Microsatellites Within Genes: Structure, Function, and Evolution

You-Chun Li,† Abraham B. Korol,† Tzion Fahima,† and Eviatar Nevo†*

†Institute of Evolution, University of Haifa, Haifa, Israel; *Department of Plant Sciences, University of Arizona

Recently, increasingly more microsatellites, or simple sequence repeats (SSRs) have been found and characterized within protein-coding genes and their untranslated regions (UTRs). These data provide useful information to study possible SSR functions. Here, we review SSR distributions within expressed sequence tags (ESTs) and genes including protein-coding, 3′-UTRs and 5′-UTRs, and introns; and discuss the consequences of SSR repeat-number changes in those regions of both prokaryotes and eukaryotes. Strong evidence shows that SSRs are *nonrandomly* distributed across protein-coding regions, UTRs, and introns. Substantial data indicates that SSR expansions and/or contractions in protein-coding regions can lead to a gain or loss of gene function via frameshift mutation or expanded toxic mRNA. SSR variations in 5′-UTRs could regulate gene expression by affecting transcription and translation. The SSR expansions in the 3′-UTRs cause transcription slippage and produce expanded mRNA, which can be accumulated as nuclear foci, and which can disrupt splicing and, possibly, disrupt other cellular function. Intronic SSRs can affect gene transcription, mRNA splicing, or export to cytoplasm. Triplet SSRs located in the UTRs or intron can also induce heterochromatin-mediated–like gene silencing. All these effects caused by SSR expansions or contractions within genes can eventually lead to phenotypic changes. SSRs within genes evolve through mutational processes similar to those for SSRs located in other genomic regions including replication slippage, point mutation, and recombination. These mutational processes generate DNA changes that should be connected by DNA mismatch repair (MMR) system. Mutation that has escaped from the MMR system correction would become new alleles at the SSR loci, and then regulate and/or change gene products, and eventually lead to phenotype changes. Therefore, SSRs within genes should be subjected to stronger selective pressure than other genomic regions because of their functional importance. These SSRs may provide a molecular basis for fast adaptation to environmental changes in both prokaryotes and eukaryotes.

Simple Sequence Repeat Abundance in Transcribed Regions

Numerous lines of evidence have demonstrated that genomic distribution of simple sequence repeats (SSRs) is *nonrandom*, presumably because of their effects on chromatin organization, regulation of gene activity, recombination, DNA replication, cell cycle, mismatch repair (MMR) system, etc. (see review, Li et al. 2002). SSRs may provide an evolutionary advantage of fast adaptation to new environments as evolutionary tuning knobs (Kashi, King, and Soller 1997; Trifonov 2003). These SSRs with putative functions may be located in gene or regulatory regions. However, the reviews published to date have not clearly discussed SSR polymorphism and evolution *in gene*, because available information about SSR locations on chromosomes has been limited.

Recently, however, many reports have demonstrated that a large number of SSRs are located in transcribed regions of genomes, including protein-coding genes and expressed sequence tags (ESTs) (e.g., Morgante, Hanafey, and Powell, 2002), although in general, repeat numbers and total lengths of SSRs in these regions are relatively small (Kantety et al. 2002; Thiel et al. 2003). For example, it has been found that ~12% of identified SSRs in Japanese pufferfish (Edwards et al. 1998), 10% in primate (Jurka and Pethiyagoda 1995), 15% in rabbit (van Lith and van Zutphen 1996), and 9.1% and 10.6%, respectively, in pig and chicken (Moran 1993) are located in the protein-coding genes or open reading frames (ORFs). In cereals (maize, wheat, barley, sorghum, and rice) 1.5%–7.5% of ESTs consist of SSRs (Kantety et al. 2002; Thiel et al. 2003). These ESTs have a range of functions such as metabolic enzymes, structural and storage proteins, disease signaling, and transcription factors, suggesting some role(s) of SSRs in plant metabolism and gene evolution. In protein-coding regions of all known proteins, 14% proved to contain repeated sequences, with a three times higher abundance of repeats in eukaryotes as in prokaryotes (Marcotte et al. 1999). Noteworthy, prokaryotic and eukaryotic repeat families are clustered to nonhomologous proteins. This may indicate that repeated sequences emerged after these two kingdoms had split (Marcotte et al. 1999). Although SSR sequences are relatively rare in prokaryotes, studies based on computer analysis of microbial whole genome sequences revealed overrepresentation of a few SSR motifs in several microbial species such as the *Neisseria* species, *Haemophilus influenzae*, *H. parainfluenzae*, *Moraxella catarrhalis* (see review by van Belkum et al. 1998).

Debates over whether SSRs play any functional role in organism development, adaptation, survival, and evolution are never-ending. The available location of specific SSRs in known genes and ESTs permits the unraveling of the biological significance of SSR distribution, expansion, and contraction on the function of the genes themselves. This article reviews the accumulating data on SSR distribution and structure, and phenotypic effects of SSR expansion and contraction within coding regions, 3′-UTRs, 5′-UTRs, and introns.

SSRs in Coding Regions

Nonrandom Distribution

Numerous SSRs exist in ORFs of higher eukaryotes including *Drosophila*, *Caenorhabditis elegans*, mammals, humans, plants, and yeast (Tóth, Gáspári, and Jurka 2000;

Key words: microsatellite, gene, gene-inactivation, truncated protein, phenotype, adaptation.

E-mail: nevo@research.haifa.ac.il.

Mol. Biol. Evol. 21(6):991–1007. 2004
DOI:10.1093/molbev/msh073
Advance Access publication February 12, 2004

Table 1
Frequency (%) of 10 Subclasses of Triplet SSRs with Similar Physical and Chemical Properties

	Primates	Rodents	Mammals	Vertebrates	Arthropods	*C. elegans*	*A. thaliana*	Barley	Sugarcane	*S. cerevisiae*	Fungi
AAC	0.7	0.0	0.0	0.0	14.0	10.7	22.6	2.1	4.7	22.1	28.1
AAG	5.1	1.9	2.5	5.2	0.0	26.3	28.3	7.2	5.1	20.8	14.4
AAT	0.0	0.0	0.0	0.0	0.6	6.2	0.4	0.9	0.0	17.4	8.9
ACC	5.2	11.8	3.5	5.6	9.1	21.1	8.0	6.0	9.0	1.7	6.8
ACG	0.4	0.0	7.4	1.8	1.3	3.2	1.8	11.3	20.5	3.8	3.1
ACT	0.4	0.0	0.0	0.0	0.0	1.9	1.2	3.0	0.0	0.7	1.3
AGC	33.8	57.1	42.9	40.9	60.9	10.1	9.7	14.3	8.1	15.9	17.6
AGG	17.1	9.2	25.0	25.5	5.3	4.5	9.5	15.1	12.0	3.0	7.1
ATC	2.3	1.4	2.1	1.8	3.7	13.3	12.3	3.8	1.3	14.6	5.8
CCG	35.0	18.6	17.6	19.1	5.0	2.6	6.3	36.4	39.3	0.0	6.8
Total repeat length (bp)	1,126	1,557	876	826	1,566	308	1,119	*470	*234	706	381

NOTE.—Triplet classification was cited from Jurka and Pethiyagoda (1995). Results calculated based on data in: sugarcane—Cordeiro et al. 2001 (repeat number ≥ 5); barley—Thiel et al. 2003 (repeat number ≥ 5); others— Tóth, Gáspári, and Jurka 2000 (repeat number ≥ 4); total repeat length = repeat base pairs per DNA megabase; an asterisk indicates total triplet SSRs.

Katti, Ranjekar, and Gupta 2001; Kantety et al. 2002; Morgante, Hanafey, and Powell 2002). SSR occurrence in coding regions seems to be limited by non-perturbation of the reading frame. This has been proved by the following facts: (1) in a human cDNA database, more than 92% of the predicted SSR polymorphisms within coding sequences have repeat-unit sizes that are a multiple of three (Wren et al. 2000); (2) in many species, exons (unlike other genomic regions) contain rare dinucleotide and tetranucleotide SSRs, but have many more triplet and hexanucleotide SSRs than other repeats (Field and Wills 1996; Edwards et al. 1998; Metzgar, Bytof, and Wills 2000; Wren et al. 2000; Young, Sloan, and van Riper 2000; Cordeiro et al. 2001; Morgante, Hanafey, and Powell 2002). Triplet repeats show approximately twofold greater frequency in exonic regions than in intronic and intergenic regions in all human chromosomes except the Y chromosome (Subramanian, Mishra, and Singh 2003). In prokaryotic genes related to adaptation or responses to stress of *Escherichia coli* K12, mononucleotide and trinucleotide SSRs are significantly overrepresented, whereas dinucleotide and tetranucleotide repeats are underrepresented (Rocha, Matic, and Taddei 2002). Such dominance of triplets over other repeats in coding regions may be explained on the basis of the suppression of non-trimeric SSRs in coding regions, possibly caused by frameshift mutations (Metzgar, Bytof, and Wills 2000).

The presence of SSRs in coding regions shows bias to some specific nucleotide composition. Thus, A/T repeats are more frequent (11.8% of 45,425 coding sequences: CDSs) than G/C repeats (0.7%) in human coding sequences (Olivero et al. 2003). Exons and ESTs show higher frequency for GA/CT repeat than for AT repeat in *Arabidopsis thaliana* (Morgante, Hanafey, and Powell 2002) and cereals (Kantety et al. 2002; Morgante, Hanafey, and Powell 2002). The AC/GT repeats in plants were more rare than in animal genomes (Tóth, Gáspári, and Jurka 2000; Morgante, Hanafey, and Powell 2002). This pattern may be related to high frequencies of certain amino acids in plants than in animals (Tóth, Gáspári, and Jurka 2000).

Triplet repeats in exons can be grouped into 10 motif subclasses (table 1), each representing six overlapping and complementary unit patterns (Jurka and Pethiyagoda 1995). In the animal kingdom, AGC was the most common motif (40.9%–60.9%). In plants, the most frequent triplet motif is AAG subclass (28.3%–42.1%) in *A. thaliana*, grape, and endophytes (table 1). In cereal species, however, the most common triplet is CCG in all the species, ranging from 32% in wheat to 49% in sorghum (Varshney et al. 2002; Thiel et al. 2003), to 39.3% in sugarcane (Cordeiro et al. 2001). The abundance of CCG repeats is a specific feature of monocot genomes, and it may be due to their increased GC content (Morgante, Hanafey, and Powell 2002). The AAT motifs were the least common ($<1\%$) in monocot species (Cordeiro et al. 2001; Varshney et al. 2002; Thiel et al. 2003) and in other species listed in table 1. This may be explained by the fact that TAA-based variants code for stop codons that have a direct effect on protein synthesis in eukaryotes.

Frequencies of Different Codon Repeats Vary Considerably Depending on the Type of Encoded Amino Acid

In plants, the most common codon repeats are codons for Lys in *Arabidopsis* and codons for Arg in sugarcane (table 2). In *Drosophila*, *C. elegans*, and yeast, the most common codon repeats are CAA and CAG encoding Gln in complete genome coding DNA sequences (Katti, Ranjekar, and Gupta 2001). It is interesting that those expansions of codon repeat corresponding to small/hydrophilic amino acids are more tolerated (with ≥ 14 repeat times) than are hydrophobic amino acids (with shorter repeat times) (Katti, Ranjekar, and Gupta 2001). At the DNA level, the AGC, GCA, CAG, CTG, TGC, and GCT repeats (representing the same repeating DNA duplex) are quite similar, and their frequencies can be expected to be comparable. In fact, however, *Drosophila* coding regions display a strong bias to CAG (Gln: 77.5% of a total of 1,909 of this repeat group), and very rare for CTG (leucine: 0.6%) and TGC (cysteine: 0.2%) (Katti, Ranjekar, and Gupta 2001). In yeast proteins, the most abundant amino acid repeats are codons of Gln, Asn, Asp, Glu, and Ser (Richard and Dujon 1997; Alba, Santibáñez-Koref, and Hancock 1999). Different amino acid repeats are concentrated in different classes of proteins. Acidic and

Table 2
Frequency (%) of Biased Distribution of Codon Repeats in Different Species

Codon	Encoded Amino Acid	*Drosophila*	*C. elegans*	Yeast	*Arabidopsis*	Sugarcane
GGA/GGG/GGC/GGT	Glycine (Gly)	4.7	6.6	0.8	5.1	0.0
GCA/GCG/GCC/GCT	Alanine (Ala)	9.2	6.3	2.7	9.1	8.2
ATA/ATC/ATT	Isoleucine (Ile)	0.3	0.6	0.0	6.6	0.0
CCA/CCG/CCC/CCT	Proline (Pro)	1.8	13.3	1.4	8.1	12.0
TAC/TCT/TCC/TCTAGC/AGT	Serine (Ser)	8.4	3.8	7.0	0.0	0.0
ACA/ACG/ACC/ACT	Threonine (Thr)	4.0	4.1	0.8	6.7	20.5
AAC/AAT	Asparagine (Asn)	5.8	3.2	16.4	10.1	1.3
CAC/CAG	Glutamine (Gln)	52.0	16.8	25.3	0.0	4.8
GAC/GAT	Aspartic acid (Asp)	2.6	14.0	16.8	0.0	0.0
GAA/GAG	Glutamic acid (Glu)	5.5	12.7	16.8	0.0	0.0
AAA/AAG	Lysine (Lys)	1.6	10.1	4.6	37.9	5.1
CGA/CGG/CGC/CGT/AGA/AGG	Arginine (Arg)	0.1	1.2	0.8	6.5	39.3
CAC/CAT	Histidine (His)	3.1	3.1	2.7	0.0	9.0
Others[a]		1.2	3.2	3.9	1.2	0.0
Total occurrences of repeats		2,993	773	483	942	234

Frequency was calculated based on data in: Katti, Ranjekar and Gupta (2001); Morgante, Hanafey, and Powell (2002); Cordeiro et al. (2001)

[a] Including codons for valine, leucine, cysteine, methionine, phenylalanine, and tryptophan.

polar amino acid repeats are significantly associated with transcription factors and protein kinases, whereas Ser repeats are significantly associated with membrane transporter proteins (Alba, Santibáñez-Koref, and Hancock 1999). Interestingly, in yeast, the longest triplet repeats ($\geq$ 75 bp) are often found in nuclear-protein genes (Richard and Dujon, 1997). Strong bias in favor of certain limited sets of amino acids in different proteins or cell locations showed that triplet repeats in ORFs were nonrandom with respect to the ORFs and DNA strands (Richard and Dujon 1997). Similarly, in humans and mice, repeat-containing genes were enriched in certain amino acids such as Pro, Gln, His, and Ser (for codons, see table 2; Hancock, Worthey, and Santibáñez-Koref 2001). Likewise, CGG, CCG, CAG, and GAA repeats coding for $(Ala)_n$, $(Gly)_n$, $(Pro)_n$, $(Gln)_n$, and $(Lys)_n$ are abundant in primate genes (Borštnik and Pumpernik 2002). The above evidence may suggest that functional selection acts on amino acid reiteration in the encoded proteins (Alba, Santibáñez-Koref, and Hancock 1999; Katti, Ranjekar, and Gupta 2001), but this selection did not reflect underlying biases in base composition.

The abundance of CAG repeats in yeast coding regions (Alba, Santibáñez-Koref, and Hancock 1999; Katti, Ranjekar, and Gupta 2001) parallels its abundance in mammalian exons (Stallings 1994). However, AAT repeats, also very abundant in yeast coding regions, are rare in the exons of mammals, and GGC repeats, relatively abundant in mammalian exons (Stallings 1994), are uncommon in yeast genes. This may be because Asn repeats (AAT) are not tolerated in mammals, and that the same is true for Gly repeats (GGC) in yeast. Asn repeats appear to be rare in vertebrates but more common in invertebrate, yeast, and plant proteins (Stallings 1994). Alternatively, these differences could be due to differences in the slippage process between the groups, or they may reflect the low GC content of the yeast genome (Richard and Dujon 1997).

Phenotypic Effect of SSRs in Coding Regions

Simple sequence repeat variation within genes should be very critical for normal gene activity because encoding SSR expansion or contraction directly affects the corresponding gene products and even causes phenotypic changes. In eukaryotes, SSR effects in coding regions on phenotypes have been extensively studied only in human diseases, revealing abundant evidence on human neuronal disorders and cancers.

Exon CAG Repeat Expansion Produces Toxic Mutant Proteins Causing Human Diseases

Human repeat expansion diseases are predominantly neurological and are caused by instability and expansion of triplet motifs within or near genes (reviewed in Cummings and Zoghbi 2000; Masino and Pastore 2002). The largest class of these diseases results from the expansion of coding CAG repeats that are translated into extended $(Gln)_n$ tracts within the corresponding proteins. These dominantly inherited diseases include Huntington's disease (HD), dentatorubro-pallidoluysian atrophy (DRPLA), spinobulbar muscular atrophy (SBMA), and spinocerebellar ataxia (SCA1, SCA2, SCA3, SCA6, and SCA7; table 3). All eight disorders are progressive, typically striking in midlife, and causing increased neuronal dysfunction and eventually neuronal loss 10–20 years after the onset of symptoms. Several other features characterize this group of diseases: the greater the number of CAG repeats on expanded alleles, the earlier the age of onset and the more severe the disease. The repeats show both somatic and germline instability (see review: Zoghbi and Orr 2000; Lima et al. 2001). Successive generations of affected families experience anticipation, or earlier age of onset, and more rapid disease progression as a result of intergenerational repeat instability that is particularly marked in paternal transmissions. CAG contraction within androgen receptor gene was involved in cancer or other diseases (table 3).

How could the expanded CAG repeats cause these diseases? There is strong evidence demonstrating that the expanded $(Gln)_n$ stretch confers either a gain- or change-of-function onto the corresponding proteins (see reviews: Galvão et al. 2001; Ranum and Day 2002). In most cases, a toxic gain of function of the mutant protein was

Table 3
SSRs Within Coding Regions, and Their Function and Phenotypic Effects in Several Species

Species	Repeat	Gene	SSR Function and Related Phenotypic Effect	Reference
Human	(CAG)$_n$	*HD*	Expansion causes Huntington's disease (HD)	Zoghbi and Orr (2000)
	(CAG)$_n$	*DRPLA*	Causes dentatorubro-pallidoluysian atrophy (DRPLA)	Nakamura et al. (2001)
	(CAG)$_n$	*KR*	Causes spinobulbar muscular atrophy (SBMA)	Mao et al. (2002)
	(CAG)$_n$	*SCA*	Causes spinocerebellar ataxias (SCA, types 1 to 7)	
	(CAG)$_n$	Androgen receptor (AR)	Shorter repeat increases hepatitis B virus (HBV)–related hepatocellular carcinoma risk	Yu et al. (2001; 2002) Bennett et al. (2002)
			Modifies prostate cancer risk and progression	Buchanan et al. (2001)
			CAG repeat size links to AR activity, abnormally large	Coetzee and Irvine (2002)
			AR-CAG sizes result in SBMA with partial androgen insensitivity, which is related to Kennedy's disease	Dejager et al. 2002
	(CGC)$_n$	Poly(A)-binding protein 2 (PABP2)	Oculopharyngeal muscular dystrophy	Brais et al. (1998)
	(A)$_n$	MMR genes *hMSH2; hMLH1, hMSH6* *hPMS1; hPMS2*	Frameshift caused by (A)$_n$ size changes inactivates MMR genes and causes human cancers	Duval and Hamelin (2002) Vassileva et al. (2002)
	(A)$_n$	MMR genes: MBD4/MED1	Frameshift caused by (A)n size changes inactivates MMR genes and causes human cancers	Yamada et al. (2002*a*)
	(A)$_n$	Signal transduction genes: *TGFRII, IGFIIR, ACTRII, WISP* *GRB-14, AXIN-2*	Tumor-suppressive function	Markowitz et al. (1995) Souza et al. (1996)
	(A)$_n$	Apoptosis: *BAX, caspase 5, APAF-1, BCL-10, FAS*	Tumor-suppressive function	Rampino et al. (1997) Schwartz et al. 1999)
	(A)$_n$	Transcriptional regulation: *TCF-4, CDX2*	Tumor-suppressive function	Duval et al. (1999)
	(A)$_n$	Immune surveillance: *2M*	Tumor-suppressive function	Bicknell et al. (1996)
	(A)$_n$	Cell cycle: *PTEN, RIZ, Hg4-1*	Tumor-suppressive function	Guanti et al. (2000) Zhou et al. (2002)
	(A)$_n$	Response to DNA damage: *BLM, CHK1, RAD-50*	Tumor-suppressive function	Duval and Hamelin (2002)
Drosophila	(CAG)$_n$	Homeobox gene *DLX6*	Triplet expansion leads to cell death	Ferro, dell'Eva, and Pfeffer (2001)
H. influenzae	CAAT	Virulence genes	Mediating phase variation to adapt to host environmental changes	Karlin, Mrazek, and Campbell (1997)
	GCAA	Virulence genes		Weiser, Love, and Moxon (1989)
	CAAC	Virulence genes		van Belkum et al. (1998)
	GACA	Virulence genes		
	AGCT	Virulence genes		
	TTTA	Virulence genes		
	CAAT	Virulence genes: *lic1, lic2, lic3*	Affect LPS expression and phase variation	Roche and Moxon (1995) Weiser, Love and Moxon (1989)
	AGTC	MMR gene: *poll*	Facilitate adaptive switching	Bayliss, van de Ven, and Moxon (2002)
H. somnus	CAAT	LOS component gene	Mediating LOS phase variation and adaptation	Inzana et al. (1997)

Table 3
Continued

Species	Repeat	Gene	SSR Function and Related Phenotypic Effect	Reference
	CAAT	LPS synthesis-related gene	Mediating LPS phase variation and adaptation	Peak et al. (1996)
Neiserria spp.	GCAA	Virulence genes	Mediating phase variation and adaption	
M. catarrhalis	CAAC	Virulence genes	Mediating phase variation and adaption	
M. hyorhinis	AGT	Lipoprotein gene *MG307*	Regulates gene translation and influences activity of this surface antigen	Rocha and Blanchard (2002)
N. gonorrhoeae	$(G)_n$	LOS component: *lsi2*	Responsible for LOS-specific phenotypic change	Burch, Danaher and Stein (1997)
Chlamydia pneumoniae	$(G)_n$	Membrane protein-gene: *pmp10*	Involved in virulence and pathogenesis of *Chlamydia*	Grimwood et al. (2001)
	$(C)_n$	Outer membrane proteins (*Ppp*)	Involved in the pathogenesis of *C. pneumoniae*	Rocha et al. (2002)

Note.—LPS: lipopolysaccharides; LOS: lipooligosaccharide.

demonstrated. For instance, both cell culture and animal studies clearly demonstrate that a long $(Gln)_n$ tract is toxic to neurons and peripheral cells alike (Galvão et al. 2001). Additional peptide sequences from expanded CAG repeats must contribute to the late onset of these diseases and selective neuronal vulnerability. This neuronal selectivity disappears in the earliest juvenile-onset cases, when the $(Gln)_n$ tract becomes disproportionately large relative to the rest of the protein; there may be a threshold for Gln repeat length beyond which it becomes the predominant toxic moiety (Zoghbi and Orr 2000). It is also possible that RNAs containing CAG expanded repeats may either interfere with processing of the primary transcript, resulting in a deficit of the corresponding protein, or interact with RNA-binding proteins, altering their normal activity (see review: Galvão et al. 2001). Recent experiments suggest that, in addition to the ubiquitin/proteasome pathway, mutant proteins with expanded $(Gln)_n$ stretches are involved in the lysosomal pathway for protein degradation, and that this processing mechanism may serve as a target for new therapeutic approaches to CAG repeat diseases (Yamada, Tsuji, and Takahashi 2002).

Monomer SSR Variation in Coding Regions Inactivates MMR Genes via Frameshift, but Leads Also to Other Truncated Proteins in Tumor Cells

Microsatellite instability (MSI) occurs in about 90% of hereditary nonpolyposis colorectal cancers (HNPCC) as well as in about 15% of sporadic tumors of the colon (Mark Redston 2001), in numbers of gastric (Yamada et al. 2002*a*), lung (Zienoddiny et al. 1999), and endometrial cancers (Vassileva et al. 2002). It has been proved that many MSI tumors are caused by mutational inactivation of the different MMR genes listed in table 3. The mutational inactivation was caused mainly by a frameshift occurring within the $(A)_n$ tracts located in exons of both major and minor MMR genes (for review, see Duval and Hamelin 2002), except the methylation of the *hMLH1* promoter (Yanagisawa et al. 2000).

Likewise, a number of SSR-containing genes (listed in table 3) are frequently affected by the MSI in tumor cells. Like the MMR genes, these genes also contain $(A)_n$ tracts in coding regions, which can lead to frameshift mutation in MMR-defective cells (Duval and Hamelin 2002). The proteins encoded by these genes display tumor-suppressive functions and, thus, represent major targets of mutator pathway-associated carcinogenesis (Schwartz et al. 1999). Most MSI–High-frequency (MSI-H) tumors had acquired frameshift mutations in more than one gene among hMSH3, hMSH6, BAX, IGFIIR, TGFbetaIIR, E2F4, and BRCA2 (Johannsdottir et al. 2000).

The human hTCF4 gene interacts functionally with β-catenin in the Wnt signaling pathway. Alternative use of different reading frames in the exon 17 of hTCF-4 generates different protein isoforms with agonist or antagonist transactivating activities (Duval et al. 2000). A 1-bp deletion in an $(A)_9$ repeat within exon 17, as well as other frameshift mutations, results in decreasing the proportion of the long COOH-terminal hTCF-4 isoform, which contains two binding domains for c-terminal binding protein, a protein implicated in the repression of the TCF family transcriptional activity. Thus, loss of the TCF-4 capacity to interact with COOH-terminal binding protein could have an important effect in colorectal carcinogenesis by modifying Wnt-signaling (Duval et al. 2000).

SSR Variation in Coding Regions Affects Gene Expression and Pathogenesis in Prokaryotes

The presence of SSRs in prokaryotes is rare, but most that do occur are related to pathogenic organisms; their variation in repeat numbers can also cause phenotypic changes (reviewed in van Belkum et al. 1998). These SSR motifs were reminiscent of the presence of repetitive elements consisting of uptake signal sequences, intergenic dyad sequences, and multiple tetranucleotide iteration (Karlin, Mrazek, and Campbell 1997). *Haemophilus* *influenzae* (*Hi*), an obligate upper respiratory tract commensal/pathogen, uses phase variation (PV) to adapt to host environment changes. Switching occurs by slippage of SSR repeats within genes coding for virulence molecules (Hood et al. 1996). Most such SSRs in *Hi* are tetranucleotide repeats, which are listed in table 3. The high prevalence of tetranucleotides mediating PV is an exceptional feature of the *Hi* genome. For instance, Weiser, Love, and Moxon (1989) found that different patterns of lipopolysaccharide expression (LPS: LPS phase variation functions as an adaptation mechanism enabling bacteria to escape the immune system attack and to translocate across various physical barriers: van Putten 1993) and the molecular switch leading to this phenotypic variability appeared to be dependent on the translational capacity of the gene *lic1* mRNA, which is caused by variable numbers of a CAAT motif within this gene. Variation in the overall number of this CAAT unit moves one of the three ATG codons in or out of the protein synthesis reading frame, and it then directly affects protein synthesis and the primary amino acid sequence. More examples are listed in table 3.

Variation in the opacity surface proteins (Opa) occurs by *recA*-independent rearrangements in the coding repeat sequence. In this region, shifting of the translational reading frame occurs because of changes in the repetitive DNA track (Murphy et al. 1989). These changes occur independently in any of the *opa* genes, which account for the production of several different Opa proteins simultaneously. The relationship between the superficial Opa protein composition of a bacterial isolate with invasiveness into the human epithelium has been demonstrated experimentally (Makino, van Putten, and Meyer 1991). A variable number of CTCTT motifs in the *opa* leader peptide moves the reading of the gene in or out of the frame. This sequence is peculiar because a triple-helix conformation is likely to occur, and the CTCTT repeat region appears to be hypersensitive to single-strand-specific nuclease. The number of repeats varies continuously at low frequency in vivo. Once environmental selection influences survival of one of the "minority SSR types," this type will "translate" its selective advantage into overgrowth of the existing population (Makino, van Putten, and Meyer 1991).

In general, the examples mentioned in this section and in table 3 emphasize the importance of SSR elements in many aspects of *adaptive* behavior in bacteria. SSR variations enable bacteria to respond to diverse environmental factors, and many of them are clearly related to bacterial pathogenesis and virulence. The contingent genes containing SSRs show high mutation rates, allowing the bacteria to act swiftly on deleterious environmental conditions (Moxon et al. 1994). Some of the SSRs seem to play an essential role in controlling surface exposition of active protein domains and antigenic variation.

SSRs in UTR Regions

SSR Distribution in 3′-UTRs and 5′-UTRs

In transcribed regions, according to available large-scale observation in humans and *Arabidopsis* databases, UTRs harbor more SSRs than the coding regions (Wren et al. 2000; Morgante, Hanafey, and Powell 2002). The 5′-UTRs and 3′-UTRs contain more monomer and dimer motifs than those in coding regions in humans, and they contain higher amounts of both monomers and triplets in the 5,000 *Arabidopsis* genes (table 4).

For dimeric repeats in UTRs, plants and animals show completely different biases. In *Arabidopsis*, the UTRs, in particular the 5′-UTRs, exhibit a strong bias toward AG/CT (Morgante, Hanafey, and Powell 2002). However, 3′-UTRs show a bias to AC/GT in catfish (55.9% of 34 SSRs; Liu et al. 2001) and in humans (52.3% of 44 SSRs, Wren et al. 2000). In human genes, SSRs in the 3′-UTRs also display an obvious bias to $(A/T)_7$ (27.0%) compared to $(G/C)_7$ (0.7%) (Olivero et al. 2003).

The 5′-UTRs contained more triplets than the 3′-UTRs in humans (31.1% vs. 4.6%; Wren et al. 2000); in barley (67% vs. 26%; Thiel et al. 2003), and in *Arabidopsis* (51.4% vs. 33.4%; Morgante, Hanafey, and Powell 2002). The 5′-UTRs also exhibit a strong bias toward specific triplet repeats in different mammalian genomes (Stallings 1994). For instance, of 136 triplet repeats identified in 5′-UTRs of human cDNA, 100 were CGG or CCG (Wren et al. 2000), serving as binding sites for nuclear proteins (Richards et al. 1993; Stallings 1994).

Intron SSR Distribution

Introns have a repeat-unit profile similar to that of genomic DNA: the majority of intronic SSRs are monomers, and/or dimers in different taxonomic groups or species (table 4).

Introns also display sequence composition biases in different repeat classes (Tóth, Gáspári, and Jurka 2000). For instance, a strong bias toward $(A/T)_n$ was found among monomer repeats in introns of different taxonomic groups and species (table 4). Among 12 possible dimeric repeats, introns show a bias (46.0%–67.5%) to AC/GT repeats in primates, rodentia, mammalia, vertebrata, arthropoda, and fungi, but a bias (40.8%–96.4%) to CG/GC repeats in *C. elegans*, embryophyta, and *Saccharomyces cerevisiae*. The functional significance of the above biases is unclear.

For triplet repeats in introns, ACG repeats are absolutely absent in primates, rodentia, mammalia, vertebrata, embryophyta, and fungi; and $(CCG)/(CGG)_n$ is almost completely absent from introns in all of the above taxonomic groups except fungi (Tóth, Gáspári, and Jurka 2000). The absence of CCG and ACG from introns could be explained by the presence of the highly mutable CpG dinucleotide within these motifs. Additionally, CCG repeats may also be selected against because of the requirements of the splicing machinery. Long CCG sequences could compete with this region in recruiting splicing machinery components, resulting in inadequate splicing. Furthermore, CCG repeats, which exhibit considerable hairpin-forming and quadruplex-forming potential, may influence the secondary structure of the pre-mRNA molecule, modulate the efficiency and accuracy of splicing, and then interfere with the formation of mature mRNA (Coleman and Roesser 1998; Tóth, Gáspári, and Jurka 2000).

Phenotypic Effect of SSRs from UTRs

The aforementioned data demonstrated that SSR distribution in the 5′-UTRs and 3′-UTRs and introns vary among species and/or taxonomic groups. The conservation of UTR repeats correlates inversely to their length, with longer repeats generally being more polymorphic than shorter repeats, irrespective of 3′-UTRs or 5′-UTRs (Suraweera et al. 2001). The UTR repeats are very often deleted in MSI-H tumors, and the average length of deletion within mononucleotide repeats in MSI-H tumors correlates strongly and positively with the length of the repeat regardless of their location in 5′-UTRs, introns, or 3′-UTRs (Suraweera et al. 2001). Some experiments demonstrated that the SSR repeat numbers located in 5′-UTRs, 3′-UTRs, and introns can regulate gene expression (see reviews: Kashi, King, and Soller 1997; Li et al. 2002; Trifonov 2003). More and more associations are currently being revealed between SSR variations in the UTRs and introns and phenotypic changes both in vitro and/or in vivo (see below).

Effects of SSR in 5′-UTR on Gene Expression and Phenotype

SSR elements in the 5′-UTRs are required for some gene expression. The human calmodulin-1 gene (hCALM 1) contains a stable $(CAG)_7$ repeat in its 5′-UTR (Toutenhoofd et al. 1998). Experiments have demonstrated that deleting this repeat decreased the gene expression by 45%, whereas repeat expansions to 20 and 45 repeats, or the insertion of a scrambled $(C, A, G)_7$ sequence did not alter gene expression (Toutenhoofd et al. 1998). These data indicate that the endogenous repeat element is required for full expression of hCALM1, and that some triplet repeat expansions in the 5′-UTR of protein-coding genes may be well tolerated and may even optimize the gene expression (Toutenhoofd et al. 1998). Expansion of the $(CTG)_n$ in the 5′-UTR of a reporter gene impeded efficient translation in vitro and in vivo, because the formation of stable hairpins by expansion of $(CUG)_n$ runs in the 5′-UTR of a mRNA progressively inhibit the scanning step of translation initiation (Raca et al. 2000).

SSR variations in 5′-UTRs influence gene expression and lead to protein adaptation. An SSR locus is found between the genes *hifA* and *hifB*, encoding fimbrial subunit proteins in *H. influenzae* (van Ham et al. 1993). Reversible phase variation is due to changes in the number of TA repeats, which space the −35 and −10 region of the dual promotor controlling *hifA* and *hifB*. This has a clear impact on transcriptional effectiveness (van Belkum et al. 1998). Streelman and Kocher (2002) reported that GT repeat polymorphisms in the *Tilapia* prolactin 1 (prl 1) 5′-UTR promoter are associated with differences in *prl 1* gene expression and the growth response of salt-challenged fishes. Individuals homozygous for long GT alleles express less prl 1 in fresh water but more prl 1 in half-seawater than fishes with other genotypes. This work provides in vivo evidence that differences in SSR length among individuals may indeed affect gene expression and that variance in expression has concomitant physiological

Table 4
Frequency (%) of SSRs in Nontranslated Regions of Genes

Repeat	*Arabidopsis*		Human			Intron								
Unit (bp)	5-UTR	3-UTR	5-UTR	3-UTR		Primates	Rodents	Mammals	Vertebrata	Arthropoda	*C. elegans*	Embrophyta	*S. cerevisiae*	Fungi
1	10.1	31.5	31.3	48.3		53.8	18.1	34.4	22.0	21.5	27.1	36.2	53.8	41.0
2	33.1	10.0	14.6	28.6		19.6	48.4	38.5	47.7	36.8	29.1	34.7	25.7	20.0
3	51.4	33.4	31.1	4.6		5.5	9.3	6.9	12.9	16.5	12.1	15.1	9.2	18.8
4	5.4	25.1	2.3	6.3		11.6	13.7	8.3	11.7	5.5	5.1	4.1	2.1	5.4
5	0.0	0.0	9.6	6.4		6.8	6.3	6.2	4.4	8.6	7.8	7.2	4.7	7.2
6	0.0	0.0	5.3	2.4		2.7	4.2	5.6	1.3	11.2	18.7	2.8	4.5	7.6
≥7			6.0	3.4										
Total SSR	634	373	438	1,659	bp[a]	7,690	12.052	5,718	6,697	4,420	1,886	3,813	5,600	5.061
					A	99.7	86.8	88.6	70.0	89.6	65.4	97.2	99.4	90.0
					C	0.3	13.2	11.4	30.0	10.4	34.6	2.8	0.6	10.0
					bp	4,127	2,182	1,957	1,476	950	512	1,380	3,012	2,075
					AC	67.2	64.9	61.2	67.5	46.0	27.5	12.7	1.8	50.4
					AG	17.8	29.8	31.7	12.9	28.8	31.5	33.6	1.8	13.8
					AT	14.7	5.1	5.6	19.6	25.1	40.8	53.7	96.4	35.7
					CG	0.3	0.2	1.5	0.0	0.0	0.0	0.0	0.0	0.0
					bp	1,506	5,837	2,202	3,193	1,627	549	1,322	1,437	1,013
Ref.	Morgante et al. (Wren et al. 2000)					Tóth, Gáspári, and Jurka (2000)								

[a] Repeat length (bp) per megabase of DNA.

consequences. Such results suggest that dinucleotide SSRs represent an underappreciated source of genetic variation for regulatory evolution (van Belkum et al. 1998).

SSRs in the 5′-UTRs serve as protein binding sites, thereby regulating gene translation and protein component and function. For instance, the transcription factor CCAAT/enhancer binding protein β (C/EBPβ) plays a significant role in the regulation of hepatocyte growth and differentiation. A single C/EBPβ mRNA encoded by an intronless gene produces several protein isoforms, presumably through alternative usage of AUG codon (Calkhoven et al. 1994). Experimental results show the presence of two CUG repeat binding sites in the 5′ region of C/EBPβ mRNA between the first and second AUG codons. One binding site contains CUG repeats in the 5′-UTR, whereas the other one contains CCG repeats in the ORF sequence, which has previously been shown to be involved in regulation of alternative translation of C/EBPβ mRNA (Calkhoven et al. 1994). Different RNA binding domains of a CUG repeat binding protein (CUGBP1) bind to both the CUG and CCG repeats. The binding of CUGBP1 to the 5′ region of C/EBPβ mRNA results in generation of low molecular weight C/EBPβ isoforms. It is possible that this interaction may stabilize a structure that favors translational initiation at downstream AUG codons (Timchenko et al. 1999).

SSRs in 5′-UTRs can regulate gene expression and phenotypically cause human diseases. The CGG repeat in 5′-UTR of the fragile X mental retardation-1 (FMR1) gene is expanded in families with fragile X syndrome, with $(CGG)_{\geq 200}$ resulting in mental retardation due to the absence of the FMR protein (FMRP), and significantly diminished FMRP levels in carriers is negatively correlated with the CGG repeat number (Kenneson et al. 2001). The $(CGG)_{40–200}$ are also related to different diseases (table 5).

SSR Expansions in 3′-UTRs Lead to Transcription Slippage

Transcription of a CAG/CTG triplet repeats in the 3′-UTR of a URA3 reporter gene in yeast leads to transcription of mRNA several kilobases longer than the expected size. These long mRNAs form more readily when CAG rather than CTG repeats are transcribed. These large mRNA molecules are formed by transcription slippage (Fabre, Dujon, and Richard 2002), a phenomenon reported to be usually stimulated by short mononucleotide and dinucleotide repeats at the 3′-UTR, both in vitro (Deng and Shuman 1997) and in vivo in *E. coli* (Xiong and Reznikoff 1993) and, possibly, in mammalian systems (Davis et al. 1997). It is generally assumed that during transcription, transient pausing of the RNA polymerase complex promotes backward slippage and leads to resynthesis of the same RNA sequence (Jacques and Kolakofsky 1991). CAG/CTG repeats form secondary structures in vitro (Gacy et al. 1995), and it was shown that CTG repeats induce transient pauses in RNA elongation after the first, second, and sixth to ninth triplet repeated in vitro unit (Parsons, Sinden, and Izban 1998). Stable in vivo secondary structures formed by mRNA are important for transcription slippage. Fabre, Dujon, and Richard (2002) propose that stalling of the RNA Pol II complex by secondary structures formed on the template DNA strand would promote slippage and give rise to longer transcripts containing reiterations of the repeated sequence. Several rounds of stalling/slippage/synthesis would probably be required to reach long mRNA.

What are the consequences of transcription slippage induced by SSR expansions at the 3′-UTR? The following examples should give us some clue. Myotonic dystrophy type 1 (DM1) is a dominant neuromuscular disorder caused by the expansion of CTG repeat in the 3′-UTR of the DM protein kinase (DMPK) genes (see review by Ranum and Day 2002). The expression of mutant DMPK mRNA containing an expanded repeat in the 3′-UTR inhibits differentiation of cultured myoblasts (Amack and Mahadevan 2001) and transgenic models in which $(CTG)_{>250}$ expressed at the RNA level, causes myotonia and muscular dystrophy (Seznec et al. 2001). The pathogenic effect of the CTG expansion in mRNA is to accumulate as nuclear foci (Davis et al. 1997; Mankodi et al. 2002) and to disrupt splicing and possibly other cellular functions (Mankodi et al. 2002). A recent experiment proved that expression of the mutant DMPK 3-UTR mRNA carrying long repeat as $(CTG)_{90}$ disrupted the nerve-growth-factor–induced Meurite outgrowth (Quintero-Mora et al. 2002). Moreover, expression of the CTG expansion has a negative *cis* effect on protein production, so that the mRNA with expanded CTG repeat in 3′-UTR has been considered to be a cause of suppression of neuronal differentiation (Quintero-Mora et al. 2002).

Intronic SSRs Affect Gene Transcription, mRNA Correct Splicing, or Export to Cytoplasm

Intronic SSRs can also regulate gene transcription. For instance, a $(TCAT)_n$ locus "HUMTH01" located in the first intron of the tyrosine hydroxylase (TH) gene acts as a transcription regulatory element in vitro (Meloni et al. 1998). Up to a maximum of eight repeats of this motif and its flanking region are highly conserved in the first intron of the TH gene of several nonhuman primate genera (Meyer et al. 1995), suggesting that evolutionary constraints may act on this sequence. The presence in the human population of a perfect and imperfect variant of $(TCAT)_{10}$ showing transcriptional regulatory activity is not due to genetic drift, but it may be relevant to the expression of normal and/or pathological genetic traits (Meloni et al. 1998). An intronic SSR can also behave as a co-regulator with SSRs in the 5′-UTR for gene expression. For example, the human type I collagen alpha2 (COL1A2) gene has one $(CA)_n$ repeat in the 5′-UTR and a $(GT)_n$ tract in its first intron. Experiment showed that transcriptional activity of the COL1A2 gene was enhanced by the co-presence of both repeats, but not by either repeat alone (Akai, Kimura, and Hata 1999). Intronic SSRs also can regulate gene expression level and lead to human diseases, such as Friedreich's ataxia (FRDA), SCA10, and breast carcinogenesis (table 5).

Table 5
SSRs Located in 5′- and 3′-UTRs And Introns and Their Functions and Phenotypic Effects in Humans and Bacteria

Motif	Location	Gene	SSR function and phenotypic effect	Reference
CAG	5′-UTR	Human calmodulin-1 gene (*hCALM1*)	Required for hCALM1 full expression,	Toutenhoofd et al. (1998)
A	5′-UTR	*M. hyorhinis* lipoprotein genes: *vlpA, vlpB, vlpC*	Affect transcription efficiency and are involved in antigen variation	van Belkum et al. (1998)
TA	5′-UTR	*H. influenzae hifA* and *hifB* encoding fimbrial subunit proteins	Influence gene expression and lead to protein adaptation and phase variation	van Belkum et al. (1998)
GT	5′-UTR	*Tilapia* prolactin 1 (prl 1)	Influence gene expression and growth response of salt-challenged fishes	Streelman and Kocher (2002)
CUG	5′-UTR	*C/EBPbeta*	Serve as protein-binding sites and regulating gene translation and protein component and function	Calkhoven et al. (1994) Timchenko et al. (1999)
CGG	5′-UTR	Fragile X mental retardation-1 (*FMR-1*)	(CGG)$_{>200}$ results in loss of FMR-1 function and cause human mental retardation	Kenneson et al. (2001)
			(CGG)$_{40–200}$ related in fragile-X-like cognitive/psychosocial impairment	Franke et al. (1998)
			(CGG)$_{40–60}$ associated with other fragile-X-like pheno types and woman ovarian dysfunction	Youings et al. (2000) Allingham-Hawkins et al. (1999)
GCC	5′-UTR	*FMR-2*	Reduced FMR2 causing abnormal neuronal gene regulation	Cummings and Zoghbi (2000)
CAG	5′-UTR	*PPP2R2B*	(CAG)$_{55–78}$ results in toxic effect at RNA level or alters gene expression and causes human SCA12 disease	O'Hearn et al. (2001)
CTG	3′-UTR	Dystrophia myotonin (DM) protein kinase	Expansion causes DM1 disease	Ranum and Day (2002)
CTG	3′-UTR	SCA8 gene	Expansion causes SCA8 disease	Koob et al. (1999)
TCAT	Intron	Tyrosine hydroxylase (TH) gene	Acts as transcription regulatory element and relevant to expression of pathogenesis	Meloni et al. (1998)
GAA	Intron	Friedreich's ataxia (FRDA) gene	GAA expansion inhibits *FRDA* expression or interferes mRNA formation and lead to FRDA disease	Ohshima et al. (1998) Sakamoto et al. (2001)
CA	Intron	Epidermal growth factor receptor (*egfr*)	CA repeat enhances *egfr* transcription and involved in breast carcinogenesis	Tidow et al. (2003)
T	Intron	ATM gene	Shortening repeat tract leads to aberrant splicing and abnormal transcription in colon tumor cells	Ejima, Yang, and Sasaki (2000)
CCTG	Intron	Zinc finger protein 9 (*ZNF9*)	Expansion causes nuclear retention of transcript and leads to DM2 disease	Liquori et al. (2001)
ATTCT	Intron	*SCA10* gene	Expansion leads to change of function and results in SCA10 disease	Matsuura et al. (2000)

Intronic polymorphism can result in abnormal splicing. The GGG repeats located at the 5′ end of human introns proved to be involved in splice site selection, and this GGG repeat may act as an enhancer of the splicing reaction at the level of spliceosome assembly (Sirand-Pugnet et al. 1995). Intronic splicing enhancers have been identified that can mediate tissue-specific exon inclusion. Although the intronic enhancers previously identified in genes generally exhibited a more complex sequence (see Gabellini 2001), in many cases some simple sequences such as SSRs were shown to act as intronic enhancers. For instance, the length polymorphism of $(TG)_m$ and $(T)_n$ repeats, both located at the intron 8/exon 9 splice acceptor site of the cystic fibrosis transmembrane conductance regulator (CFTR) gene, whose mutations cause cystic fibrosis (CF), control splicing of the CFTR mRNA (Pagani et al. 2000). The role for this polymorphic SSR site in the intron 8 of the CFTR gene is regulation of the alternative splicing and the skipping of exon 9 (Pagani et al. 2000). It has been found that short $(T)_5$, compared with $(T)_7$ or $_9$, significantly increases the alternative splicing of exon. Moreover, the polymorphic $(TG)_m$ locus juxtaposed upstream of the $(T)_n$ tract can further modulate exon 9 skipping, but only when activated by the $(T)_5$ allele (Niksic et al. 1999).

The human intron 2 of the Na^+/Ca^{2+} exchanger 1 (*NCX1*) gene contains GT repeats of variable length. Molecular dissection of the 5′-intron 2 sequence showed that the GT repeat is required for splicing activation. The remainder of the 5′-intron 2 segment was completely inactive. Moreover, the intron 2 segments with $(GT)_{10\text{–}16}$ are equally effective in promoting splicing. The minimal $(GT)_n$ length required to enhance splicing remains a question for future investigation (Gabellini 2001).

In previous studies of the cellular distribution of long triplet repeat-containing transcripts, it was shown that although the long transcripts were correctly spliced and polyadenylated, they were not correctly exported to the cytoplasm. $(CUG)_n$-containing mRNAs have been reported to accumulate in the nuclei of DM cultured myoblasts (Davis et al. 1997) and in the nuclei of muscle cells of transgenic mice expressing untranslated CUG repeats (Mankodi et al. 2002). The human DM2 is caused by the expansion of a CCTG repeat in the first intron of the zinc finger protein 9 (*ZNF9*) gene (Liquori et al. 2001). The longest normal allele observed was $(CCTG)_{26}$, whereas the expanded range is extremely broad (~75–11,000 CCTGs, mean ~5000) with a high degree of somatic mosaicism, and somatic instability is indicated by the multiple expansion sizes often found in blood (Liquori et al. 2001). Like the DM1, $(CCUG)_n$-containing RNA foci are found in DM2 muscle, indicating nuclear retention of transcripts (Liquori et al. 2001).

SCA10 disease is caused by expansion of an ATTCT motif located in intron 9 of the *SCA10* gene (Matsuura et al. 2000). The expanded *SCA10* gene encodes a novel protein (475 amino acids) with no recognizable motifs or predicted structures (Matsuura et al. 2000). Parallels between SCA10, DM1, DM2, and, possibly, SCA8 suggest that a gain of function mechanism at the RNA level could be involved (Ranum and Day 2002). Unlike the SCA8, the gene harboring the mutation for SCA10 is ubiquitously expressed, indicating that if a toxic RNA mechanism is involved, secondary proteins that interact with the ATTCT motif may confer organ-specific pathogenicity (Matsuura et al. 2000).

SSR Expansion in UTRs and Introns Leads to Gene Silencing or a Loss of Function

Gene expression is crucial to the maintenance of differentiated cell types in multicellular organisms, whereas aberrant silencing can lead to disease. The organization of DNA into euchromatin and heterochromatin is implicated in gene silencing. In euchromatin, DNA wraps around histones, creating nucleosomes. Further condensation of chromatin, associated with large blocks of repetitive DNA sequences, is known as heterochromatin. Position effect variegation (PEV) occurs when a gene is located abnormally close to heterochromatin silencing the affected gene in a proportion of cells (Dillon and Festenstein 2002). It has been reported that the relatively short GAA or CTG repeat expansions found in the 3′-UTR of DM1 and the intron of the FRDA gene mediate heterochromatin-protein-1–sensitive variegated gene silencing on a linked transgene in mice (Saveliev et al. 2003). Silencing was correlated with a decrease in promoter accessibility and was enhanced by the classical PEV modifier heterochromatin protein 1 (HP1). Notably, triplet repeat-associated variegation was not restricted to the classical heterochromatic region, but it occurred irrespective of chromosomal location (Saveliev et al. 2003). Because this phenomenon shares important features with PEV, the mechanisms underlying heterochromatin-mediated silencing might have a role in gene regulation at many sites throughout the mammalian genome. They may modulate the extent of gene silencing, and hence disease severity in several triplet diseases such as DM1 and FRDA, among others (Saveliev et al. 2003).

SSR Evolutionary Mechanism Within Genes

The SSR distribution is a function of the dynamics and history of genome evolution and of selective constraints (Morgante, Hanafey, and Powell 2002). Like the SSRs in untranscribed regions, the SSRs in genes also show a higher mutation rate (instability) than nonrepetitive regions. The fact that in human genes the most frequently encountered polymorphism is the repeat length polymorphism, and that it has its roots in repeat elongation/shortening events, indicates that such processes are important ingredients of molecular evolution (Borštnik and Pumpernik 2002). The repeat elongation/shortening processes also lead to the increase of biological complexity, which is considered to be the hallmark of biological evolution. In a variety of widely diverged eukaryotes, including *S. cerevisiae*, *S. pombe*, *C. elegans*, *Drosophila*, plants, primates, and *Mus* the rate of generation of excess triplet SSRs is not significantly affected by coding status, suggesting that both coding and noncoding triplet SSRs are subjected to similar rates of repeat expansion (Metzgar, Byfot, and Wills 2000). SSR evolution in coding genes and regulatory regions should share mutational processes similar to those of SSRs in untranscribed regions.

Molecular Mutation Mechanism

Previously, a few mutational mechanisms have been invoked to explain the high mutation rate of SSR:DNA slippage during DNA replication (Tachida and Iizuka 1992) and recombination (unequal crossover and gene conversion) between DNA strands (Harding, Boyce, and Clegg 1992; see also our review: Li et al. 2002). Recently, some studies have suggested that equilibrium distributions of SSR repeat lengths are a result of balance between slippage events and point mutation (Kruglyak et al. 1998; Ellegren 2002). Replication slippage favors growth, whereas point mutations break down a long repeat array into two or more shorter ones. Changes in the relative frequencies of slippage and point mutation might have direct effect on the accumulation of long SSR runs in genomes, as the length distribution of SSRs in a genome has been suggested to reflect this balance, a higher relative rate of slippage giving rise to longer SSRs (Kruglyak et al. 1998). Relative rates of slippage and point mutation might be altered by changes in the efficiency of MMR and proofreading during DNA replication (Ellegren 2002) and other potential differences in genome structure or organization between species (Alba, Santibáñez-Koref, and Hancock 2001). Among those factors, MMR efficiency is particularly critical for SSR slippage rates, because there are known to be differences between species in the abilities of their MMR machinery to detect loops of different length resulting from slippage during replication (Parniewski et al. 2000; Yamada et al. 2002*b*). In addition, MMR genes can be inactivated by frameshift via the $(T)_n$ located in their coding sequences, and lead to more MSI (see above, *SSRs In Coding Regions*). In molecular evolution, MMR activity is critical for SSR length fluctuation. In humans, any inactivation of MMR genes could cause tremendous MSI and phenotypically occurring cancers or other diseases (see above, *SSRs In Coding Regions*).

Role of Natural Selection

Biased SSR Distribution in Genes

Purely neutral structures would be expected for SSRs to be randomly distributed within genes including coding regions, UTRs, and introns. Or, SSR distributions would reflect the balance of replication slippage and point mutation without external forces (Kruglyak et al. 1998; Santibáñez-Koref, Gangeswaran, and Hancock 2001; Ellegren 2002). In fact, however, SSR distributions in these regions are nonrandom and strongly biased. The low frequency of dinucleotide and tetranucleotide repeats and the enhanced frequency of triplet repeats in the coding sequences of many organisms (Tóth, Gáspári, and Jurka 2000; Wren et al. 2000; Cordeiro et al. 2001) are signs of the effects of selection, indicating that those SSRs are selected against possible frameshift mutation. The action of selection on triplet repeats is best seen by considering their distribution between the different strands and reading frames within ORFs, and between *coding* and *noncoding* regions of the genome. Numerous investigations have indicated that triplet repeats show strong reading frame and strand preferences (e.g., Richard and Dujon 1997; Alba, Santibáñez-Koref, and Hancock 1999) caused by selection. Strongly biased distributions of triplet repeats and amino acid repeats have also been found in different functional protein groups and cell locations (see above, *SSRs In Coding Regions*), suggesting that repeats of these kinds are subject to strong selection (Alba, Santibáñez-Koref, and Hancock 1999).

Alba, Santibáñez-Koref, and Hancock (1999) revealed intriguing associations between the most common amino acid repeats and cellular components which appear to be part of the cell-signaling system, and it has been speculated that changes in length of repeats in such systems could alter their behavior and therefore contribute to their evolutionary diversification (Hancock 1993; Richard and Dujon 1997), perhaps involving molecular coevolution between proteins (Hancock 1993). Such diversification could be relatively rapid on an evolutionary time scale because of the high mutation rate of SSRs (Hancock 1999). It is likely that genes containing SSRs have been found to be mutated at these repeats and result in frameshift in MSI-H tumors from different sites (see review by Duval and Hamelin 2002). Accumulation of such alterations appears to be the molecular mechanism by which MSI-H cells accumulate functional changes with putative oncogenic effects. These mutations occur at variable frequencies in many genes encoding protein signal transduction, transformation growth factors, apoptosis, MMR, transcriptional regulation, or immune surveillance cell-cycle response to DNA damage (table 3). These mutations can influence genes with a putative role in human carcinogenesis involved in different or similar pathways and are thus thought to be affected by *inactivating* or *activating* events selected for in these cancers in a recessive or dominant manner (Duval and Hamelin 2002).

Natural Selection and SSR Function in Genes

When SSR repeats lie within protein coding regions, UTRs, and introns, any changes by replication slippage and other mutational mechanisms may lead to changes in protein function. There are numerous lines of evidence (see above, *SSRs In Coding Regions* and *SSRs in UTR Regions*) indicating that changes in lengths of triplet or amino acid repeats could affect protein function, and frameshifts within coding regions caused by SSR expansion or contraction could (1) cause gain of function and loss of function or gene silencing and (2) induce novel protein, bacterial pathogenesis, and virulence. Variations in repeat number of SSR located in the 5′-UTRs and 3′-UTRs and introns can cause significant effects on gene expression—e.g., mRNA splicing or translation—and lead to phenotypic changes with altered selective values (see above, *SSR in UTR Regions*). For instance, in *Escherichia coli*, hundreds of genes related to DNA repair, recombination, and physiological adaptation to different stresses contain high density of small SSRs, which can induce mutator phenotypes by affecting repair efficiency and/or DNA metabolism (Rocha, Matic, and Taddei 2002). Genes containing $(Gln)_n$ that are more conserved in length between the human and mouse tend to show low nonsynonymous substitution rates, while genes containing more evolutionarily labile repeats, tend to have higher substitution rates (Hancock, Worthey, and Santibáñez-Koref 2001). This indicates that the level of selection acting on a gene containing a repeat has a significant impact on DNA repeat evolution. Effect of selection could include disfavoring uninterrupted structures, as these are more likely to change in length with phenotypic consequences such as triplet expansion disease in humans (Alba, Santibáñez-Koref, and Hancock 2001).

Environmental Stress and SSR-Regulated Adaptation

An organism has to adapt to environmental change for its survival. A certain degree of stress caused by fluctuations in the environment is a necessary starting point for every adaptational change. Earlier investigations speculated that eukaryotes incorporating more DNA repeats might provide a molecular device for faster adaptation to environmental stresses (Kashi, King, and Soller 1997; Marcotte et al. 1999; Wren et al. 2000; Li et al. 2002; Trifonov 2003). This speculation has been supported by an increasing number of experiments. For instance, in *S. cerevisiae*, SSRs are overrepresented among ORFs encoding for regulatory proteins (e.g., transcription factors and protein kinases) rather than for structural ones, indicating the role of SSRs as a factor contributing to *fast evolution of adaptive phenotypes* (Young, Sloan, and van Riper 2000). Although in prokaryotes SSRs are not so abundant as in eukaryotes, most of the SSRs in bacteria are located in virulence genes and/or regulatory regions, and they affect pathogenesis and bacterial adaptive behavior, indicating the signature of natural selection (Hood et al. 1996; Peak et al. 1996; van Belkum et al. 1998; Field and Wills 1998). The contingency genes containing SSRs show high mutation rates, allowing the bacterium to act swiftly on deleterious environmental conditions (Moxon et al. 1994). As in eukaryotes, SSR variations enable bacteria to respond to diverse environmental factors.

Trifonov (2003) suggests that environmental challenges cause various stress reactions, including changes in the copy number in the tandem runs. This suggestion is supported by the fact that DNA damage caused by external stresses such as UV irradiation, γ–irradiation, t-butyl hydrogen peroxide, oxidative damage, etc. can induce slippage mutations and increase mutation rates in SSR sequences (Jackson, Chen, and Loeb 1998; Chang et al. 2002; Slebos et al. 2002). An immediate response to environmental challenges is retuning of the expression of many genes and multigene functions influenced by the repeats. The copy numbers of the variable tuners linked to specific genes most relevant to given environmental changes are under selection pressure. Accordingly, some of the copy-number tuning responses to the stress result in observable phenotypic changes (Trifonov 2003).

Trifonov (2003) reviewed tuning functions of SSRs and minisatellites for a number of genes in eukaryotes and prokaryotes, suggesting that tuning by the repeat copy number is a rather general phenomenon. The SSRs and minisatellites may act as "tuning knobs" (King, Soller, and Kashi 1997; Trifonov 2003) for modulation of gene expression or other functions as gradually as discrete numbers of the repeats in the tandem runs would allow. The larger the number of the repeats in the run, and the weaker the influence of any individual repeat, the finer the tuning. The data actually represent the middle part of the causal route from environmental changes to gross phenotypic response—expansion/contraction of the repeats and respective changes in gene expression patterns, leaving out the initial stress component and the final stage of the organismal phenotype change. The genes themselves (classical genotype) remain unchanged (Trifonov 2003).

Copy number of SSR repeats could influence the phenotype, both at the moment of the change and after sexual spread of the new change. A particular mechanism of the molecular change could be any intracellular DNA turnover mechanism including the tuning by tandem repeats. The molecular drive concept, suggested for multigene families, is applicable as well to any other repeating sequences involved in regulation of gene expression (Dover 1986). The DNA turnover, however, can be viewed as another kind of mutation (resulting in deletions and insertions of the repeats), in which case the copy number-phenotype relationship would be a simple case of Darwinian selection (Trifonov 2003). The specific adaptive response can only be established by a selection process spanning several sexual generations or, at the cellular level, several cellular divisions. The changes of tandem repeat tuners for adaptationally competent genes cause the adaptive response and are, thus, under selection pressure. In other words, the changes in the tuners of relevant genes in desirable directions and respective changes of the gene activities toward relaxation of the stress impact are selected for. The selection pressure thus may result in systematic directional change of the respective repeat number that leads, finally, to desirable activity levels of the adaptationally relevant genes and relaxation of the stress—i.e., adaptation (Trifonov 2003). Future studies could increasingly unravel the significant evolutionary role of SSRs in regulating gene expression under diverse environmental stresses.

Conclusions

It has been demonstrated that SSRs are much more abundant in the UTRs or regulatory regions than in other genomic regions of plants (Morgante, Hanafey, and Powell 2002) and bacteria (Moxon et al. 1994; Field and Wills 1998; Metzgar, Bytof, and Wills 2000). Substantial evidence shows that SSRs are *nonrandomly* distributed across protein-coding sequences, UTRs, and introns. SSR variations in these regions could cause a frameshift, a fluctuation of gene expression, inactivation of gene activity, and/or a change of function, and eventually phenotypic changes (fig. 1). In humans, SSR variation in coding regions, UTRs, and introns can cause neuronal diseases, cancers, SCA, and DM diseases, among others. In some cases, MSI even affects the effectiveness of medical treatment on human cancers (Kim et al. 2001), even though more studies should be conducted for clarification and efficient utilization of this mechanism. In plants and other species, outcomes of SSR variation within their genes remain to be studied further, despite the large number of studies reporting the SSR distribution in their ESTs or genes.

In bacteria, particularly pathogenic bacteria, infection processes require that the bacteria adapt to several host environments. Initial colonization, crossing epithelial and endothelial barriers, survival in circulation, and trans-

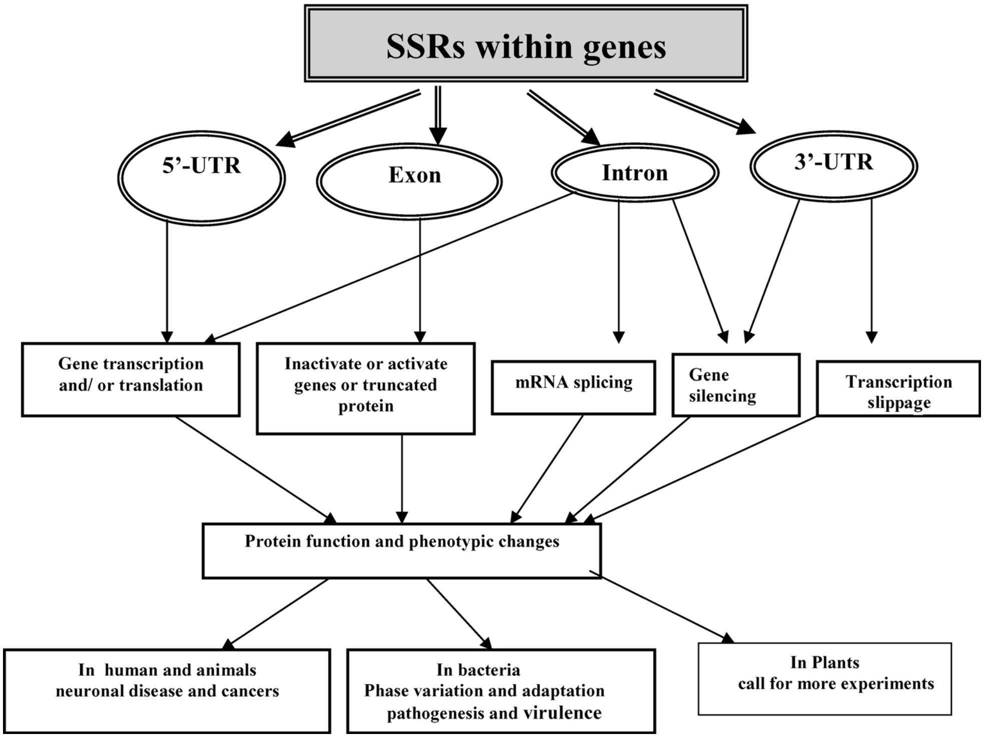
SSRs within genes
5'-UTR
Exon
Intron
3'-UTR
Gene transcription and/ or translation
Inactivate or activate genes or truncated protein
mRNA splicing
Gene silencing
Transcription slippage
Protein function and phenotypic changes
In human and animals neuronal disease and cancers
In bacteria Phase variation and adaptation pathogenesis and virulence
In Plants call for more experiments

FIG. 1.—SSR regulatory functions within exons, 3′-UTRs, 5′-UTRs, and introns.

location across, for instance, the blood-brain barrier, are all processes that require specific virulence traits (Roche and Moxon 1995). The possibility of varying pathogenicity factors to meet these requirements could possibly be achieved through SSR modulation, as has been described for a multitude of different genes. Variation through single-strand mispairing or recombination processes allows regulatory or adaptive function to be specifically activated or repressed (van Belkum et al. 1998).

SSR evolution in genes should share similar mutational processes, including replication slippage, point mutation, and recombination, but SSRs within genes should be subjected to stronger selection pressure than other regions because of their functional significance in regulating gene expression and function. These mutational processes provide mutation resources for the MMR system. If SSR mutations within genes escape from MMR correction, these mutations can cause phenotypic changes. The link between changing copy number of SSRs and phenotypes is provided by an accumulating number of experimental observations showing a dependence of gene expression and other functions on the copy number of the associated repeats (see above, *SSRs In Coding Regions* and *SSR in UTR Regions*). If SSR changes result in selectable phenotypic variation, selection can naturally start to act. It has been demonstrated that SSRs in protein-coding regions are under strong selection (Richard and Dujon 1997; Alba, Santibáñez-Koref, and Hancock 1999). The presence and variation of SSRs in upstream regulatory elements might affect the expression of ORFs in either an on/off or a quantitative manner. Morgante, Hanafey, and Powell (2002) estimated that in the *Arabidopsis* genome 5′-UTRs are under very strong positive selection, maintaining the optimum repeat number of SSRs with an almost threefold higher frequency than at any other genomic region; repeats from the 3′-UTRs are under moderate positive selection. In other regulatory regions such as introns, SSRs are also under selection pressure that keeps SSRs in a proper size range, at least for those functional loci. In sum, we have shown that SSRs within genes are substantially involved in regulatory evolutionary processes.

Acknowledgments

We thank Prof. E. Trifonov for his constructive comments on this review, and Robin Permut for editing the manuscript.

Literature Cited

Akai, J., A. Kimura, and R. I. Hata. 1999. Transcriptional regulation of the human type I collagen alpha2 (COL1A2) gene by the combination of two dinucleotide repeats. Gene **239**:65–73.

Alba, M. M., M. F. Santibáñez-Koref, and J. M. Hancock. 1999. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. J. Mol. Evol. **49**:789–797.

———. 2001. The comparative genomics of polyglutamine repeats: extreme difference in the codon organization of repeat-encoding region between mammals and *Drosophila*. J. Mol. Evol. **52**:249–259.

Allingham-Hawkins, D. J., R. Babul-Hirji, D. Chitayat et al. (>25 coauthors). 1999. Fragile X premutation is a significant risk factor for premature ovarian failure: the International

Collaborative POF in Fragile X study—preliminary data. Am. J. Med. Genet. **83**:322–325.

Amack, J. D., and M. S. Mahadevan. 2001. The myotonic dystrophy expanded CUG repeat tract is necessary but not sufficient to disrupt C2C12 myoblast differentiation. Hum. Mol. Genet. **10**:1879–1887.

Bayliss, C. D., T. van de Ven, and E. R. Moxon. 2002. Mutations in polI but not mutSLH destabilize *Haemophilus influenzae* tetranucleotide repeats. EMBO J. **21**:1465–1476.

Bicknell, D. C., L. Kaklamanis, R. Hampson, W. F. Bodmer, and P. Karran. 1996. Selection for beta 2-microglobulin mutation in mismatch repair-defective colorectal carcinomas. Curr. Biol. **6**:1695–1697.

Borštnik, B., and D. Pumpernik. 2002. Tandem repeats in protein coding regions of primate genes. Genome Res. **12**:909–915.

Brais, B., J. P. Bouchard, Y. G. Xie et al. (19 co-authors). 1998. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat. Genet. **18**:164–167.

Buchanan, G., R. A. Irvine, G. A. Coetzee, and W. D. Tilley. 2001. Contribution of the androgen receptor to prostate cancer predisposition and progression. Cancer Metastasis Rev. **20**:207–233.

Burch, C. L., R. J. Danaher, and D. C. Stein. 1997. Antigenic variation in *Neisseria gonorrhoeae*: production of multiple lipooligosaccharides. J. Bacteriol. **79**:982–986.

Calkhoven, C. F., P. R. Bouwman, L. Snippe, and G. Ab. 1994. Translation start site multiplicity of the CCAAT/enhancer binding protein alpha mRNA is dictated by a small 5′ open reading frame. Nucleic Acids Res. **22**:5540–5547.

Chang, C. L., G. Marra, D. P. Chauhan, H. T. Ha, D. K. Chang, L. Ricciardiello, A. Randolph, J. M. Carethers, and C. R. Boland. 2002. Oxidative stress inactivates the human DNA mismatch repair system. Am. J. Physiol. Cell. Physiol. **283**:C148–C154.

Coetzee, G., and R. Irvine. 2002. Size of the androgen receptor CAG repeat and prostate cancer: does it matter? J. Clin. Oncol. **20**:3572–3573.

Coleman, T. P., and J. R. Roesser. 1998. RNA secondary structure: an important *cis*-clement in rat calcitonin/CGRP pre-messenger RNA splicing. Biochemistry **37**:15941–15950.

Cordeiro, G. M., R. Casu, C. L. McIntyre, J. M. Manners, and R. J. Henry. 2001. Microsatellite markers from sugarcane (*Saccharum* spp) ESTs across transferable to erianthus and sorghum. Plant Sci. **160**:1115–1123.

Cummings, C. J., and H. Y. Zoghbi. 2000. Trinucleotide repeats: mechanisms and pathophysiology. Annu. Rev. Genomics Hum. Genet. **1**:281–328.

Davis, B. M., M. E. McCurrach, K. L. Taneja, R. H. Singer, and D. E. Housman. 1997. Expansion of a CUG trinucleotide repeat in the 3′ untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts. Proc. Natl. Acad. Sci. USA. **94**:7388–7393.

Dejager, S., H. Bry-Gauillard, E. Bruckert, B. Eymard, F. Salachas, E. LeGuern, S. Tardieu, R. Chadarevian, P. Giral, and G. Turpin. 2002. A comprehensive endocrine description of Kennedy's disease revealing androgen insensitivity linked to CAG repeat length. J. Clin. Endocrinol. Metab. **87**:893–901.

Deng, L., and S. Shuman. 1997. Elongation properties of vaccinia virus RNA polymerase: pausing, slippage, 3′ end addition, and termination site choice. Biochemistry **36**:15892–15899.

Dillon, N., and R. Festenstein. 2002. Unravelling heterochromatin: competition between positive and negative factors regulates accessibility. Trends Genet. **18**:252–258.

Dover, G. 1986. Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet. **2**:59–165.

Duval, A., J. Gayet, X. P. Zhou, B. Iacopetta, G. Thomas, and R. Hamelin. 1999. Frequent frameshift mutations of the TCF-4 gene in colorectal cancers with microsatellite instability. Cancer Res **59**:4213–4215.

Duval, A., and R. Hamelin. 2002. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer Res. **62**:2447–2454.

Duval, A., S. Rolland, E. Tubacher, H. Bui, G. Thomas, and R. Hamelin. 2000. The human T-cell transcription factor-4 gene: structure, extensive characterization of alternative splicings, and mutational analysis in colorectal cancer cell lines. Cancer Res. **60**:3872–3879.

Edwards, Y. J., G. Elgar, M. S. Clark, and M. J. Bishop, 1998. The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: perspectives in functional and comparative genomic analyses. J. Mol. Biol. **278**:843–854.

Ejima, Y., L. Yang, and M. S. Sasaki. 2000. Aberrant splicing of the ATM gene associated with shortening of the intronic mononucleotide tract in human colon tumor cell lines: a novel mutation target of microsatellite instability. Int. J. Cancer **86**:262–268.

Ellegren, H. 2002. Microsatellite evolution: a battle between replication slippage and point mutation. Trends Genet. **18**:70.

Fabre, E., B. Dujon, and G. F. Richard. 2002. Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. Nucleic Acids Res. **30**:3540–3547.

Ferro, P., R. dell'Eva, and U. Pfeffer. 2001. Are there CAG repeat expansion-related disorders outside the central nervous system? Brain Res. Bull. **56**:259–264.

Field, D., and C. Wills. 1996. Long, polymorphic microsatellites in simple organisms. Proc. R. Soc. London Ser. B. **263**:209–251.

———. 1998. Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. Proc. Natl. Acad. Sci. USA **95**:1647–1652.

Franke, P., M. Leboyer, M. Gansicke, et al. (12 coauthors). 1998. Genotype-phenotype relationship in female carriers of the premutation and full mutation of FMR-1. Psychiatry Res. **80**:113–127.

Gabellini, N. 2001. A polymorphic GT repeat from the human cardiac Na^+ Ca^{2+} exchanger intron 2 activates splicing. Eur. J. Biochem. **268**:1076–1083.

Gacy, A. M., G. Goellner, N. Juranic, S. Macura, and C. T. McMurray. 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell **81**:533–540.

Galvão, R., L. Mendes-Soaresa, J. Câmaraa, I. Jacoa, and M. Carmo-Fonseca. 2001. Triplet repeats, RNA secondary structure and toxic gain-of-function models for pathogenesis. Brain Res. Bull. **56**:191–201.

Grimwood, J., L. Olinger, and R. S. Stephens. 2001. Expression of *Chlamydia pneumoniae* polymorphic membrane protein family genes. Infect Immun. **69**:2383–2389.

Guanti, G., N. Resta, C. Simone, F. Cariola, I. Demma, P. Fiorente, and M. Gentile. 2000. Involvement of PTEN mutations in the genetic pathways of colorectal cancerogenesis. Hum. Mol. Genet. **9**:83–87.

Hancock, J. M. 1993. Evolution of sequence repetition and gene duplication in the TATA-binding protein TBP (TFIID). Nucleic Acids Res. **21**:2823–2830.

———. 1999. Microsatellites and other simple sequences: genomic context and mutational mechanisms. Pp. 1–9 *in*

D. B. Goldstein and C. Schlötterer, eds., Microsatellites: evolution and applications. Oxford University Press, Oxford, U.K.

Hancock, J. M., E. A. Worthey, and M. F. Santibáñez-Koref. 2001. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. Mol. Biol. Evol. **18**:1014–1023.

Harding, R. M., A. J. Boyce, and J. B. Clegg. 1992. The evolution of tandemly repetitive DNA: recombination rules. Genetics **132**:847–859.

Hood, D. W., M. E. Deadman, M. P. Jennings, M. Bisercic, R. D. Fleischmann, J. C. Venter, and E. R. Moxon. 1996. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. Proc. Natl. Acad. Sci. USA **93**:11121–11125.

Inzana, T. J., J. Hensley, J. McQuiston, A. J. Lesse, A. A. Campagnari, S. M. Boyle, and M. A. Apicella. 1997. Phase variation and conservation of lipooligosaccharide epitopes in *Haemophilus influenzae*. Infect. Immun. **65**:4675–4681.

Jacques, J. P., and D. Kolakofsky. 1991. Pseudo-templated transcription in prokaryotic and eukaryotic organisms. Genes Dev. **5**:707–713.

Jackson, A. L., R. Chen, and L. A. Loeb. 1998. Induction of microsatellite instability by oxidative DNA damage. Proc. Natl. Acad. Sci. USA **95**:12468–12473.

Johannsdottir, J. T., J. G. Jonasson, J. T. Bergthorsson, L. T. Amundadottir, J. Magnusson, V. Egilsson, and S. Ingvarsson. 2000. The effect of mismatch repair deficiency on tumourigenesis; microsatellite instability affecting genes containing short repeated sequences. Int. J. Oncol. **16**:133–139.

Jurka, J., and C. Pethiyagoda. 1995. Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol. **40**:120–126.

Kantety, R. V., M. La Rota, D. E. Matthews, and M. E. Sorrells. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol. **48**:501–510.

Karlin, S., J. Mrazek, and A. M. Campbell. 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. **179**:3899–3913.

Kashi, Y., D. King, and M. Soller. 1997. Simple sequence repeats as a source of quantitative genetic variation. Trends Genet. **13**:74–78.

Katti, M. V., P. K. Ranjekar, and V. S. Gupta. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. **18**:1161–1167.

Kenneson, A., F. Zhang, C. H. Hagedorn, and S. T. Warren. 2001. Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. Hum. Mol. Genet. **10**:1449–1454.

Kim, G. P., L. Colangelo, C. Allegra, et al. 2001. Prognostic role of microsatellite instability in colon cancer. Proc. Am. Soc. Clin. Oncol. **20**:1666.

King, D. G., M. Soller, and Y. Kashi. 1992. Evolutionary tuning knobs. Endevor **21**:36–40.

Koob, M. D., M. L. Moseley, L. J. Schut, K. A. Benzow, T. D. Bird, J. W. Day, and L. P. Ranum. 1999. An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). Nat. Genet. **21**:379–384.

Kruglyak, S., R. T. Durrett, M. D. Schug, and C. F. Aquadro. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutation. Proc. Natl. Acad. Sci. USA **95**:10774–10778.

Li, Y. C., A. B. Korol, T. Fahima, A. Beiles, and E. Nevo. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol. Ecol. **11**:2453–2465.

Liquori, C. L., K, Ricker, M. L. Moseley, J. F. Jacobsen, W. Kress, S. L. Naylor, J. W. Day, and L. P. Ranum. 2001. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science **293**:864–867.

Liu, Z., P. Li, K. Kocabas, A. Karsi, and Z. Ju. 2001. Microsatellite-containing genes from the channel catfish brain: evidence of trinucleotide repeat expansion in the coding region of nucleotide excision repair gene RAD23B. Biophys. Res. Commun. **289**:317–324.

Makino, S., J. P. M. van Putten, and T. F. Meyer. 1991. Phase variation of the opacity outer membrane protein controls invasion by *N. gonorhoeae* into human epithelial cells. EMBO J. **10**:1307–1315.

Mankodi, A., M. P. Takahashi, H. Jiang, C. L. Beck, W. J. Bowers, R. T. Moxley, S. C. Cannon, and C. A. Thornton. 2002. Expanded CUG repeats trigger aberrant splicing of ClC-1 chloride channel pre-mRNA and hyperexcitability of skeletal muscle in myotonic dystrophy. Mol. Cell. **10**:35–44.

Mao, R., A. S. Aylsworth, N. Potter et al. (11 co-authors). 2002. Childhood-onset ataxia: testing for large CAG-repeats in SCA2 and SCA7. Am. J. Med. Genet. **110**:338–345.

Marcotte, E. M., M. Pellegrini, T. O. Yeates, and D. Eisenberg. 1999. A census of protein repeats. J. Mol. Biol. **293**:151–160.

Markowitz, S., J. Wang, L. Myeroff et al. (>10 co-authors). 1995. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science **268**:1336–1338.

Mark Redston, M. D. 2001. Carcinogenesis in the GI tract: from morphology to genetics and back again. Mod. Pathol. **14**:236–245.

Matsuura, T., T. Yamagata, D. L. Burgess et al. (17 co-authors). 2000. Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. Nat. Genet. **26**:191–194.

Meloni, R., V. Albanese, P. Ravassard, F. Treilhou, and J. Mallet. 1998. A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. Hum. Mol. Genet. **7**:423–428.

Metzgar, D., J. Bytof, and C. Wills. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. **10**:72–80.

Meyer, E., P. Wiegand, S. P. Rand, D. Kuhlmann, M. Brack, and B. Brinkmann. 1995. Microsatellite polymorphisms reveal phylogenetic relationships in primates. J. Mol. Evol. **41**:10–14.

Moran, C. 1993. Microsatellite repeats in pig (*Sus domestica*) and chicken (*Gallus domesticus*) genomes. J. Hered. **84**:274–280.

Morgante, M., M. Hanafey, and W. Powell. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet. **30**:194–200.

Moxon, E. R., P. B. Rainey, M. A. Nowak, and R. E. Lenski. 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. **4**:24–33.

Murphy, G. L., T. D. Connell, D. S. Barritt, M. Koomey, and J. G. Cannon. 1989. Phase variation of gonococcal protein II: regulation of gene expression by slipped strand mispairing of a repetitive DNA sequences. Cell **56**:539–547.

Nakamura, K., S. Y. Jeong, T. Uchihara, M. Anno, K. Nagashima, T. Nagashima, S. Ikeda, S. Tsuji, and I. Kanazawa. 2001. SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. Hum. Mol. Genet. **10**:1441–1448.

Niksic, M., M. Romano, E. Buratti, F. Pagani, and F. E. Baralle. 1999. Functional analysis of cis-acting elements regulating the alternative splicing of human CFTR exon 9. Hum. Mol. Genet. **8**:2339–2349.

Ohshima, K., L. Montermini, R. D. Wells, and M. Pandolfo. 1998. 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**:14588–14595.

Olivero, M., T. Ruggiero, N. Coltella, A. Maffe, R. Calogero, E. Medico, and M. F. Di Renzo. 2003. Amplification of repeat-containing transcribed sequences (ARTS): a transcriptome fingerprinting strategy to detect functionally relevant microsatellite mutations in cancer. Nucleic Acids Res. **31**:e33.

O'Hearn, E., S. E. Holmes, P. C. Calvert, C. A. Ross, and R. L. Margolis. 2001. SCA-12: Tremor with cerebellar and cortical atrophy is associated with a CAG repeat expansion. Neurology **56**:299–303.

Pagani, F., E. Buratti, C. Stuani, M. Romano, E. Zuccato, M. Niksic, L. Giglio, D. Faraguna, and F. E. Baralle. 2000. Splicing factors induce cystic transmembrane regulator exon 9 skipping through a nonevolutionary conserved intronic element. J. Biol. Chem. **275**:210141–210147.

Parniewski, P., A. Jaworski, R. Wells, and R. Bowater. 2000. Length of CTG CAG repeats determines the influence of mismatch repair on genetic instability. J. Mol. Biol. **299**:865–874.

Parsons, M. A., R. R. Sinden, and M. G. Izban. 1998. Transcriptional properties of RNA polymerase II within triplet repeat-containing DNA from the human myotonic dystrophy and fragile X loci. J. Biol. Chem. **273**:26998–27008.

Peak, I. R. A., M. P. Jennings, D. W. Hood, M. Bisercic, and E. R. Moxon. 1996. Tetrameric repeat units associated with virulence factor phase variation in *Haemophilus* also occur in *Neiserria* spp. and *Moraxella catarrhalis*. FEMS Microbiol. Lett. **137**:109–114.

Quintero-Mora, M. L., F. Depardon, J. Waring, R. G. Korneluk, and B. Cisneros. 2002. Expanded CTG repeats inhibit neuronal differentiation of the PC12 cell line. Biochem. Biophys. Res. Commun. **295**:289–294.

Raca, G., E. Y. Siyanova, C. T. McMurray, and S. M. Mirkin. 2000. Expansion of the $(CTG)_n$ repeat in the 5′-UTR of a reporter gene impedes translation. Nucleic Acids Res. **28**:3943–3949.

Rampino, N., H. Yamamoto, Y. Ionov, Y. Li, H. Sawai, J. C. Reed, and M. Perucho. 1997. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science **275**:967–969.

Ranum, L. P., and J. W. Day. 2002. Dominantly inherited, non-coding microsatellite expansion disorders. Curr. Opin. Genet. Dev. **12**:266–271.

Richard, G.–F., and B. Dujon. 1997. Trinucleotide repeats in yeast. Res. Microbiol. **148**:731–744.

Richards, R. I., K. Holman, S. Yu, and G. R. Sutherland. 1993. Fragile X syndrome unstable element, $p(CCG)_n$, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. Hum. Mol. Genet. **2**:1429–1435.

Rocha, E. P. C., and N. Blanchard. 2002. Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res. **30**:2031–2042.

Rocha, E. P., I. Matic, and F. Taddei. 2002. Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions? Nucleic Acids Res. **30**:1886–1894.

Rocha, E. P. C., O. Pradillon, H. Bui, C. Sayada, and E. Denamur. 2002. A new family of highly variable proteins in the *Chlamydophila pneumoniae* genome. Nucleic Acids Res. **30**:4351–4360.

Roche, R. J., and E. R. Moxon. 1995. Phenotypic variation in *H. influenzae*: the interrelationship of colony opacity, capsule and lipopolysaccharide. Microb. Pathog. **18**:129–140.

Sakamoto, N., K. Ohshima, L. Montermini, M. Pandolfo, and R. D. Wells. 2001. 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**:27171–27177.

Santibáñez-Koref, M. F., R. Gangeswaran, and J. M. Hancock. 2001. A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. Mol. Biol. Evol. **18**:2119–2123.

Saveliev, A., C. Everett, T. Sharpe, Z. Webster, and R. Festenstein. 2003. DNA triplet repeats mediate heterochromatin-protein-1-sensitive variegated gene silencing. Nature **422**:909–913

Schwartz, S. Jr., H. Yamamoto, M. Navarro, M. Maestro, J. Reventos, and M. Perucho. 1999. Frameshift mutations at mononucleotide repeats in caspase-5 and other target genes in endometrial and gastrointestinal cancer of the microsatellite mutator phenotype. Cancer Res. **59**:2995–3002.

Seznec, H., O. Agbulut, N. Sergeant et al. (15 co-authors). 2001. Mice transgenic for the human myotonic dystrophy region with expanded CTG repeats display muscular and brain abnormalities. Hum. Mol. Genet. **10**:2717–2726.

Sirand-Pugnet, P., P. Durosay, E. Brody, and J. Marie. 1995. An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chickrn β–tropomyosin pre-mRNA. Nucleic Acids Res. **23**:3501–3507.

Slebos, R. J., D. S. Oh, D. M. Umbach, and J. A. Taylor. 2002. Mutations in tetranucleotide repeats following DNA damage depend on repeat sequence and carcinogenic agent. Cancer Res. **62**:6052–6060.

Souza, R. F., R. Appel, J. Yin et al. (21 co-authors). 1996. Microsatellite instability in the insulin-like growth factor II receptor gene in gastrointestinal tumours. Nat. Genet. **14**:255–257.

Stallings, R. L. 1994. Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases. Genomics **21**:116–121.

Streelman, J. T., and T. D. Kocher. 2002. Microsatellite variation associated with prolactin expression and growth of salt-challenged Tilapia. Physiol. Genomics **9**:1–4.

Subramanian, S., R. K. Mishra, and L. Singh. 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. **4**:R13.

Suraweera, N., B. Iacopetta, A. Duval, A. Compoint, E. Tubacher, and R. Hamelin. 2001. Conservation of mononucleotide repeats within 3′ and 5′ untranslated regions and their instability in MSI-H colorectal cancer. Oncogene **20**:7472–7477.

Tachida, H., and M. Iizuka. 1992. Persistence of repeated sequences that evolve by replication slippage. Genetics **131**:471–478.

Thiel, T., W. Michalek, R. K. Varshney, and A. Graner. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor. Appl. Genet. **106**:411–422.

Tidow, N., A. Boecker, H. Schmidt, K. Agelopoulos, W. Boecker, H. Buerger, and B. Brandt. 2003. Distinct amplification of an untranslated regulatory sequence in the *egfr* gene contributes to early steps in breast cancer development. Cancer Res. **63**:1172–1178.

Timchenko, N. A., A. L. Welm, X. Lu, and L. T. Timchenko. 1999. CUG repeat binding protein (CUGBP1) interacts with the 5′ region of C/EBPbeta mRNA and regulates translation of C/EBPbeta isoforms. Nucleic Acids Res. **27**:4517–4525.

Tóth, G., Z. Gáspári, and J. Jurka. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. **10**:967–981.

Toutenhoofd, S. L., F. Garcia, D. A. Zacharias, R. A. Wilson, and E. E. Strehler. 1998. Biochim. Biophys. Acta **1398**:315–320.

Trifonov, E. N. 2003. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation. Pp. 1–24 *in* S. P. Wasser, ed., Evolutionary theory and processes: nodern horizons, papers in honor of Eviatar Nevo. Kluwer Academic Publishers. Amsterdam, The Netherlands.

van Belkum, A., S. Scherer, L. van Alphen, and H. Verbrugh. 1998. Short-sequence DNA repeats in prokaryotic genomes. Microbiol. Mol. Biol. Rev. **62**:275–293.

van Ham, S. M., L. van Alphen, F. R. Mooi, and J. P. M. van Putten. 1993. Phase variation of *H. influenzae* fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. Cell **73**:1187–1196.

van Lith, H. A., and L. F. van Zutphen. 1996. Characterization of rabbit DNA microsatellites extracted from the EMBL nucleotide sequence database. Anim. Genet. **27**:387–395.

van Putten, J. P. M. 1993. Phase variation of lipopolysaccharide directs interconversion of invasive and immuno-resistant phenotypes of *N. gonorrhoeae*. EMBO J. **12**:4043–4051.

Varshney, R. K., T. Thiel, N. Stein, P. Langridge, and A. Graner. 2002. *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell. Mol. Biol. Lett. **7**:537–546.

Vassileva, V., A. Millar, L. Briollais, W. Chapman, and B. Bapat. 2002. Genes involved in DNA repair are mutational targets in endometrial cancers with microsatellite instability. Cancer Res. **62**:4095–4099.

Weiser, J. N., J. M. Love, and E. R. Moxon. 1989. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. Cell **59**:657–665.

Wren, J. D., E. Forgacs, J. W. Fondon 3rd, A. Pertsemlidis, S. Y. Cheng, T. Gallardo, R. S. Williams, R. V. Shohet, J. D. Minna, and H. R. Garner. 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. Am. J. Hum. Genet. **67**:345–356.

Xiong, X. F., and W. S. Reznikoff. 1993. Transcriptional slippage during the transcription initiation process at a mutant lac promoter in vivo. J. Mol. Biol. **231**:569–580.

Yamada, M., S. Tsuji, and H. Takahashi. 2002. Involvement of lysosomes in the pathogenesis of CAG repeat diseases. Ann. Neurol. **52**:498–503.

Yamada, T., T. Koyama, S. Ohwada, K. Tago, I. Sakamoto, S. Yoshimura, K. Hamada, I. Takeyoshi, and Y. Morishita. 2002*a*. Frameshift mutations in the MBD4/MED1 gene in primary gastric cancer with high-frequency microsatellite instability. Cancer Lett. **181**:115–120.

Yamada, N. A., G. A. Smith, A. Castro, C. N. Roques, J. C. Boyer, and R. A. Farber. 2002*b*. Relative rates of insertion and deletion mutations in dinucleotide repeats of various lengths in mismatch repair proficient mouse and mismatch repair deficient human cells. Mutat. Res. **499**:213–225.

Yanagisawa, Y., Y. Akiyama, S. I. E. Iida, T. Nomizu, K. Sugihara, Y. Yuasa, and K. Maruyama. 2000. Methylation of the hMLH1 promoter in familial gastric cancer with microsatellite instability. Int. J. Cancer **85**:50–53.

Young, E. T., J. S. Sloan, and K. van Riper. 2000. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. Genetics **154**:1053–1068.

Youings, S. A., A. Murray, N. Dennis, S. Ennis, C. Lewis, N. McKechnie, M. Pound, A. Sharrock, and P. Jacobs. 2000. FRAXA and FRAXE: the results of a five year survey. J. Med. Genet. **37**:415–421.

Yu, M. W., Y. C. Yang, S. Y. Yang, S. W. Cheng, Y. F. Liaw, S. M. Lin, and C. J. Chen. 2001. Hormonal markers and hepatitis B virus–related hepatocellular carcinoma risk: a nested case-control study among men. J. Natl. Cancer Inst. **93**:1644–1651.

Yu, M. W., Y. C. Yang, S. Y. Yang et al. (12 co-authors). 2002. Androgen receptor exon 1 CAG repeat length and risk of hepatocellular carcinoma in women. Hepatology **36**:156–163.

Zhou, X. P., A. Loukola, R. Salovaara, M. Nystrom-Lahti, P. Peltomaki, A. de la Chapelle, L. A. Aaltonen, and C. Eng. 2002. PTEN mutational spectra, expression levels, and subcellular localization in microsatellite stable and unstable colorectal cancers. Am. J. Pathol. **161**:439–447.

Zienoddiny, S., D. Ryberg, A. F. Gazdar, and A. Haugen. 1999. DNA mismatch binding in human lung tumor cell lines. Lung Cancer **26**:15–25.

Zoghbi, H. Y., and H. T. Orr. 2000. Glutamine repeats and neurodegeneration. Annu. Rev. Neurosci. **23**:217–237.

Thomas H. Eickbush, Associate Editor

Accepted December 3, 2003