, R Williamson and S. M. Forrest: G130V, a common FRDA point mutation, appears to have arisen from a common founder. *Hum Genet* 105(4), 343-346 (1999)


Tandem repeats in human disorders


156. S. Bruneau, K. R. Johnson, M. Yamamoto, A. Kuroiwa and D. Duboule: The mouse Hoxd13 (spdh) mutation, a polyalanine expansion similar to human type II synpolydactyly (SPD), disrupts the function but not the expression of other Hoxd genes. Dev Biol 237(2), 345-353 (2001)
158. L. C. Post, E. H. Margulies, A. Kuo and J. W. Innis: Severe limb defects in Hypodactyly mice result from the


Tandem repeats in human disorders


Tandem repeats in human disorders


**Note:** Further details on the gene, mutations, and disease can be obtained from the OMIM link (http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM), using the MIM number provided

**Abbreviations:** AD: autosomal dominant, AR: autosomal recessive, XR: X-linked recessive, NR: not reported, De novo: De novo mutation in a germ cell, M: maternal, P: paternal, LOF: loss-of-function, GOF: gain-of-function

**Key Words:** Tandem repeat sequences, repeat dynamism, PolyQ disorders, PolyA Disorder, Ubiquitin-Proteasome System, neuronal dysfunction, cell death, Review

**Send correspondence to:** Dr Subramaniam Ganesh, Department of Biological Sciences and Bioengineering, Indian Institute of Technology, Kanpur 208016, India, Tel: 91-512-259-4040, Fax: 91-512-259-4010, E-mail: sganesh@iitk.ac.in

http://www.bioscience.org/current/vol13.htm