Am. J. Hum. Genet. 67:345–356, 2000

Repeat Polymorphisms within Gene Regions: Phenotypic and Evolutionary Implications

Jonathan D. Wren,[1] Eva Forgacs,[3] John W. Fondon III,[2] Alexander Pertsemlidis,[4,5] Sandra Y. Cheng,[8,9] Teresa Gallardo,[10] R. S. Williams,[5,6,10] Ralph V. Shohet,[5,10] John D. Minna,[3,5,7] and Harold R. Garner[4,5,8,9]

Programs in [1]Genetics and Development and [2]Molecular Biophysics, Southwestern Graduate School of Biomedical Sciences, [3]Hamon Center for Therapeutic Oncology Research, and Departments of [4]Biochemistry, [5]Internal Medicine, [6]Molecular Biology, and [7]Pharmacology, [8]McDermott Center for Human Growth and Development, [9]Center for Biomedical Inventions, and the [10]Ryburn Cardiac Center, The University of Texas Southwestern Medical Center, Dallas

We have developed an algorithm that predicted 11,265 potentially polymorphic tandem repeats within transcribed sequences. We estimate that 22% (2,207/9,717) of the annotated clusters within UniGene contain at least one potentially polymorphic locus. Our predictions were tested by allelotyping a panel of ~30 individuals for 5% of these regions, confirming polymorphism for more than half the loci tested. Our study indicates that tandem-repeat polymorphisms in genes are more common than is generally believed. Approximately 8% of these loci are within coding sequences and, if polymorphic, would result in frameshifts. Our catalogue of putative polymorphic repeats within transcribed sequences comprises a large set of potentially phenotypic or disease-causing loci. In addition, from the anomalous character of the repetitive sequences within unannotated clusters, we also conclude that the UniGene cluster count substantially overestimates the number of genes in the human genome. We hypothesize that polymorphisms in repeated sequences occur with some baseline distribution, on the basis of repeat homogeneity, size, and sequence composition, and that deviations from that distribution are indicative of the nature of selection pressure at that locus. We find evidence of selective maintenance of the ability of some genes to respond very rapidly, perhaps even on intragenerational timescales, to fluctuating selective pressures.

Introduction

The association between repeating microsatellite elements and polymorphism, caused by the expansion and contraction of the core repetitive unit via slipped-strand mispairing, uneven recombination, or some combination of both, has been well documented (Jeffreys et al. 1988; Zuliani and Hobbs 1990; Jakupciak and Wells 1999; Karthikeyan et al. 1999). The potential for such elements to cause disease has been highlighted by the linkage of several inherited neurological disorders to increases in the copy number of various trinucleotide repeats. For some of these diseases—such as Machado-Joseph disease (CAG repeat), Haw River syndrome (CAG repeat), Huntington disease (CAG repeat), and some forms of fragile-X syndrome (CGG repeat)—the repetitive element occurs within the coding sequence (Verkerk et al. 1991; Kawaguchi et al. 1994). For others—including Fredreich ataxia (GAA repeat), myotonic dystrophy (CAG repeat), and another form of fragile-X syndrome—the expanded repeats lie in the introns and 3′ and 5′ UTRs, respectively (Smits et al. 1993; Jansen et al. 1994; Bidichandani et al. 1998). There are excellent reviews available for those interested in the molecular basis for the instability of some of these repeats and how they contribute to disease (Wells 1996; Hancock and Santibanez-Koref 1998).

Building on the success of POMPOUS (polymorphic marker prediction of ubiquitous simple sequences), a program that we developed to identify tandem-repeat polymorphisms in genomic sequences as genetic markers (Fondon et al. 1998), we developed a program, REP-X, to increase the predictive accuracy for repeat polymorphisms in transcribed sequences. This is achieved by requiring perfect homogeneity of the repetitive unit, allowing shorter repeats, and including mononucleotide repeats. This generates fewer predictions, but ones with a higher expected probability of being polymorphic. POMPOUS and REP-X both were used to generate the initial set of predictions tested, so that the role of homogeneity in the sensitivity and specificity of polymorphism predictions could be analyzed.

We applied both of these informatics tools to the UniGene database of human cDNA sequences and selected, for further study, 146 genes predicted to harbor

Received April 10, 2000; accepted for publication June 2, 2000; electronically published July 7, 2000.

Address for correspondence and reprints: Dr. Harold R. (Skip) Garner, 5323 Harry Hines Boulevard, Dallas, TX 75390-8591. E-mail: garner@utsw.swmed.edu

0002-9297/2000/6702-0013$02.00

Table 1

Genes Allelotyped for Polymorphisms

Repeat Class [a]	Amino Homogeneity: No. of Repeats	Amino Homogeneity: Acid[b]	Amino Homogeneity: Score[c]	Gene Name[d]	Region	No. of: Alleles	No. of: Heterozygotes	No. of: Homozygotes	Heterozygosity
TGC	8	L	.92	Insulin-like growth factor II receptor	Coding	2	1	36	.03
TGC	12	L	1.00	DNA for ApoER2	Coding	2	1	33	.03
TGA	13	D	1.00	Histidine-rich calcium-binding protein	Coding	5	14	22	.39
TG	8	CV	1.00	Sperm acrosomal protein	Coding	3	9	21	.30
TCC	11	S	.94	Eph-family protein*	Coding	3	3	26	.10
CGG	7	G	1.00	CCAAT/enhancer binding protein alpha	Coding	4	3	33	.08
CGG	7	G	.90	Forkhead protein	Coding	2	2	35	.05
CCG	8	A	.92	Transforming growth factor-β	Coding	2	2	35	.05
CAG	29	Q	.97	MN1: meningioma (known polymorphism)	Coding	3	4	26	.13
CAG	21	Q	.97	DNA polymerase γ (mitochondrial): already discovered	Coding	3	6	24	.20
CAG	12	Q	.94	Myocyte-specific enhancer factor 2A gene	Coding	4	15	15	.50
AGGCCC	22	QA	.94	Putative transcription factor CA150	Coding	3	5	25	.17
AGG	8	E	1.00	Histidine-rich calcium-binding protein	Coding	4	32	0	1.00
AGG	8	E	1.00	HVEC	Coding	2	11	16	.41
ACC	15	H	.95	Polycomb 2 homologue	Coding	2	1	29	.03
ACC	12	H	.92	Serine/threonine protein kinase	Coding	2	1	35	.03
AAC	8	T	1.00	MEK kinase 1	Coding	2	6	24	.20
TTTG	7		.96	EST—similar to T cell RANTES-specific precursor	Unknown	3	11	19	.37
TG	12		.92	EST—similar to SP:SYB2_HUMAN P19065 synaptobrevin	Unknown	2	6	24	.20
CGG	8		.92	EST—similar to SP:S22371 NADH dehydrogenase	Unknown	2	3	32	.09
CA	22		1.00	TRE17 5′ extremity and unnamed adjacent to TRE17	Unknown	4	8	22	.27
AGG	13		.92	EST—similar to gb:M64497 apolipoprotein Al regulatory protein-1	Unknown	3	8	22	.27
AGAT	11		.97	EST—similar to gb:L07077 enoyl-CoA hydratase	Unknown	6	21	14	.60
AG	23		1.00	EST—similar to aminopeptidase (puromycin sensitive)	Unknown	3	27	8	.77
TTTTG	5		.96	Tumor necrosis factor receptor type 1*	3′ UTR	2	1	35	.03
TGC	11		1.00	DMR-N9 and myotonic dystrophy kinase (dm kinase) gene	3′ UTR	13	27	8	.77
TG	16		.98	Human novel growth-factor receptor	3′ UTR	3	3	29	.09
TG	24		1.00	Fibroblast growth factor 9	3′ UTR	3	6	22	.21
TG	22		1.00	Synaptotagmin I	3′ UTR	3	6	24	.20
TG	21		.98	β 3 adrenergic receptor	3′ UTR	5	37	0	.00
TG	21		1.00	Platelet CGI-PDE	3′ UTR	4	32	2	.94
TG	18		1.00	Lysosome-associated membrane protein 2 (alternative products)	3′ UTR	3	5	25	.17
TG	11		.92	Coagulation factor IX gene	3′ UTR	4	37	0	1.00
TG	10		1.00	Plasma gelsolin	3′ UTR	2	1	29	.03

CT	8		1.00	Apoptosis inhibitor survivin gene	3′ UTR	2	1	36	.03
CT	18		1.00	CLCN3–voltage gated calcium channel	3′ UTR	6	15	12	.56
CA	24		1.00	Checkpoint suppressor 1	3′ UTR	6	18	12	.60
CA	19		1.00	Sorting nexin 2	3′ UTR	6	12	5	.71
CA	17		1.00	DNA repair protein XRCC1	3′ UTR	4	6	24	.20
CA	14		1.00	Tumor necrosis factor receptor–related protein	3′ UTR	4	6	24	.20
CA	12		1.00	Ubiquitous Kruppel-like factor	3′ UTR	4	16	17	.48
ATCCC	8		1.00	Interferon regulatory factor 2	3′ UTR	3	10	19	.34
AT	25		1.00	Homeodomain protein	3′ UTR	5	8	16	.33
AT	19		1.00	Human semaphorin III family homolog	3′ UTR	4	17	7	.71
AT	12		1.00	Adenylate cyclase activating polypeptide 1	3′ UTR	2	2	33	.06
AT	14		.93	M3 muscarinic acetylcholine receptor	3′ UTR	3	2	33	.06
AG	12		1.00	Platelet-derived growth factor PDGF-a	3′ UTR	2	1	36	.03
AAAAC	10		1.00	Leukotriene B4 omega-hydroxylase	3′ UTR	3	2	28	.07
CGG	8		1.00	Very low density lipoprotein receptor	5′ UTR	4	3	34	.08
CGG	7		1.00	Phosphatase and tensin homologue (mutated in multiple advanced cancers 1)	5′ UTR	2	9	24	.27
CAG	21		1.00	MAB-21 cell fate–determining protein homolog	5′ UTR	18	32	5	.86
CAG	12		.99	Brain natriuretic protein BNP	5′ UTR	2	2	26	.07
CAG	6		1.00	ERK1 mRNA for protein serine/threonine kinase	5′ UTR	2	3	33	.08
AG	24		1.00	MDS1B (MDS1): acute myeloid leukemia–related transcript	5′ UTR	4	9	21	.30
TGG	6	V	.89	TAN-1 (*drosophila* notch homolog)*	Coding	1	0	30	
TGC	8	L	.96	Clq/MBL/SPA receptor ClqR(p): for phagocytosis	Coding	1	0	30	
TGA	10	D	.97	Transcriptional activation factor TAFII32	Coding	1	0	27	
TGA	10	D	.96	Cardiac calsequestrin	Coding	1	0	30	
CCG	6	P	1.00	Transcription factor HCSX	Coding	1	0	36	
CAG	18	Q	.90	CREB-binding protein	Coding	1	0	37	
CAG	12	S	.92	Sterol regulatory element binding protein–2*	Coding	1	0	28	
CAG	10	S	.93	Sex-determining region Y: box 11	Coding	1	0	30	
CAG	10	A	1.00	MAP kinase kinase kinase	Coding	1	0	30	
CAG	9	Q	.96	ALR mRNA	Coding	1	0	29	
CAG	7	S	1.00	Insulin receptor substrate–1	Coding	1	0	30	
CAG	7	Q	1.00	Nck, Ash, and phospholipase C γ-binding protein NAP4*	Coding	1	0	28	
AGG	21	E	.88	Extracellular matrix protein	Coding	1	0	29	
AGG	14	E	.90	Major centromere autoantigen CENP-B*	Coding	1	0	30	
AGG	14	E	.90	Potassium voltage-gated channel, Shaker-related subfamily, member 4	Coding	1	0	28	
AGG	13	E	.95	Golgin-95 (clone SY11)	Coding	1	0	26	
AGG	12	E	.92	Major centromere autoantigen CENP-B	Coding	1	0	30	
AGG	10	E	.90	Aortic carboxypeptidase-like protein ACLP	Coding	1	0	30	

(*continued*)

Table 1 Continued

Repeat Class [a]	No. of Repeats	Amino Homogeneity: Acid [b]	Amino Homogeneity: Score [c]	Gene Name [d]	Region	No. of: Alleles	No. of: Heterozygotes	No. of: Homozygotes	Heterozygosity
AGG	9	G	.92	Transformer-2 β (HTRA-2 β)	Coding	1	0	37	
AGG	8	E	.92	Mitochondrial hinge protein (repeated; same results)	Coding	1	0	30	
AGG	7	E	.91	Actin-binding protein (filamin)	Coding	1	0	36	
AG	9	RE	1.00	Putative GR6 protein*	Coding	1	0	30	
ACC	18	T	.92	Ankyrin G	Coding	1	0	26	
ACC	11	H	.94	T-type calcium channel α-1 subunit	Coding	1	0	30	
AAG	25	E	.95	p160 mRNA, partial cds	Coding	1	0	28	
TTTC	7		.96	EST—similar to gb:M20022 HLA class I histocompatibility antigen	Unknown	1	0	30	
CCG	9		.96	EST—similar to gb:X52611 transcription factor AP-2	Unknown	1	0	30	
CA	12		.92	EST—similar to ICAM-1 precursor	Unknown	1	0	30	
AT	17		.96	EST—calcium-transporting ATPase plasma membrane	Unknown	1	0	30	
AAAC	6		.92	EST—similar to gb:M33987 carbonic anhydrase I	Unknown	1	0	30	
AAAAG	8		.92	EST—similar to gb:M29874 cytochrome P450 IIB6	Unknown	1	0	30	
TTTTC	5		1.00	Tight junction (zonula occludens) protein ZO-1	3′ UTR	1	0	33	
TG	16		1.00	Insulin-like growth factor II receptor	3′ UTR	1	0	37	
TG	16		1.00	Neuronal nitric oxide synthase	3′ UTR	1	0	37	
TG	10		.95	Vacuolar proton-ATPase, subunit D	3′ UTR	1	0	30	
CT	10		.90	TGF β 1 precursor	3′ UTR	1	0	30	
CA	14		.93	Nerve growth factor receptor	3′ UTR	1	0	37	
CA	13		1.00	Lectin-like oxidized LDL receptor	3′ UTR	1	0	36	
CA	7		1.00	c-sis/platelet-derived growth factor 2	3′ UTR	1	0	37	
AT	13		.92	Orphan G protein-coupled receptor (RDC1)	3′ UTR	1	0	37	
AGG	7		1.00	Inwardly rectifying potassium channel KIR3.3	3′ UTR	1	0	37	
TCCGGC	9		.92	LCA-homolog, leukocyte antigen–related protein	5′ UTR	1	0	37	
GGC	11		1.00	Ubiquitin-conjugating enzyme E2B (yeast RAD6 homologue)	5′ UTR	1	0	29	
GGC	9		1.00	Fibroblast growth factor 18	5′ UTR	1	0	28	
CCG	7		.96	Bc12, p53 binding protein Bbp/53BP2	5′ UTR	1	0	30	
CCG	6		.94	TGF β 1 precursor	5′ UTR	1	0	30	
AGG	6		.94	Inositol polyphosphate 1-phosphatase	5′ UTR	1	0	27	
AG	18		.93	HOX 5.1 gene for HOX 5.1 protein	5′ UTR	1	0	37	

Note.—Sample of predictions located in coding regions, 5′ and 3′ UTRs, and regions for which no annotation was available was chosen for further analysis.

[a] Shown as a cyclic permutation of the repeated unit, for categorical convenience.

[b] Provided for coding-sequence predictions.

[c] Based on the repetitive unit and number of repeats.

[d] Identified from the annotated header in the UniGene entries; polymorphisms verified by sequencing in addition to allelotyping are followed by an asterisk (*).

repeat polymorphisms of a variety of types. We tested our predictions by designing primers flanking the predicted polymorphisms, PCR amplifying, and then allelotyping them for either of two panels of individuals (see the Material and Methods section).

After establishing the predictive power of our program, we surveyed amino acid repeats in the human genome, for repeat polymorphisms in coding regions. This analysis supports several new conclusions with respect to the functional and evolutionary importance of polymorphic repeat sequences.

Material and Methods

Computational Tools

The REP-X and POMPOUS programs were run on a Hewlett Packard Exemplar supercomputer running SPP-UX 5.2. For selection of our allelotyping test set and validation of predictive accuracy, both codes were run on the annotated portion of the June 1999 release of the UniGene database of expressed human sequences. For all other analyses, REP-X was run on the January 2000 release of UniGene. The longest sequence with the fewest ambiguous bases in each UniGene cluster was used for analysis. REP-X identifies repeats by comparing a sequence to itself and identifying the longest similar sequence, for each position in the sequence. Mononucleotide A/T repeats within 5% of the end of the sequences were excluded from further analysis.

Polymorphism Prediction Criteria

For a stretch of repeated nucleotides, the minimum number of occurrences of the tandemly repeated unit necessary for it to be considered polymorphic depends on its size and homogeneity (Fondon et al. 1998). Fractional numbers of repeating units were rounded to the nearest integer. To be scored as polymorphic by POMPOUS, repeated DNA sequences had to be eight units long for dimers, whereas trimers, tetramers, pentamers to nonamers, and repeat units of lengths $\geqslant$10 required seven, six, five, and four repeats, respectively. POMPOUS permits up to 10% of the nucleotides within a repeat to deviate from the core repetitive unit. REP-X predictions included monomers, dimers, trimers, tetramers, pentamers to nonamers and repeats with unit size $\geqslant$10 or larger, for which the minimum numbers of repeated units were 12, 6.5, 5.5, 4.5, 3.5, and 2.5, respectively. REP-X permits no deviations from the core repetitive element. Statistics for intronic sequences were obtained by running REP-X on a subset of the GenBank primate database that includes annotated intron sequences for humans only.

Analysis of Peptide Repeats

Comparison of peptide repeats with nucleotide repeats was done by translating the UniGene nucleotide sequences into protein sequences by use of the annotated start and stop sites. Protein sequences were scanned for occurrences of four perfectly repeated residues, and these were then extended, permitting mismatches, provided that two consecutive matches immediately followed the mismatch. These repeats were then scored for polymorphic potential on the basis of REP-X parameters. For gene fragments with a stop site, but not a start site, annotated, open-reading frames (ORFs) uninterrupted by stop codons were chosen. Ambiguous ORFs were discarded.

Statistics for repeats within genomic DNA were obtained by running REP-X on 273 MB of high-throughput genome sequencing (HTGS) human DNA sequence obtained from the National Center for Biotechnology Information.

Allelotyping

For the loci chosen for allelotyping, primers were synthesized using our Mermade oligonucleotide synthesizer (Rayner et al. 1998). For 64 of the genes in our analysis, genomic DNA was extracted, by standard methods, from 30 Epstein-Barr virus–immortalized B lymphoblastoid cell lines of small-cell, non–small-cell, and adenocarcinoma lung cancer patients. For 40 of the genes analyzed, genomic DNA was obtained from the peripheral leukocytes of 36 individuals, 12 of whom had a diagnosis of hypertrophic cardiomyopathy.

Genomic DNA was amplified by PCR using the "touchdown" methodology, with an initial denaturation step at 95°C for 10 min. This was followed by 10 touchdown cycles of 30 s at 94°C, 30 s at 70°C (with a decrease, in the annealing temperature, by 1°C each cycle), and 30 s at 72°C. This was followed by 30 cycles of 30 s at 94°C, 30 s at 60°C, and 30 s at 72°C, with a final extension at 72°C for 10 min. DNA (~50–100 ng of genomic DNA) was amplified in 20-μl reaction volumes containing 50 mM KCl, 10 mM Tris (pH 8.3), 1.5 mM $MgCl_2$, 200 μM each dNTP, 1 μM each primer, 0.5 U of Ampli*taq* Gold (PE Biosystems), and 2 μCi of [^{32}P]-dCTP (Amersham). The samples were heat denatured, snap chilled, and run on a 6.8% polyacrylamide gel (acrylamide:bis acrylamide ratio 19:1) containing 10 M urea. The gels were dried and exposed overnight using BioMax film (Kodak).

For some genes, the PCR products were also sequenced. For better separation of the different alleles, the samples were run on a 0.5 × mutation-detection enhancement gel. Shifted bands were excised from the gel, and DNA was eluted with distilled water and was ream-

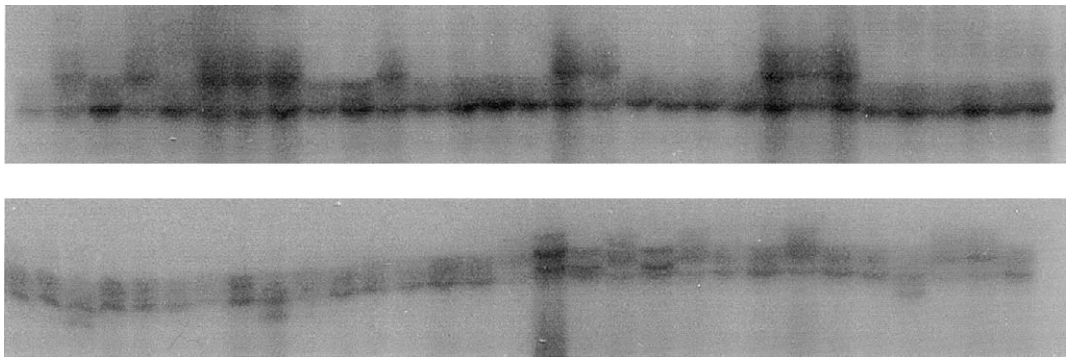

Figure 1 Two examples of REP-X prediction of polymorphisms. *Top,* Polymorphism in HVEC, encoding eight to nine polyglutamic acids residues located in the cytoplasmic portion of this transmembrane protein. Polyglutamic acid tracts have been associated with microtubule binding and factors promoting DNA conformational changes. The effects of copy-number variance are not known but could play a role in herpesvirus infectivity. *Bottom,* Frameshifting dinucleotide polymorphism located at the C-terminal end of ACRP. The alleles shown here represent all three coding frames resulting from the polymorphism.

plified using the original PCR primers. The PCR product was run on a 2% agarose gel and was purified by Genlute Agarose spin columns (Sigma). Automated bidirectional sequencing was performed by ABI 377 Dye Terminator cycle sequencing. Sequences were analyzed and were compared with the sequences downloaded from GenBank by DNAStar software (DNAStar).

Results

Polymorphism Predictions

Of the 11,265 putative polymorphic loci identified in coding regions and UTRs, 2,769 are in annotated UniGene clusters. This allowed us to categorize each locus as occurring within either a coding region or the 5′ or 3′ UTRs. A total of 146 of the 2,769 were chosen for analysis, on the basis of medical interest and as a representative sample of repeat types (table 1; GenBank accession numbers are listed in the Electronic-Database Information section), and 102 were successfully amplified within three attempts of primer design. Of these 102, 54 (53%) were verified to be polymorphic, defined as having at least two alleles among a sample of 60–74 chromosomes. Examples of results are shown in figure 1. The results for all loci tested are summarized in table 2, and, as anticipated, more-stringent homogeneity requirements resulted in higher polymorphism levels.

Effect of Homogeneity of Repeats on Polymorphism Levels

Increasing the homogeneity requirements for polymorphism predictions increased prediction accuracy. A total of 54 (53%) of 102 POMPOUS predictions were found to be polymorphic, whereas 50 (67%) of 75 REP-X predictions were found to be polymorphic (tables1 and 2). Within the polymorphisms tested, only four of 27 tandem repeats containing deviations from the canonical repeat unit were found to harbor polymorphisms. These results are consistent with the

Table 2

Polymorphism Prediction Accuracy, by Gene Region

	Region[a]				
	5′ UTR	Coding	3′ UTR	Unknown	Overall
Algorithm:[b]					
POMPOUS no. confirmed/no. predicted (%)	6/13 (46)	17/42 (40)	24/34 (70)	7/13 (54)	54/102 (53)
REP-X no. confirmed/no. predicted (%)	6/9 (67)	14/25 (56)	23/30 (77)	7/11 (64)	50/75 (67)
Average heterozygosity[c] (SD)	.28 (.30)	.22 (.25)	.38 (.34)	.37 (.24)	.32 (.30)
Average no. of alleles[c] (SD)	5.33 (6.28)	2.82 (.95)	4.00 (2.30)	3.29 (1.38)	3.69 (2.67)

[a] Transcriptional position determined using annotated start and stop sites.

[b] The two algorithms differ primarily in their homogeneity requirements (REP-X requires perfect homogeneity). Increasing homogeneity leads to more-specific predictions with a modest decrease in sensitivity.

[c] For polymorphic loci; nonpolymorphic loci are excluded. Because the distributions do not the shape of a standard normal curve, SD is not an appropriate statistical measure, but is shown to illustrate that the variability within the samples.

Table 3

Predicted Repeat Polymorphisms, by Species and Location

	Entries		Predictions			
Organism	Total	Annotated	5′ UTR	Coding	3′ UTR	Unannotated
Homo sapiens:						
Frequencies:						
Repeats[a]			1 per 3,195 bp	1 per 23,107 bp	1 per 4,544 bp	1 per 1,200 bp
ALUs[b]			1 per 9,842 bp	1 per 823,530 bp	1 per 11,276 bp	1 per 3,239 bp
No. (%)	92,219	9,717	438 (4.5)	672 (6.9)	1,659 (17.1)	32,611
Mus musculus:						
Frequency of repeats			1 per 2,682 bp	1 per 26,057 bp	1 per 2,857 bp	1 per 1,872 bp
No. (%)	75,962	5,398	250 (4.6)	309 (5.7)	978 (18.1)	12,934
Rattus norvegicus:						
Frequency of repeats			1 per 4,043 bp	1 per 32,375 bp	1 per 3,055 bp	1 per 753 bp
No. (%)	28,687	3,407	109 (3.2)	156 (4.6)	592 (17.4)	15,473

Note.—Statistics were derived from the January 2000 version of UniGene, using the REP-X criteria. Predicted polymorphisms in coding sequences occur at a frequency of 1/5–1/10 of those in UTRs. The frequency of repeats in unannotated UniGene entries deviates significantly from those in coding regions and UTRs, indicating that they contain another population of sequences.

[a] Frequencies included because there are more raw sequence data available for the 3′ than for the 5′ UTR and, therefore, more putative polymorphisms in that region. For comparison, the frequency of potentially polymorphic elements predicted (by the same criteria) in total human genomic DNA is ~1 per 2,600 bp.

[b] ALU sequences constitute a large portion of repeats in the human genome, occurring at a frequency of ~1 per 6,000 bp in genomic DNA (Deininger and Batzer 1999), and are also included for comparison.

previously observed positive correlation between repeat homogeneity and polymorphism levels (Kunst et al. 1997).

Distribution of Repetitive Unit Lengths

Once the suitability of the new polymorphism criteria for intragenic sequences was established, REP-X was used to generate predictions of repeat polymorphisms in human, mouse, and rat cDNA sequences from the January 2000 UniGene release. Because sequences determined from 3′ UTRs are overrepresented in the UniGene database (Boguski and Schuler 1995), the frequencies of repeats in each of these regions are reported per nucleotide scanned, to permit direct comparison (table 3). Furthermore, only those entries for which reliable translational start and stop sites are known are used for comparisons of coding 5′ and 3′ UTR sequences.

That UTRs harbor more repetitive and polymorphic elements than are seen in coding sequences is expected; what is surprising is the number of repeat polymorphisms occurring within the coding regions of genes. If, as with the test set, two thirds of these predictions are correct, then ~3.7% of human genes contain at least one fairly common repeat polymorphism. Note that the high frequency of repeat polymorphisms for unannotated sequences in table 3 is due primarily to mononucleotide repeats (as shown in table 4).

Coding sequences, introns, and 3′ and 5′ UTRs have characteristic distributions of repetitive unit lengths (table 4). More than 92% of the predicted polymorphisms within coding sequences have unit lengths that are a multiple of 3, which would give protection against frameshift mutations (but see Ohno 1984). However, this does leave 0.5% (51) of the annotated data set entries with potentially frameshifting loci.

Peptide Repeats and DNA Repeats

Specific amino acids have an increased proclivity to form homopolymeric runs (Sumiyama et al. 1996). Because of the redundancy of the genetic code, it is not necessary for repeated tracts of amino acids to be encoded by homogeneous trinucleotide repeats (except for methionine and tryptophan homopolymers, which are very rare). We examined all peptide homopolymers of length greater than or equal to five, to determine polymorphic potential (fig. 2).

Although hydrophobic repeats tend to be located in amino-terminal signaling peptides (fig. 2), we found that some amino acids (Ile, Val, Met, Cys, Asn, Phe, Trp, and Tyr) are rarely, if at all, found repeated in human genes. Numerous potential reasons come to mind to explain these observations: in the case of Trp or Tyr, this is possible because their bulkiness could contribute to unstable structures, because of steric interference; in the case of Cys, it is likely because it could contribute to anomalous cross-linking. Other amino acids vary in the frequency with which they are encoded by potentially polymorphic elements, ranging from 6% for Arg to 62% for His. There is a tendency for homopolymeric runs of residues with more codons to have lower homogeneity in their

Table 4

Percentage Distribution of Repeats—Unit Sizes Considered Potentially Polymorphic, by Annotated Region

	5′ UTR	Coding[a]	3′ UTR	Combined	Unknown[b]	Unknown − Anomalous[c]	Intronic[d]	Genomic[e]
	(*N* = 438)	(*N* = 672)	(*N* = 1,659)	(*N* = 2,769)	(*N* = 32,611)	(*N* = 8,496)	(*N* = 3,480)	(*N* = 104,097)
	Proportion of Total Repeats							
Repeat-unit size:								
1	31.3	2.2	48.3	34.5	88.8	57.1	47.8	53.7
2	14.6	2.4	28.6	20.0	4.7	18.0	21.1	18.3
3	31.1	67.1	4.6	24.0	1.4	5.5	4.6	4.0
4	2.3	.3	6.3	4.2	2.3	8.9	11.6	11.5
5	9.6	1.5	6.4	5.7	1.5	5.9	7.8	6.9
6	5.3	12.1	2.4	5.2	0.5	2.1	2.8	2.2
7	.7	.0	.2	.2	.1	.2	.3	.3
8	.0	.2	.1	.1	<.1	.1	.2	.2
9+	5.3	14.3	3.1	6.1	.5	2.1	3.8	3.0
Total[f]	100	100	100	100	100	100	100	100
	Proportion of Repeats with a Unit Size Evenly Divisible by 3							
mod(3) = 0	37.7	92.4	7.7	33.1	2.2	8.4	8.8	7.2

[a] Repeats are dominated by mod(3) repeats that avoid frameshifting, much more so than any other region.
[b] Sequences lacking annotation, trimmed of potential poly-A tails within 5% of the sequence ends.
[c] Unannotated sequences, excluding sequences considered anomalous. "Anomalous" is defined here as an unannotated sequence containing one or more poly-A or poly-T tracts and belonging to a single sequence cluster.
[d] Because cDNA and EST sequences in UniGene lack introns, these sequences, in humans, were obtained from the intronic annotations in the GenBank primate databases.
[e] Distribution was obtained from 273 Mb of Human GenBank HTGS database sequence.
[f] Totals do not sum exactly to 100 because of rounding errors.

encoding DNA (e.g., His and Gln > Thr and Gly > Arg and Ser), although there are some deviations from this trend (e.g., Leu > Pro and Gly > Lys).

Discussion

Polymorphism Profiles of Gene Regions

The 5′ and 3′ UTRs are known to harbor more genetic variation than is seen in coding sequences, and this is borne out in our results; relative decreases in both heterozygosity levels and number of alleles were observed for the coding-sequence polymorphisms. Such regional variances may help to identify in which region of a gene an unknown expressed-sequence tag (EST) is located. This variation is presumably due primarily to two factors: the presence, in the UTRs, of repetitive sequences with regulatory functions (e.g., mRNA stability) and, within coding sequences, selection against repeat polymorphisms. Unlike the 3′ UTR, the 5′ UTR exhibits a strong bias toward specific trinucleotide repeats (Stallings 1994). Of the 136 trinucleotide repeats identified in 5′ UTRs, 101 of them were CGG or CCG (data not shown), which have been shown to serve as binding sites for nuclear proteins (Richards et al. 1993; Stallings 1994). The 3′ UTR regions display a broad distribution of repeat-unit sizes but are biased toward mononucleotide repeats (poly-A tails within 5% of the sequence ends were excluded from the analysis). Intronic sequences were found to have a repeat-unit profile very similar to that of genomic DNA.

Approximately 90% of UniGene clusters lack annotation. Each class of transcribed sequence (5′, 3′, intronic, and coding) has a distinct distribution of repeat types, frequency, and unit-size distributions. These distributions may serve to help classify sequences of unknown origin. The difference in distribution of repeat types within these "unknown" sequences and "known" genes indicates that a significant proportion of UniGene clusters may not represent genuine genes. With respect to their repetitive character, these "unknowns," in the aggregate, do not resemble transcribed DNA at all (as shown in table 4) and are explained if they contain a substantial fraction of cloning or sequencing artifacts. For example, unlike coding sequence, they are biased away from trinucleotide repeats, contain more monomers than are seen in genomic DNA, and contain a higher frequency of ALU sequences than is seen in any transcribed region. Attempts to infer coding sequences from these entries by using conventional "longest ORF" methods, as well as more-sophisticated algorithms (Burge and Karlin 1997), yielded low-confidence coding predictions and repeated amino acid profiles distinctly

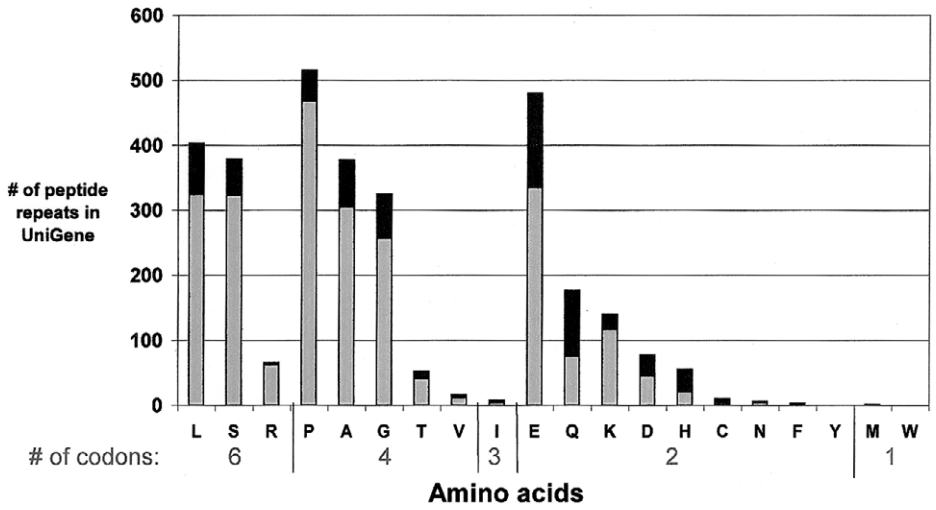

Figure 2 Amino acid repeats from transcribed UniGene entries, varying in both number and potential for polymorphism. Tandem repeats of at least five amino acids from annotated UniGene entries are shown, grouped first by the number of corresponding codons available to encode them, in descending order from left to right and then, within this group by the number of repeats. Some amino acid repeats are severely underrepresented in humans, whereas others are not. Some amino acid repeats (Q and H) tend to be encoded by a higher percentage of potentially polymorphic codon repeats (*blackened portion of bars*) than are those (R and P) that use a more heterogeneous codon set to encode the repeat (*gray portion of bars*). Amino acids encoded by more codons have a greater tendency to exhibit repeat heterogeneity, but there are significant departures from this trend (e.g., L > P and G > K).

different than those of the annotated sequences (data not shown). This is possibly due to the fact that ~34,500 (37.5%) of UniGene clusters in this "build" contain only one sequence (UniGene Build #113), which represents either very rare transcripts or sequencing artifacts. In addition, 16,513 of these single-sequence clusters contain anomalous poly-A and poly-T tracts after exclusion of 3′ poly-A tails. These single-sequence clusters have been deposited in GenBank over the years from a variety of sources, and many of them are likely single-read sequences with low-quality base calls. If these 24,115 anomalous poly-A and poly-T repeats found within the 16,513 single-sequence clusters are discarded, this leaves 8,496 polymorphic loci predictions. Then, the repeat frequency and size distribution of the new set begins to more closely resemble some of the other categories in table 4, such as the 3′ UTR or genomic sequences. If we are to assume that at least these 16,513 clusters are not true genes, then the number of valid UniGene clusters becomes 75,706, and inferring the number of human genes from the number of UniGene clusters results in an 18% overestimation. When added to the 2,769 predictions from the annotated clusters, the result is a set of 11,265 loci most likely to be repeated regions in true genes.

Evolutionary and Phenotypic Implications

The redundancy of the genetic code renders it unnecessary to use a perfect DNA repeat to encode a peptide repeat, and it is biologically intuitive to assume that evolution will tend to exploit this redundancy, to fix the number of repeated elements in a gene at some optimal level. To find evidence of this, we sought to compare the homogeneity of all peptide repeats to "expected" levels, for each amino acid. Because of DNA's natural propensity for self-similarity, random models of expected homogeneity are unsuitable (Tautz et al. 1986). By examining, within genomic DNA, the distributions of sequences that, if translated, would yield peptide repeats, we can estimate the expected levels of homogeneity for a peptide repeat in the absence of selective pressure on the encoded protein. Selection is clearly acting to influence the length of peptide repeats, so comparisons of homogeneity in genomic versus coding repeats are paired by "peptide" type and repeat length (fig. 2). As anticipated (Schmid et al. 1999), selection appears to depress polymorphism levels in repeated coding sequences, by peppering repeats of "optimal" length with synonymous substitutions (fig. 3*B* and 3*C*). However, for some specific proteins (data not shown)—and even for some entire classes of peptide repeats (fig. 3*D*)—peptide repeats appear to be under positive selection for both elevated homogeneity and, thus, higher polymorphism; for these loci, silent substitutions might indeed be deadly. It has been noted that the length of repeats with higher homogeneity tend to diverge between species such as mice and humans (Alba et al. 1999). Furthermore, these peptide repeats are more common in eukaryotes than in

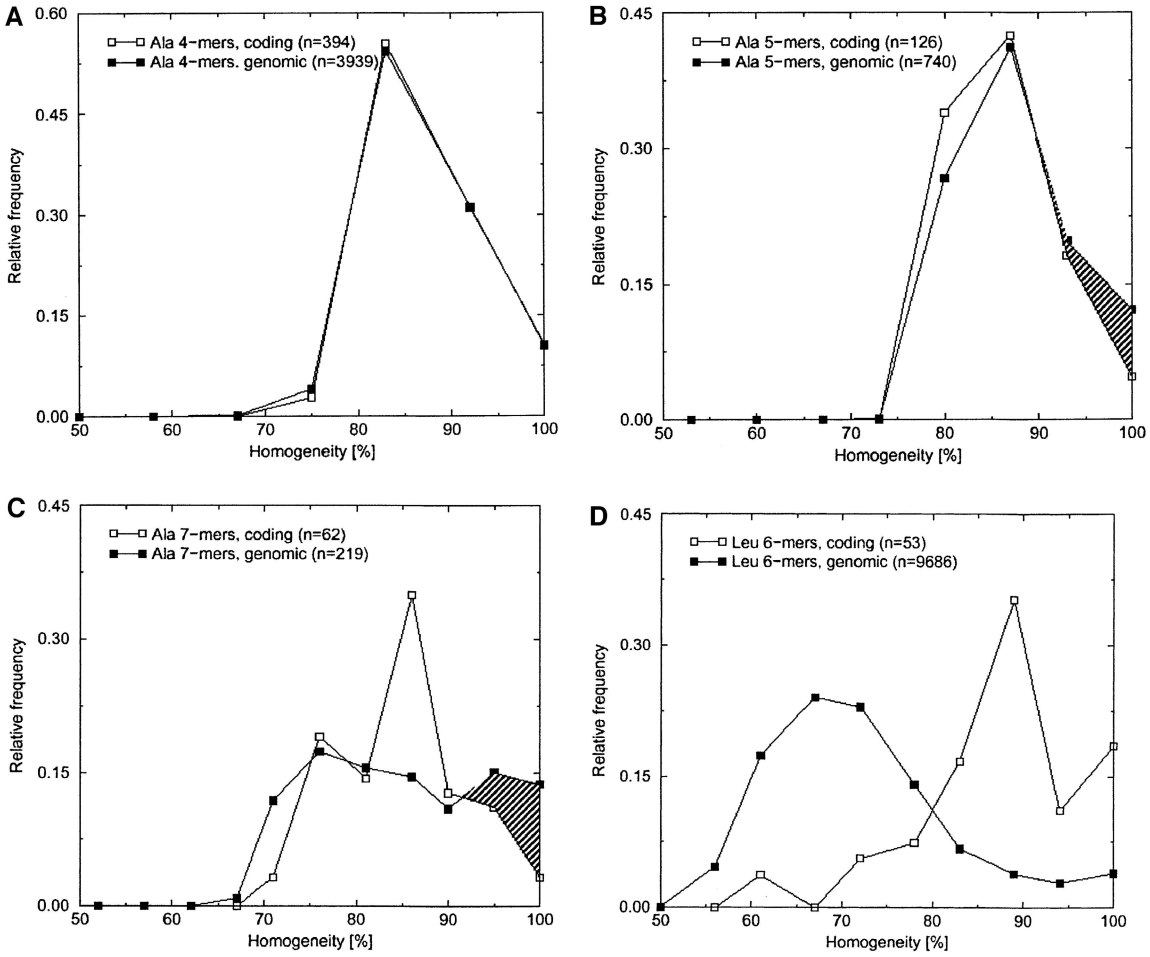

Figure 3 Selection for or against allelic plasticity, reflected in repeat homogeneity. *A,* Homogeneity distributions for DNA encoding four repeated amino acids, including alanine. These are almost identical. Because, regardless of their homogeneity, repeats of four trimers are rarely polymorphic, differences between coding DNA and genomic sequences are not expected. *B,* Homogeneity distributions for perfectly homogeneous repeats of five trimers. Although they are expected to exhibit some elevated plasticity, the effects of selection to repress this plasticity can be observed in the reduced proportion of alanine-coding pentamers that have perfect homogeneity, relative to genomic sequences (*shaded*). *C,* Homogeneity distributions for longer repeats. This trend continues and becomes more pronounced for longer repeats, wherein highly pure homopolymer-encoding repeats are underrepresented, presumably because of selection for synonymous substitutions that repress repeat expansions and contractions. *D,* Homogeneity distributions for other types of repeats, such as leucine hexamers. The distribution is shifted toward higher homogeneity in coding sequences relative to genomic sequences, suggesting that selection is functioning to increase allelic plasticity for a substantial proportion of these loci.

prokaryotes, and such hypermutable elements may be a mechanism for more-rapid protein evolution (Marcotte et al. 1999).

If allelic *diversity* is advantageous for a population, such as in genes involved in host-pathogen interactions, then balancing selection has little trouble maintaining multiple alleles, even when the alleles are relatively immutable. However, if the "optimal" number of repeated elements in a gene varies over time, the fittest allele may ultimately be the one with maximal *plasticity.* Selection may be actively maintaining this elevated plasticity, for some genes, by preserving high homogeneity in tandem repeats. We hypothesize that, for many of the highly homogeneous coding repeats predicted, by our algorithms, to be polymorphic, this may indeed be the case. And, given that the rate of expansion/contraction of pure, long tandem repeats is high enough that somatic mosaicism is commonplace (Leeflang et al. 1999), it is possible that there is considerable allelic diversity and, thus, potential competition and evolution among cells *within an individual* (somatic and/or germline).

The location and type of polymorphic repeats can facilitate the building of hypotheses about the potential functional roles that a gene region may have in physiology. Differences in tandem-repeat lengths in 5′ UTR promoter elements, for example, can lead to the mod-

ulation of the level of gene transcription, either directly (Shimajiri et al. 1999; Yamada et al. 2000) or indirectly (Mooser et al. 1995; Valenti et al. 1999), whereas AU-rich elements in the 3′ UTRs have been shown to affect mRNA stability (Gay and Babajko 2000). Given the variety of amino acid properties, there are a large number of ways in which polymorphic repeats in coding sequences could affect protein function. For example, in yeast, tandem peptide repeats are found to be overrepresented in certain functional classes of genes, such as transcription factors (Mar Alba et al. 1999).

The polymorphisms that we discovered in two of the genes in our subset—those for herpes viral entry protein C (HVEC; fig. 1, *top*) and sperm acrosomal protein (ACRP; fig. 1, *bottom*)—are useful examples of how the location of a polymorphism can be used to construct a hypothesis about its potential effect. In *HVEC,* the repetitive region encodes eight to nine glutamic acid residues located in the cytoplasmic portion of this transmembrane protein, whereas the *ACRP* gene has a polymorphic TG-dinucleotide repeat beginning near the 3′ end of its coding sequence, with stop codons in all three frames after 5, 7, or 14 residues (alternative translations are MCVCV, VCVCVRV, and CVCVCESVNAQVGI). ACRP is found in maturing and elongating spermatid heads and is suspected to be involved in penetration of the oocyte zona pellucida (Beaton et al. 1995). Although HVEC could be responsible for some of the known population variance in susceptibility to herpesvirus infection, ACRP could, similarly, have an impact on fertility.

Polymorphic repeats in genes can not only provide useful information about selection forces acting on a gene—and, thereby, aid in generating a hypothesis about the physiological role of the gene—but are also useful as extremely tightly linked markers for mapping studies. We have developed and tested a method optimized to find tandem-repeat polymorphisms in cDNA sequences, where they are considered rare (Nakamura et al. 1987). We have shown that there are a surprisingly large number of these elements undiscovered and uncharacterized in humans and rodents, some of which may provide functional information about the proteins that contain them, whereas others may provide important leads to potential disease-causing mechanisms. The details of the predicted polymorphisms in gene regions described here and in 11,265 others are available for download as a text file at The Garner Lab at UTSW.

Acknowledgments

This research was funded by Special Projects Open Research Environment grant P50CA70907, the Patrick O'Brien Montgomery Distinguished Chair, and the D.W. Reynolds Cardiovascular Clinical Research Center. We would like to thank Hewlett-Packard for the loan of an Exemplar supercomputer.

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

Garner Lab at UTSW, The, http://pompous.swmed.edu

GenBank Overview, http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html (for human databases [accession numbers Y00285, D86407, M60052, AF047437, D83492, Y11525, AF032886, M60315, X82209, U60325, U49020, AF017789, M60052, AF060231, AF013956, D86550, AF042838, T62484, T63962, R42196, X78261, T70173, R12160, T47177, X55313, L08835, M64347, D14838, M55047, X70811, U36798, U36336, K02402, X04412, U75285, X78520, U68723, AF065482, M36089, L04489, AB015132, X15949, AF022654, U38276, S83513, U29589, X06374, AB002454, D16532, U92436, U38810, AL021155, X60188, U43292, M75866, M73980, U94333, U21858, D55655, U34962, U47741, U02031, U23752, AF002715, AF010403, S62539, AB005216, AB011792, X05299, M55514, L06147, X05299, AF053944, U68063, Y00764, X53416, AF008192, U13616, AF051946, U88153, T87413, R33865, T62835, T80553, T70304, T60175, L14837, Y00285, U17327, NM_004691, X02812, M14764, AB010710, M12783, U67784, U52152, Y00815, M74525, AF075292, U58334, X02812, L08488, and X17360])

Entrez Nucleotide, http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide (for nucleotide sequences)

UniGene Database (latest release), ftp://ncbi.nlm.nih.gov/repository/UniGene/Hs.seq.uniq.Z

UniGene Build #113, ftp://ncbi.nlm.nih.gov/repository/UniGene/Hs.info

References

Alba MM, Santibanez-Koref MF, Hancock JM (1999) Conservation of polyglutamine tract size between mice and humans depends on codon interruption. Mol Biol Evol 16: 1641–1644

Beaton S, ten Have J, Cleary A, Bradley MP (1995) Cloning and partial characterization of the cDNA encoding the fox sperm protein FSA-Acr.1 with similarities to the SP-10 antigen. Mol Reprod Dev 40:242–252

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

Boguski MS, Schuler GD (1995) ESTablishing a human transcript map. Nat Genet 10:369–371

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Deininger PL, Batzer MA (1999) Alu repeats and human disease. Mol Genet Metab 67:183–193

Fondon JW III, Mele GM, Brezinschek RI, Cummings D, Pande A, Wren J, O'Brien KM, et al (1998) Computerized

polymorphic marker identification: experimental validation and a predicted human polymorphism catalog. Proc Natl Acad Sci USA 95:7514–7519

Gay E, Babajko S (2000) AUUUA sequences compromise human insulin-like growth factor binding protein-1 mRNA stability. Biochem Biophys Res Commun 267:509–515

Hancock JM, Santibanez-Koref MF (1998) Trinucleotide expansion diseases in the context of micro- and minisatellite evolution, Hammersmith Hospital, April 1–3, 1998. EMBO J 17:5521–5524

Jakupciak JP, Wells RD (1999) Genetic instabilities in (CTG.CAG) repeats occur by recombination. J Biol Chem 274:23468–23479

Jansen G, Willems P, Coerwinkel M, Nillesen W, Smeets H, Vits L, Howeler C, et al (1994) Gonosomal mosaicism in myotonic dystrophy patients: involvement of mitotic events in (CTG)n repeat variation and selection against extreme expansion in sperm. Am J Hum Genet 54:575–585

Jeffreys AJ, Royle NJ, Wilson V, Wong Z (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature 332:278–281

Karthikeyan G, Chary KV, Rao BJ (1999) Fold-back structures at the distal end influence DNA slippage at the proximal end during mononucleotide repeat expansions. Nucleic Acids Res 27:3851–3858

Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, Kawakami H, et al (1994) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet 8:221–228

Kunst CB, Leeflang EP, Iber JC, Arnheim N, Warren ST (1997) The effect of FMR1 CGG repeat interruptions on mutation frequency as measured by sperm typing. J Med Genet 34: 627–631

Leeflang EP, Tavare S, Marjoram P, Neal CO, Srinidhi J, MacFarlane H, MacDonald ME, et al (1999) Analysis of germline mutation spectra at the Huntington's disease locus supports a mitotic mutation mechanism. Hum Mol Genet 8:173–183 [erratum: Hum Mol Genet 8:717]

Mar Alba M, Santibanez-Koref MF, Hancock JM (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

Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D (1999) A census of protein repeats. J Mol Biol 293:151–160

Mooser V, Mancini FP, Bopp S, Petho-Schramm A, Guerra R, Boerwinkle E, Muller HJ, et al (1995) Sequence polymorphisms in the apo(a) gene associated with specific levels of Lp(a) in plasma. Hum Mol Genet 4:173–181

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, et al (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622

Ohno S (1984) Repeats of base oligomers as the primordial coding sequences of the primeval earth and their vestiges in modern genes. J Mol Evol 20:313–321

Rayner S, Brignac S, Bumeister R, Belosludtsev Y, Ward T, Grant O, O'Brien K, et al (1998) MerMade: an oligodeoxyribonucleotide synthesizer for high throughput oligonucleotide production in dual 96-well plates. Genome Res 8: 741–747

Richards RI, Holman K, Yu S, Sutherland GR (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

Schmid KJ, Nigro L, Aquadro CF, Tautz D (1999) Large number of replacement polymorphisms in rapidly evolving genes of drosophila: implications for genome-wide surveys of dna polymorphism. Genetics 153:1717–1729

Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, Sasaguri Y (1999) Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. FEBS Lett 455:70–74

Smits AP, Dreesen JC, Post JG, Smeets DF, de Die-Smulders C, Spaans-van der Bijl T, Govaerts LC, et al (1993) The fragile X syndrome: no evidence for any recent mutations. J Med Genet 30:94–96

Stallings RL (1994) Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases. Genomics 21: 116–121

Sumiyama K, Washio-Watanabe K, Saitou N, Hayakawa T, Ueda S (1996) Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals. J Mol Evol 43:170–178

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Valenti K, Aveynier E, Leaute S, Laporte F, Hadjian AJ (1999) Contribution of apolipoprotein(a) size, pentanucleotide TTTTA repeat and C/T(+93) polymorphisms of the apo(a) gene to regulation of lipoprotein(a) plasma levels in a population of young European Caucasians. Atherosclerosis 147: 17–24

Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, Reiner O, et al (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65:905–914

Wells RD (1996) Molecular basis of genetic instability of triplet repeats. J Biol Chem 271:2875–2878

Yamada N, Yamaya M, Okinaga S, Nakayama K, Sekizawa K, Shibahara S, Sasaki H (2000) Microsatellite polymorphism in the heme oxygenase-1 gene promoter is associated with susceptibility to emphysema. Am J Hum Genet 66: 187–195

Zuliani G, Hobbs HH (1990) A high frequency of length polymorphisms in repeated sequences adjacent to Alu sequences. Am J Hum Genet 46:963–969