



# GENOME RESEARCH

## Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties

Albino Bacolla, Jacquelynn E. Larson, Jack R. Collins, et al.

*Genome Res.* 2008 18: 1545-1553 originally published online August 7, 2008

Access the most recent version at doi:[10.1101/gr.078303.108](https://doi.org/10.1101/gr.078303.108)

---

**References** This article cites 53 articles, 14 of which can be accessed free at:  
<http://genome.cshlp.org/content/18/10/1545.full.html#ref-list-1>

**Open Access** Freely available online through the *Genome Research* Open Access option.

**License** Freely available online through the Genome Research Open Access option.

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Copyright © 2008, Cold Spring Harbor Laboratory Press

# Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties

Albino Bacolla,<sup>1,5</sup> Jacquelynn E. Larson,<sup>1</sup> Jack R. Collins,<sup>2</sup> Jian Li,<sup>3</sup> Aleksandar Milosavljevic,<sup>3</sup> Peter D. Stenson,<sup>4</sup> David N. Cooper,<sup>4</sup> and Robert D. Wells<sup>1</sup>

<sup>1</sup>*Institute of Biosciences and Technology, Center for Genome Research, Texas A&M University Health Science Center, Houston, Texas 77030, USA;* <sup>2</sup>*Advanced Biomedical Computing Center, Advanced Technology Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702, USA;* <sup>3</sup>*Department of Molecular and Human Genetics and Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA;* <sup>4</sup>*Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, United Kingdom*

Microsatellites are abundant in vertebrate genomes, but their sequence representation and length distributions vary greatly within each family of repeats (e.g., tetranucleotides). Biophysical studies of 82 synthetic single-stranded oligonucleotides comprising all tetra- and trinucleotide repeats revealed an inverse correlation between the stability of folded-back hairpin and quadruplex structures and the sequence representation for repeats  $\geq 30$  bp in length in nine vertebrate genomes. Alternatively, the predicted energies of base-stacking interactions correlated directly with the longest length distributions in vertebrate genomes. Genome-wide analyses indicated that unstable sequences, such as CAG:CTG and CCG:CGG, were over-represented in coding regions and that micro/minisatellites were recruited in genes involved in transcription and signaling pathways, particularly in the nervous system. Microsatellite instability (MSI) is a hallmark of cancer, and length polymorphism within genes can confer susceptibility to inherited disease. Sequences that manifest the highest MSI values also displayed the strongest base-stacking interactions; analyses of 62 tri- and tetranucleotide repeat-containing genes associated with human genetic disease revealed enrichments similar to those noted for micro/minisatellite-containing genes. We conclude that DNA structure and base-stacking determined the number and length distributions of microsatellite repeats in vertebrate genomes over evolutionary time and that micro/minisatellites have been recruited to participate in both gene and protein function.

[Supplemental material is available online at [www.genome.org](http://www.genome.org).]

DNA microsatellites, tandem arrays of simple repeats such as mono-, di-, tri-, and tetranucleotides, are common in eukaryotic genomes (Lander et al. 2001; Waterston et al. 2002; Sharma et al. 2007). However, different types of sequences display widely variable abundances within each class (such as the trinucleotide and tetranucleotide repeats, triNRs and tetraNRs, respectively) (Subramanian et al. 2003), particularly at lengths  $\geq 30$  nt. For example, >16,000 tracts of mononucleotides comprised of  $\geq 30$  As or Ts are present in the human genome, but only seven analogous tracts of Gs or Cs are found (Bacolla et al. 2006).

<sup>5</sup>CpG<sup>3'</sup>-(CpG)-containing repeats are generally rare, suggesting that cytosine methylation and subsequent deamination leading to T:A transitions from <sup>5m</sup>C:G base pairs (Walsh and Xu 2006), a frequent cause of human gene mutation (Mort et al. 2008), may have been involved (Kelkar et al. 2008). However, because the rates of cytosine methylation (Bacolla et al. 2001) and <sup>5m</sup>C deamination (Lindahl and Nyberg 1974; Frederico et al. 1993) decrease with increasing DNA stability, CpG-containing triNR and tetraNR are expected to display varying transition rates according to their C+G content (Elango et al. 2008). These rela-

tionships remain to be clarified. Other repeats,<sup>6</sup> such as ATGC, AGCT, ACCC, and ACT, are also rare, although they do not contain the CpG step. Therefore, the underlying mechanisms of biased sequence representation remain poorly understood.

Another characteristic of simple repeats is the lengths attained genome-wide by specific sequences such as AAG repeats, which are consistently the longest by quite a margin within the triNR family (Clark et al. 2006). This behavior is also enigmatic.

The number of repeat units is polymorphic at certain loci, and this variability can play specific roles both in physiology and pathology (for review, see Supplemental Table 1). Also, more than 20 neurological diseases (Supplemental Table 1) are caused by the expansion of triNRs within the coding or untranslated regions of genes (for review, see Wells and Ashizawa 2006; Mirkin 2007; Orr and Zoghbi 2007; Kovtun and McMurray 2008). Among these, recessive Friedreich ataxia (FA) is atypical since, contrary to all other triNR diseases, the expanded GAA:TTC tract

## <sup>5</sup>Corresponding author.

E-mail [abacolla@ibt.tamhsc.edu](mailto:abacolla@ibt.tamhsc.edu); fax (713) 677-7689.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.078303.108>. Freely available online through the *Genome Research* Open Access option.

<sup>6</sup>Abbreviations: Concerning nucleic acid nomenclature, we designated a double-stranded genomic triNR or tetraNR by its unique sequence (Tables 1, 2) with no specification as to the reading frame or strand composition. Accordingly, AGC includes all genomic tracts composed of AGC:GCT, GCA:TGC, CAG:CTG, GCT:AGC, TGC:GCA, and CTG:CAG duplex DNA, where the colon separates the complementary strands. In contrast, we specify single-stranded DNA oligonucleotides and their reading frame by d(AGC)<sub>n</sub>, for example. A subscript (*n*) indicates the number of repeating units. Hydrogen-bonded nucleotides are also indicated by a colon, that is, A:T.

(up to >1000 repeat copies) in the first intron of the *FXN* gene is somatically stable in the transmitting parents (De Biase et al. 2006; Pandolfo 2006). The reasons for this behavior are unclear. Elevated microsatellite alterations at selected tetraNRs (EMAST) have been noted in mismatch repair-proficient cancers of the respiratory tract, skin, and bladder, and they involve preferentially biased purine:pyrimidine (R:Y) sequences (for review, see Supplemental Table 2). Hence, since length instabilities are believed to arise from unrepaired bulges during DNA polymerase slippage over the repeats (Sammalkorpi et al. 2007), R:Y tracts appear to selectively escape repair. However, the underlying mechanisms remain speculative (Yang 2006).

Herein, we show that the relative abundances of triNR and tetraNR sequences in nine vertebrate genomes are inversely proportional to the capacity of their single strands to fold back into hairpin or quadruplex structures of varying thermodynamic stabilities. The sequences that form the longest tracts comprise R:Y-rich sequences, and their length distributions in the human genome correlate directly with the strength of base-stacking in

teractions within the R-rich strand. Hence, certain DNA secondary structures have prevented the accumulation of specific repeating sequences in genomes over evolutionary time, whereas strong base-stacking interactions have favored their expansion. These results are discussed in the context of human disease-associated repeat polymorphism and genome-wide analyses, which suggest the recruitment of micro/minisatellites by transcription factors and other regulatory genes to perform specific functions, particularly in the nervous system.

## Results

### Certain long tetraNRs are absent from the human genome

The sequence representation of tetraNR sequences in the human genome was analyzed by comparing tracts comprising  $\geq 8$  identical units (Table 1). For some nucleotide combinations, the number of tracts exceeded 5000 copies, whereas for others, no tracts were found (Table 1). This diversity, spanning nearly four

**Table 1. TetraNR abundance in the human genome for tracts with  $\geq 8$  units,  $T_a$  values, and DNA structure**

No. of tracts	Unique genomic sequence	Forward oligonucleotide sequence (FO)	$T_a$ (FO)	DNA structure (FO)	Reverse oligonucleotide sequence (RO)	$T_a$ (RO)	DNA structure (RO)
0	ACCG	d(ACCG) <sub>9</sub>	39.4	H	d(CGGT) <sub>9</sub>	50.6	H
0	ACGT	d(ACGT) <sub>9</sub>	78.4	H			
0	AGCG	d(AGCG) <sub>9</sub>	86.7	H	d(CGCT) <sub>9</sub>	21.8	H
0	ATCG	d(ATCG) <sub>9</sub>	77.5	H	d(CGAT) <sub>9</sub> <sup>b</sup>	76.4	H
0	ATGC	d(ATGC) <sub>9</sub>	80.1	H	d(GCAT) <sub>9</sub> <sup>b</sup>	78.6	H
0	CCCG	d(CCCG) <sub>9</sub>	35.8	H	d(CGGG) <sub>9</sub> <sup>c</sup>	N	Q
0	CCGG	d(CCGG) <sub>9</sub> <sup>c</sup>	N	H			
1	AACG	d(AACG) <sub>9</sub>	39.9	H	d(CGTT) <sub>9</sub>	29.0	H
2	AGCT	d(AGCT) <sub>9</sub>	76.7	H			
3	ACGC	d(ACGC) <sub>9</sub>	46.3	H	d(GCGT) <sub>9</sub>	59.9	H
3	ACTG	d(ACTG) <sub>9</sub>	31.0	H	d(CAGT) <sub>9</sub>	N	C
4	ACCC	d(ACCC) <sub>9</sub>	24.7	H	d(GGGT) <sub>9</sub> <sup>c</sup>	N	Q
6	AAGT	d(AAGT) <sub>9</sub>	S	C	d(ACTT) <sub>9</sub>	N	C
8	ACGG	d(ACGG) <sub>9</sub>	74.8	H	d(CCGT) <sub>9</sub>	22.6	H
9	AGCC	d(AGCC) <sub>9</sub>	39.2	H	d(GGCT) <sub>9</sub>	56.3	H
10	AACT	d(AACT) <sub>9</sub>	N	C	d(AGTT) <sub>9</sub>	S	C
24	AATT	d(AATT) <sub>9</sub>	53.3	H			
34	ACTC	d(ACTC) <sub>9</sub>	N	C	d(GAGT) <sub>9</sub>	40.9	H
53	ACCT	d(ACCT) <sub>9</sub>	N	C	d(AGGT) <sub>9</sub>	24.4	H
54	AAGC	d(AAGC) <sub>9</sub>	37.9	H	d(GCTT) <sub>9</sub>	41.4	H
71	AATC	d(AATC) <sub>9</sub>	S	C	d(GATT) <sub>9</sub>	25.5	H
84	AGGC	d(AGGC) <sub>9</sub>	57.1	H	d(GCCT) <sub>9</sub>	N	C
86	AACC	d(AACC) <sub>9</sub>	N	C	d(GGTT) <sub>9</sub>	N	C
138	ACAG	d(ACAG) <sub>9</sub>	19.5	H	d(CTGT) <sub>9</sub>	16.7	H
539	AGGG	d(AGGG) <sub>9</sub> <sup>c</sup>	N	Q <sup>d</sup>	d(CCCT) <sub>9</sub>	26.7	H
558	ACAT	d(ACAT) <sub>9</sub>	N	C	d(ATGT) <sub>9</sub>	24.7	H
688	AATG	d(AATG) <sub>9</sub>	51.0	H	d(CATT) <sub>9</sub>	N	C
1092	AAAC	d(AAAC) <sub>9</sub>	S	C	d(GTTT) <sub>9</sub>	N	C
1718	ATCC	d(ATCC) <sub>9</sub>	N	C	d(GGAT) <sub>9</sub>	21.0	H
4431	AAGG	d(AAGG) <sub>9</sub>	39.3	H	d(CCTT) <sub>9</sub>	N	H
5134	AGAT	d(AGAT) <sub>9</sub>	28.5	H	d(ATCT) <sub>9</sub>	N	C
6310	AAAG	d(AAAG) <sub>9</sub>	S	C	d(CTTT) <sub>9</sub>	N	C
8530	AAAT	d(AAAT) <sub>9</sub>	S	C	d(ATTT) <sub>9</sub>	N	C

The columns are as follows: Column 1, number of tetraNRs  $\geq 8$  units in the human genome; column 2, unique tetraNR sequence; column 3, sequence of the forward oligonucleotides; column 4, annealing temperature ( $T_a$ ) for the forward oligonucleotide sequences; column 5, structure of the forward oligonucleotides; columns 6–8, same as columns 3–5, respectively, but for the reverse oligonucleotide sequences.

<sup>a</sup>Self-complementary sequence with no staggered ends for which  $T_a$  measurements for the forward and reverse oligonucleotide sequences are identical.

<sup>b</sup>Self-complementary sequence with staggered ends for which  $T_a$  measurements for the forward and reverse oligonucleotide sequences are not identical.

<sup>c</sup>Sequences studied by  $T_a$  analyses with variable salt concentration assay conditions and/or by circular dichroism (CD).

<sup>d</sup>Structure in the presence of 100 mM K<sup>+</sup> and 10 mM Mg<sup>2+</sup>. Other structures are possible at different metal ion concentrations (Supplemental material).

(N) Less than 8% total hypochromicity during both the heating and cooling steps without a detectable first-derivative peak; (S) greater than 8% total hypochromicity during either the heating or cooling steps without a detectable first-derivative peak; (C) single-stranded coil; (H) hairpin; (Q) quadruplex.

orders of magnitude, was intriguing. Based on molecular modeling (Supplemental material), we postulated that the range of tetraNR abundance values might be related to the capacity of certain sequences to fold back upon themselves in the single-stranded state to form quasi-stable non-B DNA conformations, either during DNA replication or transcription. The more stable folded conformations could serve as substrates for DNA repair and might therefore be excised (Wojciechowska et al. 2006; Mirkin 2007). Alternatively, these structures could be bypassed by the DNA replication complex (Iyer et al. 2000; Wells et al. 2005; Mirkin 2007; Zahra et al. 2007). Hence, those sequences that adopted the more stable non-B DNA conformations would tend to be lost over evolutionary time and consequently display reduced length distributions. In contrast, those sequences unable to adopt stable folded conformations would tend to survive, giving rise to extended length distributions.

### Annealing temperature determinations

To assess whether hairpin stability correlated with tetraNR abundance in the human genome, the annealing temperature ( $T_a$ ) values of 62 single-stranded synthetic oligonucleotides, 36 nt in length and corresponding to all tetraNR duplex DNAs (Table 1), were obtained by temperature-dependent absorption spectroscopy (TDAS). Thirty-six oligonucleotides displayed  $T_a$  values ranging from 86.7°C to 16.7°C. The highest  $T_a$  values were observed for the self-complementary sequences, that is, d(ACGT)<sub>9</sub>, d(ATCG)<sub>9</sub>, d(ATGC)<sub>9</sub>, and d(AGCT)<sub>9</sub>, with the exception of d(CCGG)<sub>9</sub>, for which no cooperative transition was observed. These genomic tetraNRs were either extremely rare ( $\leq 2$  occurrences) or absent. In contrast, the self-complementary d(AATT)<sub>9</sub> molecule displayed a much lower  $T_a$  value (53.3°C), which was also lower than for molecules whose hairpins were stabilized by Watson-Crick CG:CG or GC:GC doublets, such as d(AGCG)<sub>9</sub>, d(GCGT)<sub>9</sub>, d(ACGG)<sub>9</sub>, d(GGCT)<sub>9</sub>, and d(AGGC)<sub>9</sub>. Hence, adjacent CG:CG and GC:GC base pairs contributed the most stability to the tetraNR hairpins.

For the d(CCCCT)<sub>9</sub> oligonucleotide, a strong hysteresis effect was observed, with melting temperature ( $T_m$ ) values of  $\sim 80.0^\circ\text{C}$  and  $T_a$  values of  $\sim 30.0^\circ\text{C}$ . Additional TDAS determinations (Supplemental material) indicated that this behavior was due to slow hairpin formation by partially protonated cytosine pairs (C<sup>+</sup>:C). Therefore, only the  $T_a$  value at pH 7.0 was considered for further analyses.

Six oligonucleotides (Table 1, S) displayed a nearly linear decrease in absorbance with decreasing temperature, indicative of base-stacking within single-stranded helices, but little or no hydrogen bonding. The remaining 20 oligonucleotides manifested no temperature-dependent changes in absorbance (Table 1, N), implying that these sequences existed as either random, single-stranded coils or very stable hydrogen-bonded structures with  $T_a$  and  $T_m$  values  $>94^\circ\text{C}$ . To resolve these ambiguities, further TDAS and CD investigations were conducted (Supplemental material; Supplemental Fig. 1), which revealed that the d(CCGG)<sub>9</sub> and d(XGGG)<sub>9</sub> (X = C, T, or A) oligonucleotides formed highly stable hairpin and quadruplex structures under physiologic K<sup>+</sup> and Mg<sup>2+</sup> concentrations, respectively. These four sequences were rare in the human genome, and the numbers of quadruplex-forming repeats correlated inversely with structure stability (Rachwal et al. 2007). In contrast, the 22 sequences for which no folded-back structures could be revealed were among the most abundant genomic tetraNRs (Table 1).

In summary, the highest  $T_a$  values were found for genomic tracts that were only rarely represented, whereas low or absent  $T_a$  values were observed for the tetraNRs that were most abundant in the reference human genome sequence, thereby confirming the predictions made through modeling. Hence, hairpin and quadruplex stabilities are a robust predictor of tetraNR abundance in the human genome.

### $T_a$ and tetraNR abundance are inversely correlated

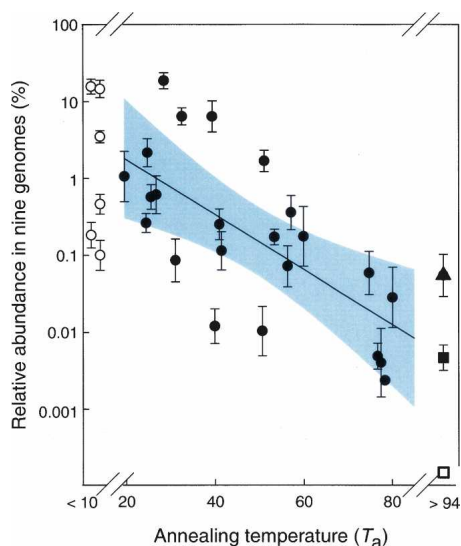
In order to determine whether a correlation existed between tetraNR abundance and hairpin/quadruplex formation, we plotted the highest  $T_a$  value for each pair of forward/reverse tetraNR sequences versus the log of the tetraNR abundance in the human genome (Table 1; see Supplemental material for a rationale of this analysis). A linear and inverse correlation was found ( $r = -0.607$ ,  $P = 0.028$ ), a result that established a clear relationship between genomic tetraNR abundance and non-B-DNA-structure formation. To verify whether the correlation was solely due to the low number of CpG-containing tetraNRs, we estimated the numbers of CpG-containing tetraNRs  $\geq 8$  units as if methylation-mediated and deamination-dependent C:G  $\rightarrow$  T:A transitions had not occurred (Supplemental material). The correlation remained significant ( $P = 0.007$ ).

Next, we reasoned that if the  $T_a$  versus tetraNR abundance relationship were to be dictated solely by DNA structural properties, analogous trends would be evident in other genomes despite differences in DNA repair systems. Analyses of eight additional vertebrate genomes indicated that in spite of large variations in the absolute numbers of tetraNRs of  $\geq 8$  units (144–126,473), both the rank order and the relative abundance of the different sequences were similar to those exhibited by the human genome. Moreover, the relative abundances for any given sequence remained unchanged irrespective of whether the results for one or the other chromosomal strand were analyzed. This suggested that genomic tetraNR abundance was critically dependent on DNA sequence composition. The relative tetraNR abundances for the nine combined vertebrate genomes correlated strongly with the  $T_a$  values ( $r = -0.706$ ,  $P = 0.0001$ ,  $\alpha_{0.05} = 0.98$ ) (Fig. 1). Significantly, no tracts  $\geq 8$  units were found for CCGG, the most thermostable hairpin-forming tetraNR sequence, in any of the nine genomes. The  $r^2$  values decreased hyperbolically when shorter tetraNR lengths were considered (Supplemental Fig. 2); nevertheless, all regressions remained significant ( $P = 0.03$  for tract lengths  $\geq 3$  units). This trend is in agreement with the dependence of base-paired stem-loop stability upon DNA length (Supplemental Fig. 3) and suggested that (1) longer tracts were deleted more efficiently than shorter tracts; and (2) some shorter tracts might have arisen from longer tracts as a result of DNA replication errors. Finally, no correlation ( $r = -0.21$ ,  $P = 0.32$ ) was found for tetraNRs  $\geq 4$  units in four nonvertebrate genomes that exhibited shorter length distributions than those noted in vertebrate genomes (Methods).

In summary, the stability of DNA secondary structures was a key determinant for the abundance of tetraNRs in vertebrate genomes and acted as an important modulator of sequence instability over evolutionary time.

### $T_a$ determinations in triNRs

Does hairpin formation determine the abundance of other simple repeats? To address this question, we considered the dis-



**Figure 1.** Correlation between  $T_a$  and tetraNR abundance in nine vertebrate genomes. Each of the 33 symbols represents one of the unique tetraNR sequences listed in Table 1 (column 2). For each sequence, the  $T_a$  (x-axis) is given by the highest value found for either the forward (column 4) or the reverse (column 7) oligonucleotide, whereas the relative abundance (y-axis) is the log of the mean fraction found for each tetraNR sequence relative to the total number of tetraNRs  $\geq 8$  units in the nine vertebrate genomes listed in Methods. The (solid line) regression was calculated only for those (solid circles) tetraNR sequences for which either the forward or reverse oligonucleotides displayed a  $T_a$  value within the measurable temperature range (10°C to 94°C). (Vertical lines) Standard errors; (blue) 99% confidence interval; (open circles) tetraNRs devoid of temperature-dependent structural transitions; (filled square) CCGG and (filled triangle) TGGG are tetraNRs with  $T_a$  values  $>94^\circ\text{C}$ ; (open square) CCGG with a  $T_a$  value of  $>94^\circ\text{C}$  but not present in any genome.

tribution of triNRs  $\geq 10$  units in the nine vertebrate genomes (Table 2). The numbers varied, with the ACG sequence being the rarest and the AAT sequence being the most abundant, hence displaying the same characteristics observed for the tetraNRs. The  $T_a$  values for the 20 single-stranded triNR oligonucleotides revealed that the 10 genomic triNRs can be divided into two groups: group 1, comprising the ACG, CCG, and AGC sequences, for which at least one strand displayed a high  $T_a$  value

(58.5°C–76.7°C); and group 2, comprising the other seven repeats, for which  $T_a$  values were low ( $\leq 27^\circ\text{C}$ ) (Supplemental material).

The group 1 triNRs were generally of lower abundance than the group 2 triNRs. One notable exception was the AGC sequence, which occurred about 10 times more frequently than expected. This discrepancy was intriguing given the expansion of CAG:CTG repeats (the AGC triNR contains the CAG:CTG sequence) in human genes as a frequent cause of neurological diseases (for review, see Wells and Ashizawa 2006; Mirkin 2007; Orr and Zoghbi 2007; Kovtun and McMurray 2008). Since CUG repeat-containing RNAs participate in splicing regulation and CAG-encoded polyglutamine tracts play a role in protein structure and function (Orr and Zoghbi 2007), the high proportion of the AGC triNR across all vertebrate genomes may reflect the action of positive selection over evolutionary time.

The highest  $T_a$  values for each of the triNR sequences correlated negatively with triNR abundance ( $r = -0.585$  for occurrences in the human genome and  $-0.685$  for the percentages in the vertebrate genomes). In summary, considering the composite data from the tetraNRs and triNRs, we conclude that hairpin-forming capacity rather than primary DNA sequence per se determined the relative abundance of simple repeating sequences over evolutionary time.

#### Intragenic micro/minisatellites

To test the prediction that the AGC triNR has been under positive selection, we compared the intergenic versus intragenic distributions of triNRs and tetraNRs in the human genome. At lengths  $\geq 30$  bp, both the AGC and CCG triNRs were highly ( $\sim 30\%$ ) over-represented within coding regions (Supplemental Figs. 4A, 5A), indicating selection for both sequences. At lengths  $\geq 12$  bp, the group 1 triNRs and the CCGG, AGCG, CCGG, and ACCG tetraNRs were also over-represented ( $>10\%$ ) within coding regions (Supplemental Figs. 4B, 5B). Since all these repeats are rare at lengths  $\geq 30$  bp (Tables 1, 2) and are associated with high ( $>50^\circ\text{C}$ )  $T_a$  values, these results support the notion that selection acted so as to preserve inherently unstable sequences within coding regions.

To assess the functional relevance of this conclusion, we searched for and analyzed all micro/minisatellite-containing (diNRs–11-mer repeats) cDNAs in the human genome. Of the approximately 2300 nonredundant genes, strong enrichment ( $P$ -

**Table 2.** TriNR abundances and  $T_a$  values

Unique genomic sequence	No. of tracts in human genome	Percent of all triNRs in vertebrate genomes (%)	Forward oligo (FO)	$T_a$ (FO)	Reverse oligo (RO)	$T_a$ (RO)
ACG	3	0.1	d(ACG) <sub>12</sub>	58.5	d(CGT) <sub>12</sub>	50.6
ACT	89	3.4	d(ACT) <sub>12</sub>	N	d(AGT) <sub>12</sub>	27.1
CCG	118	1.1	d(CCG) <sub>12</sub>	48.4	d(CGG) <sub>12</sub>	76.7
AGC	152	4.5	d(AGC) <sub>12</sub>	58.3	d(GCT) <sub>12</sub>	59.9
ACC	188	3.4	d(ACC) <sub>12</sub>	N	d(GGT) <sub>12</sub>	N
AGG	229	8.5	d(AGG) <sub>12</sub>	S	d(CCT) <sub>12</sub>	19.2
ATC	458	6.0	d(ATC) <sub>12</sub>	6.5	d(GAT) <sub>12</sub>	23.0
AAG	734	8.8	d(AAG) <sub>12</sub>	S	d(CTT) <sub>12</sub>	N
AAC	1075	10.7	d(AAC) <sub>12</sub>	N	d(GTT) <sub>12</sub>	N
AAT	3668	46.4	d(AAT) <sub>12</sub>	S	d(ATT) <sub>12</sub>	14.2

The columns are as follows: Column 1, unique triNR sequence; column 2, number of triNRs  $\geq 10$  units in the human genome; column 3, the percentage of each triNR relative to all triNRs for eight vertebrate genomes (human, chimp, mouse, rat, dog, chicken, *Fugu*, and zebrafish) was averaged. In the cow genome assembly browser, the AGC tracts accounted for  $\sim 80\%$  of all triNRs. Hence, these data were excluded from the analyses; column 4, sequence of the forward oligonucleotides; column 5,  $T_a$  for the forward oligonucleotides; column 6, sequence of the reverse oligonucleotides; column 7,  $T_a$  for the reverse oligonucleotides.

(N) Less than 8% total hypochromicity during both the heating and cooling steps without a detectable first-derivative peak; (S) greater than 8% total hypochromicity during either the heating or cooling steps without a detectable first-derivative peak.

**Table 3.** Intrinsic base-stacking and repeat tract length

Unique genomic sequence	10 longest tracts (mean bp)	Single-stranded oligo	TDAS slope ( $\times 10^4$ )	Bases X–Y	Base pairs X–Y (vacuum)	Base pairs X–Y (water)	Base pairs X–Y
				$\Delta G(v)$ (kcal/mol)	$\Delta E$ (kcal/mol)	$\Delta E$ (kcal/mol)	$\Delta H^{\circ}$ (kcal/mol)
(Average value for X–Y base or base pair steps)							
AAAG	221	d(AAAG) <sub>9</sub>	20.4	–7.78	–13.70	–10.75	–7.80
AAGG	124	d(AAGG) <sub>9</sub>	16.6	–8.03	–12.82	–10.22	–7.90
ATCC	95	d(GGAT) <sub>9</sub>	12.6	–7.25	–13.12	–10.17	–7.97
AGAT	68	d(AGAT) <sub>9</sub>	11.6	–7.08	–12.87	–10.27	–7.60
AAAT	61	d(AAAT) <sub>9</sub>	11.2	–6.81	–13.87	–11.02	–7.40
AAAC	46	d(AAAC) <sub>9</sub>	10.5	–6.63	–14.47	–10.60	–8.02
$r^2$			0.96	0.62	0.03	0.004	0.02
AAG	141	d(AAG) <sub>12</sub>	17.8	–7.86	–13.37	–10.50	–7.87
AAT	57	d(AAT) <sub>12</sub>	9.7	–6.58	–13.60	–10.87	–7.33
AAC	52	d(AAC) <sub>12</sub>	7.2	–6.34	–14.40	–10.30	–8.17

The columns are as follows: Column 1, unique genomic sequence, as per Tables 1 and 2; column 2, average length in base pairs of the 10 longest tracts found in the human genome; column 3, oligonucleotide sequence of the R-rich strand of the corresponding unique genomic sequence studied by TDAS; column 4, slope value of the TDAS curve; column 5, average theoretical free energy contribution to nearest-neighbor base stacking in single-stranded DNA (from Table 4 [ $\epsilon_i = 2$ ] of Friedman and Honig 1995); column 6, average contribution to nearest-neighbor base pair stacking energy in double-stranded canonical B-DNA from quantum-chemical calculations in vacuo (from Table 4 [ $\Delta E^{AB/CD}$  ( $\Delta E^{\ddagger}$ )] of Sponer et al. 2006); column 7, as column 6, but in the hydrated state (data from Table 6 [ $\Delta \Delta G_{\text{sol}}$  (water) with B3LYP/6-31G(d) data] of Sponer et al. 2006 were added to the Table 4 [ $\Delta E^{AB/CD}$  ( $\Delta E^{\ddagger}$ )] data); column 8, experimental stacking enthalpies for nearest-neighbor base pairs in double-stranded canonical B-DNA (from Table 1 of SantaLucia and Hicks 2004).

$r^2$ , Regression values obtained using the column 2 data as the  $x$  independent variable.

values down to  $10^{-40}$ ) was found for genes involved in the regulation of transcription and cellular functions, synaptic activity, axon guidance, and the MAPK and WNT signaling pathways, with no differences with respect to location (5'-UTR, ORF, or 3'-UTR) (Supplemental Table 3). Hence, genes involved in cell regulatory/signaling functions, particularly in the nervous system, actively recruited simple repeat sequences. diNRs and triNRs were the most abundant (>1000 copies) (Supplemental Fig. 6A). However, diNR tracts localized preferentially to the 3'-UTR region, whereas triNRs localized preferentially in the ORF (Supplemental Fig. 6B). In addition, poly(Q) and poly(E) were the most commonly encoded amino acids, whereas no poly-aromatic amino acids (F, Y, and W) were found (Supplemental Fig. 6C). These results support the hypothesis that codons used for 3-C unbranched amino acids (Supplemental Fig. 6D) were preferentially recruited to perform protein functions, perhaps disordered regions involved in allosteric interactions (Perutz et al. 2002; Liu et al. 2006; Minezaki et al. 2006; Hilser and Thompson 2007; Friedman et al. 2008). Poly(Q) and poly(E) runs were encoded predominantly by the first translated exon, whereas the poly(A), poly(L), poly(G), and poly(P) amino acids were encoded mostly by subsequent exons (Supplemental Fig. 6C, inset). This partitioning suggests a differential localization within the folded native protein. Finally, to assess the extent of evolutionary conservation of homopolymeric amino acid runs, we analyzed three genes (*TBP*, *MEF2A*, and *POU4F2*); strong conservation (Supplemental Fig. 7) of two CAG:CTG and one GGC:GCC repeat tracts encoding poly(Q) and poly(G), respectively, was found. In summary, simple repeats have been recruited by a large network of regulatory genes, mostly related to nervous system activity, to affect both gene and protein function.

### Repeat length distributions and base-stacking

In addition to the disparity in relative abundance (Tables 1, 2), genomic tetraNRs and triNRs also manifest variable length dis-

tributions in the human genome (Supplemental Fig. 8). A compilation of tetraNR length distributions in the nine vertebrate genomes examined revealed that only a few sequences (e.g., AGAT, ATCC, AAGG, and AAAG) display extended length distributions (Supplemental Fig. 9). Inspection of the tetraNR (Supplemental Figs. 8A, 9) and triNR (Supplemental Fig. 8B; Clark et al. 2006) sequences that formed the longest distributions in the primate lineage (human and chimpanzee) reveals their R:Y-rich nature, suggestive of a role for R:Y asymmetry in supporting extended repeat lengths over evolutionary time.

Analyses of the TDAS profiles indicated a consistent and pronounced “S-behavior” (Tables 1, 2) for the R-rich single-stranded oligonucleotides that corresponded to the longest tetraNR and triNR distributions in the human genome (Supplemental Fig. 8). Because the “S-behavior” is elicited by base-stacking interactions (Applequist and Damie 1966; Powell et al. 1972; Cantor and Schimmel 1980; Friedman and Honig 1995), we considered whether base-stacking interactions might facilitate genomic expansion. Analysis of the TDAS hypochromicity curves (Table 3) yielded qualitative evidence for a direct relationship between the TDAS slope values and the genomic length distributions.

To verify that the slope values faithfully reflected the energetics of base-stacking, the average theoretical free energy contributions (Friedman and Honig 1995) to the nearest-neighbor base-stacking ( $\Delta G_v$ ) were evaluated (Table 3). For both the tetraNR ( $r^2 = 0.76$ ,  $P = 0.02$ ) (data not shown) and triNR oligonucleotides, ( $\Delta G_v$ ) correlated negatively with the slope values (Table 3; Supplemental material), supporting the view that the dynamics of base-stacking are accurately represented by the TDAS slope values (Applequist and Damie 1966). In conclusion, both the TDAS curves and the theoretical  $\Delta G_v$  calculations indicated a role for base-stacking in promoting long triNR and tetraNR lengths in the human genome.

To investigate these relationships quantitatively, we determined the average number of base pairs for the 10 longest tracts for each of the six genomic tetraNR and three triNR sequences

(Table 3). With the sole exception of the AAAG sequence, both the TDAS slopes and the ( $\Delta G_{\text{tr}}$ ) values followed the rank order of genomic mean repeat lengths. The mean values for the tetraNR tract lengths correlated positively with both the TDAS slopes ( $r^2 = 0.96$ ,  $P = 0.007$ ) and the ( $\Delta G_{\text{tr}}$ ) values ( $r^2 = 0.95$ ,  $P = 0.005$  without AAAG;  $r^2 = 0.62$ ,  $P = 0.06$  with AAAG). In contrast, no correlations were found when the stacking energies in double-stranded (Cantor and Schimmel 1980; Hunter and Lu 1997; Isaksson et al. 2004; SantaLucia and Hicks 2004; Sponer et al. 2006), rather than single-stranded, DNA were analyzed ( $r^2 \leq 0.02$ ,  $P = \text{not significant [NS]}$ ). In summary, nearest-neighbor base-stacking interactions within the R-rich strand of genomic simple repeat sequences played a key role in acquiring (and subsequently maintaining) considerable lengths over the course of vertebrate evolution.

## Discussion

We previously documented the variable abundances of simple repeating sequences in vertebrate genomes (Bacolla et al. 2006). Herein, model-building studies suggested that non-B DNA structures could be responsible for this behavior, and subsequent coil-to-helix transition analyses on repeating tetraNR and triNR oligonucleotides revealed an inverse relationship between the capacity of these sequences to form DNA secondary structures and their abundance in nine vertebrate genomes. These relationships also revealed the action of positive selection in maintaining unstable repeats within coding regions in the human genome and the recruitment of simple repeats by genes involved in regulatory and signaling pathways, particularly of the nervous system. Furthermore, nearest-neighbor base-stacking interactions correlated directly with the repeat sequences that manifested a bias toward expansion. These remarkably simple findings crystallize the concept that intrinsic structural features of DNA played a fundamental role as cellular recognition targets over evolutionary time.

Figure 2 shows a model for the relationship between stable hairpin formation and the “evolutionary fitness” of DNA motifs. If a repeating sequence can form a stable hairpin during the process of DNA replication or transcription, this self base-paired tract may be bypassed by DNA replication (left side, a) or induce DSBs

(left side, b), which then trigger deletions resulting, over evolutionary time, in the loss of the underlying duplex DNA. Alternatively, if the sequence forms a less stable hairpin, it may generate longer repeat-containing alleles as a consequence of DNA slippage (Bowater et al. 1997; Lin et al. 2006; Wang and Vasquez 2006; Wells and Ashizawa 2006; Mirkin 2007; Wells 2007). These longer alleles will be maintained in the population and will go on to generate further length polymorphisms (Fig. 2, right side, alleles 1–3). Functional repeat polymorphisms within human genes may be associated with variable phenotypic traits (filled bars), such as blood pressure, heart rate, and muscular tension (Supplemental Table 1). Such traits have the potential to confer either increased fitness (allele 2) or allele-specific susceptibility to complex diseases (allele 3) or repeat expansion disorders (Supplemental Table 1).

## Relationships with gene function and disease

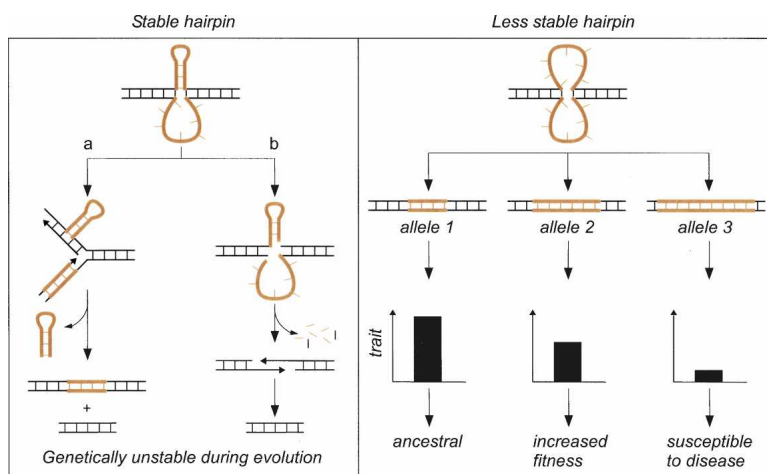
An analysis of genes in which the association between triNR and tetraNR length polymorphisms and phenotypic variation and/or susceptibility to inherited disease in the general population was reported (Supplemental Table 1) revealed enrichment in development, transcriptional regulation, cell signaling, and nervous system activities (Supplemental Table 4). Hence, the ability of gene-associated simple repeats to expand and contract within relatively short evolutionary time frames could have contributed to gene and protein structure/function in vertebrate genomes (Supplemental Table 3; Legendre et al. 2007). In the process, however, repeat polymorphism may also have generated new risk factors for disease susceptibility. Gene classes involved in cell adhesion and cell–cell communication are also characterized by long intronic R:Y tracts (Bacolla et al. 2006) and are associated with high meiotic recombination rates (Frazer et al. 2007; Freudenberg et al. 2007). Hence, repeating DNA and meiotic recombination may have facilitated selective pressure over evolutionary time.

In the context of trinucleotide repeat expansion disorders, the CAG:CTG and CCG:CGG repeats have been preserved in coding regions in the human genome in spite of their inherent instability; the mechanisms involved remain speculative. In the case of FA, the frequency of carriers with expanded GAA:TTC

repeats is estimated at ~1:500 in the general population (De Biase et al. 2006); to our knowledge, this is the only example of a triNR/tetraNR of such length to be stably maintained in the human population. The unique base-stacking behavior of the GAA:TTC repeat is likely to contribute to its maintenance over the generations. Similarly, strong base-stacking interactions may also be responsible for the preferential instability at AAAG and AAGG repeats in specific malignancies (Supplemental Table 2; Ahrendt et al. 2000; Xu et al. 2001).

## Role of base-stacking in repeat expansion

Base-stacking is emerging as a key component of DNA repair since the efficiency of this activity by diverse enzyme systems is inversely proportional to the



**Figure 2.** Model for a relationship among repeat abundance, DNA structure, repeat polymorphism, and variable phenotype. (Orange) triNRs or tetraNRs. Note that overall repeat lengths are not drawn to scale and that the choice of decreased gene expression with increasing repeat length is arbitrary.

target stacking strength (for review, see Yang 2006). Base-stacking may potentiate the expansion of repeat sequences by promoting replication slippage and at the same time protecting any ensuing secondary structures from repair (Jucker et al. 1996). These properties may contribute to avoidance of DNA repair and could be responsible for the unique lengths and high mutation rates (Kelkar et al. 2008) observed for the AAG, AAAG, and AAGG repeating sequences in the human genome.

### Mechanisms of sequence loss

The nature of the mechanisms that lead to DNA structure-dependent sequence loss is unclear. However, hairpin formation is an obligatory step in programmed V(D)J recombination, where the ARTEMIS/DNA-PKcs (ARTEMIS also known as DCLRE1C) complex cleaves the RAG-induced hairpin structures in human B- and T-cells (Ma et al. 2002). In other tissues, such as mouse liver, muscle, heart, and kidney, the ARTEMIS/DNA-PKcs activity cleaves the terminal hairpins of recombinant adeno-associated virus (rAAV) particles used in gene therapy, thereby contributing to viral replication (Inagaki et al. 2007). An alternative pathway to process hairpins is the Holliday-junction resolvase activity (Inagaki et al. 2007), which comprises RAD51C in humans (for review, see Sharan and Kuznetsov 2007). Hence, at least two enzymatic activities are known to process hairpin structures in mammals.

Cleaved hairpins/cruciforms may trigger loss of the underlying DNA sequences by two mechanisms. First, opening of the single-stranded loops by the ARTEMIS/DNA-PKcs complex or incisions at the 3-/4-way junctions by the Holliday junction resolvase may initiate DNA repair and exonucleolytic cleavage, followed by single-strand annealing (Al-Minawi et al. 2008) or non-homologous end-joining (Inagaki et al. 2007), respectively. Second, cleaved hairpins may impose a mutational burden with an ensuing growth disadvantage for the organism (Tanaka et al. 2007). Analogously, an excessive reservoir of hairpin-forming sequences may overwhelm cellular defense mechanisms and lead to deleterious rearrangements (Bacolla and Wells 2004; Wang and Vasquez 2006; Wells 2007).

In summary, DNA secondary structures played a key role in determining the number and length distributions of microsatellite repeats in vertebrate genomes over evolutionary time. Concomitantly, microsatellite length polymorphism may have served to modulate both gene expression and protein function, thereby contributing to, but at times also hampering, cellular regulatory circuitries.

## Methods

### Oligonucleotides

The sequences of the 82 single-stranded oligonucleotides used in this study are given in Tables 1 and 2. HPLC-purified synthetic oligonucleotides containing nine copies of each tetraNR and 12 copies of each triNR (to give 36-base molecules in all cases) sequences were purchased from Sigma Genosys.

### Temperature-dependent absorption spectroscopy (TDAS)

#### Standard assay conditions

Oligonucleotides (0.6–0.8 OD<sub>260</sub>/mL, 1–3 μM) were dissolved in buffer 1 (50 mM KCl, 0.5 mM MgCl<sub>2</sub>, 0.4 mM Na-phosphate at pH 7.0) and equilibrated overnight at 25°C. The optical absor-

bance at 260 nm was measured on a Cary 3 Bio UV-Vis spectrophotometer equipped with a Cary temperature controller with heating (melting curve) from 10°C to 94°C and cooling (annealing curve) from 94°C to 10°C (the temperature range was extended to 4°C for the triNR oligos) at a rate of 0.75°C/min. Minimum hysteresis effects were observed, except when noted. The melting curves often showed small peaks not present during the cooling step, suggesting the slow formation of multiple DNA conformations during the overnight incubation. For these reasons, and because the annealing step simulated more closely the biologically relevant folding of single strands into hairpins than the melting step, only the cooling curves were used to determine the midpoint of transitions. We defined as  $T_m$  the temperature at which the midpoint of transition occurred during the melting step (helix-to-coil transition) and  $T_a$  the temperature at which the midpoint of transition occurred during the annealing step (coil-to-helix transition).

#### Other pH assay conditions

For the d(CCCT)<sub>9</sub> and d(CCT)<sub>12</sub> oligonucleotides, for which strong hysteresis effects were observed, TDAS measurements were performed at three different pH values in the following buffers: 0.5 mM Tris-acetate (pH 4.5), 50 mM KCl, 0.5 mM MgCl<sub>2</sub>; 0.5 mM Tris-acetate (pH 7.0), 50 mM KCl, 0.5 mM MgCl<sub>2</sub>; 0.5 mM Tris-HCl (pH 8.0), 50 mM KCl, 0.5 mM MgCl<sub>2</sub>, and both the  $T_m$  and  $T_a$  values were determined.

#### Salt concentration assay conditions

Variable salt concentration assay conditions were used to distinguish between nonhelical single-stranded coils and hydrogen-bonded, highly structured, helices with  $T_m$  and  $T_a$  values >94°C. These TDAS measurements were performed in buffer 2 (10 mM Tris-HCl, 10 μM EDTA at pH 7.4) with or without KCl (1, 5, 10, 50, 100, 500 mM), NaCl (1, 10, 50, 100 mM), LiCl (1, 10, 50 mM), MgCl<sub>2</sub> (0.5 and 10 mM), and ±50% formamide.

#### $T_m$ and $T_a$ determinations

The raw absorbance curves were smoothed with a Lowess function ( $f = 0.10$ – $0.20$ ) and used to obtain the first derivative with cubic spline functions. The first derivative curves were interpolated with peak curves (SigmaPlot 8.02; SPSS Inc.), and the best fits were chosen to obtain the  $T_m$  and  $T_a$  values from the peak parameter ( $r^2 = >0.95$  in most cases).

#### Slopes of hypochromicity curves

For the d(AAAT)<sub>9</sub>, d(AAAG)<sub>9</sub>, d(AAAC)<sub>9</sub>, d(AAC)<sub>12</sub>, d(AAT)<sub>12</sub>, and d(AAG)<sub>12</sub> oligonucleotides, which displayed near-linear TDAS hypochromicity curves, the slopes of the cooling curves were taken over the entire temperature range (10°C–94°C). For the d(AAGG)<sub>9</sub> and d(AGAT)<sub>9</sub> oligonucleotides, the slopes of the cooling curves were taken from 94°C to the temperature that preceded the annealing transitions; these portions of the curves were near-linear. For the d(AAGG)<sub>9</sub> oligonucleotide, two near-linear segments were observed: one less steep from 94°C to ~60°C, and a second, steeper one from ~60°C to ~30°C, which was used to obtain the slope value. In all cases,  $r^2 \geq 0.94$  for the near-linear segments.

#### Circular dichroism (CD)

CD studies were performed to monitor quadruplex formation by the d(XGGG)<sub>9</sub> oligonucleotides (where X = A, C, or T). The ellipticity was monitored on a Jasco J-720 spectropolarimeter over the 220–320-nm range for oligonucleotide solutions (0.6–0.8 OD<sub>260</sub>/mL, 1–3 μM) at 25°C by using the variable salt concentration

assay conditions (see above) with and without K<sup>+</sup>, Na<sup>+</sup>, Li<sup>+</sup> (1–100 mM), and Mg<sup>2+</sup> (0.5 and 10 mM) ions.

### Repeat searches

Computer searches were performed on a variety of eukaryotic genome sequences to retrieve all genomic triNR and tetraNR tracts comprising at least three tandem units (e.g., ACGACGACG or ACGGACGGACGG). All tracts with 3, 4, 5 units, and so on (no upper limits were set) were binned, and within each bin the total number of tracts was given by all possible reading frames and strand complementarities, to yield the 33 unique tetraNR sequences and the 10 unique triNR sequences (Tables 1 and 2, respectively, “Unique genomic sequence”). Computer searches (Collins et al. 2003) were performed on the following genomes: Nine vertebrate genomes: human (*Homo sapiens*, hg18, 18 March 2006, NCBI Build 36.1, database version 44.36f), chimpanzee (*Pan troglodytes*, panTro2), mouse (*Mus musculus*, mm8), rat (*Rattus norvegicus*, rn4), dog (*Canis familiaris*, canFam2), cow (*Bos taurus*, bosTau2), chicken (*Gallus gallus*, galGal3), zebrafish (*Danio rerio*, danRer4), and *Fugu* (*Takifugu rubripes*, fr2); four nonvertebrate genomes: plants [*Arabidopsis thaliana*, tair7 and rice (*Oryza sativa* ssp. japonica, release 5)], nematode worm (*Caenorhabditis elegans*, ce2), and yeast (*Saccharomyces cerevisiae*, sacCer1).

### Database searches

Studies describing associations between human intragenic tetraNR and triNR length polymorphisms and inherited disease or the occurrence of tetraNR instability in cancer were retrieved from the Human Gene Mutation Database (<http://www.hgmd.org>; Stenson et al. 2003) and from manual PubMed searches.

### Gene enrichment analyses

Functional category enrichment analyses were performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) at <http://david.abcc.ncifcrf.gov>. Only the gene categories most enriched in each proteomic/genomic database were reported.

### Acknowledgments

This work was supported by the NIH (ES11347), Friedreich's Ataxia Research Alliance, Seek-a-Miracle Foundation, and the Robert A. Welch Foundation to R.D.W. and in part by the Intramural Research Program of the NIH, NCI, and Federal funds from the NCI, NIH to J.R.C. (contract no. N01-CO-12400), and financial support from BIOBASE GmbH to D.N.C. We thank J.E.L. for 40 years of research on non-B DNA structures. All research materials from the R.D.W. laboratory have been transferred to S. Mirkin (Sergei.Mirkin@tufts.edu). We thank Xiaolian Gao of the University of Houston for the use of her facilities and Jin Jen of the NCI for helpful discussions.

### References

- Ahrendt, S.A., Decker, P.A., Doffek, K., Wang, B., Xu, L., Demeure, M.J., Jen, J., and Sidransky, D. 2000. Microsatellite instability at selected tetranucleotide repeats is associated with p53 mutations in non-small cell lung cancer. *Cancer Res.* **60**: 2488–2491.
- Al-Minawi, A.Z., Saleh-Gohari, N., and Helleday, T. 2008. The ERCC1/XPF endonuclease is required for efficient single-strand annealing and gene conversion in mammalian cells. *Nucleic Acids Res.* **36**: 1–9.
- Appelquist, J. and Damie, V. 1966. Thermodynamics of the one-stranded helix-coil equilibrium in polyadenylic acid. *J. Am. Chem. Soc.* **88**: 3895–3900.
- Bacolla, A. and Wells, R.D. 2004. Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.* **279**: 47411–47414.
- Bacolla, A., Pradhan, S., Larson, J.E., Roberts, R.J., and Wells, R.D. 2001. Recombinant human DNA (cytosine-5) methyltransferase. III. Allosteric control, reaction order, and influence of plasmid topology and triplet repeat length on methylation of the fragile X CCG.CCG sequence. *J. Biol. Chem.* **276**: 18605–18613.
- Bacolla, A., Collins, J.R., Gold, B., Chuzhanova, N., Yi, M., Stephens, R.M., Stefanov, S., Olsh, A., Jakupciak, J.P., Dean, M., et al. 2006. Long homopurine\*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Res.* **34**: 2663–2675.
- Bowater, R.P., Jaworski, A., Larson, J.E., Parniewski, P., and Wells, R.D. 1997. Transcription increases the deletion frequency of long CTG.CAG triplet repeats from plasmids in *Escherichia coli*. *Nucleic Acids Res.* **25**: 2861–2868.
- Cantor, C.R. and Schimmel, P.R. 1980. *Biophysical chemistry*. W.H. Freeman & Co., New York.
- Clark, R.M., Bhaskar, S.S., Miyahara, M., Dalgliesh, G.L., and Bidichandani, S.I. 2006. Expansion of GAA trinucleotide repeats in mammals. *Genomics* **87**: 57–67.
- Collins, J.R., Stephens, R.M., Gold, B., Long, B., Dean, M., and Burt, S.K. 2003. An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics* **82**: 10–19.
- De Biase, I., Rasmussen, A., and Bidichandani, S.I. 2006. Evolution and instability of the GAA triplet-repeat sequence in Friedreich's Ataxia. In *Genetic instabilities and neurological diseases* (eds. R.D. Wells and T. Ashizawa), pp. 305–319. Elsevier/Academic Press, San Diego.
- Elango, N., Kim, S.H., Vigoda, E., and Yi, S.V. 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. *PLoS Comput. Biol.* **4**: e1000015. doi: 10.1371/journal.pcbi.1000015.
- Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.
- Frederico, L.A., Kunkel, T.A., and Shaw, B.R. 1993. Cytosine deamination in mismatched base pairs. *Biochemistry* **32**: 6523–6530.
- Freudenberg, J., Fu, Y.H., and Ptacek, L.J. 2007. Enrichment of HapMap recombination hotspot predictions around human nervous system genes: Evidence for positive selection? *Eur. J. Hum. Genet.* **15**: 1071–1078.
- Friedman, R.A. and Honig, B. 1995. A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophys. J.* **69**: 1528–1535.
- Friedman, M.J., Wang, C.E., Li, X.J., and Li, S. 2008. Polyglutamine expansion reduces the association of TATA-binding protein with DNA and induces DNA binding-independent neurotoxicity. *J. Biol. Chem.* **283**: 8283–8290.
- Hilser, V.J. and Thompson, E.B. 2007. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl. Acad. Sci.* **104**: 8311–8315.
- Hunter, C.A. and Lu, X.J. 1997. DNA base-stacking interactions: A comparison of theoretical calculations with oligonucleotide X-ray crystal structures. *J. Mol. Biol.* **265**: 603–619.
- Inagaki, K., Ma, C., Storm, T.A., Kay, M.A., and Nakai, H. 2007. The role of DNA-PKcs and Artemis in opening viral DNA hairpin termini in various tissues in mice. *J. Virol.* **81**: 11304–11321.
- Isaksson, J., Acharya, S., Barman, J., Cheruku, P., and Chattopadhyaya, J. 2004. Single-stranded adenine-rich DNA and RNA retain structural characteristics of their respective double-stranded conformations and show directional differences in stacking pattern. *Biochemistry* **43**: 15996–16010.
- Iyer, R.R., Pluciennik, A., Rosche, W.A., Sinden, R.R., and Wells, R.D. 2000. DNA polymerase III proofreading mutants enhance the expansion and deletion of triplet repeat sequences in *Escherichia coli*. *J. Biol. Chem.* **275**: 2174–2184.
- Jucker, F.M., Heus, H.A., Yip, P.F., Moors, E.H., and Pardi, A. 1996. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.* **264**: 968–980.
- Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., and Makova, K.D. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.* **18**: 30–38.
- Kovtun, I.V. and McMurray, C.T. 2008. Features of trinucleotide repeat instability in vivo. *Cell Res.* **18**: 198–213.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- Legendre, M., Pochet, N., Pak, T., and Verstrepen, K.J. 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* **17**: 1787–1796.

- Lin, Y., Dion, V., and Wilson, J.H. 2006. Transcription promotes contraction of CAG repeat tracts in human cells. *Nat. Struct. Mol. Biol.* **13**: 179–180.
- Lindahl, T. and Nyberg, B. 1974. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**: 3405–3410.
- Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K. 2006. Intrinsic disorder in transcription factors. *Biochemistry* **45**: 6873–6888.
- Ma, Y., Pannicke, U., Schwarz, K., and Lieber, M.R. 2002. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell* **108**: 781–794.
- Minezaki, Y., Homma, K., Kinjo, A.R., and Nishikawa, K. 2006. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *J. Mol. Biol.* **359**: 1137–1149.
- Mirkin, S.M. 2007. Expandable DNA repeats and human disease. *Nature* **447**: 932–940.
- Mort, M., Ivanov, D., Cooper, D.N., and Chuzhanova, N.A. 2008. A meta-analysis of nonsense mutations causing human genetic disease. *Hum. Mutat.* **29**: 1037–1047.
- Orr, H.T. and Zoghbi, H.Y. 2007. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* **30**: 575–621.
- Pandolfo, M. 2006. Friedreich's ataxia. In *Genetic instabilities and neurological diseases* (eds. R.D. Wells and T. Ashizawa), pp. 277–296. Elsevier/Academic Press, San Diego.
- Perutz, M.F., Pope, B.J., Owen, D., Wanker, E.E., and Scherzinger, E. 2002. Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid beta-peptide of amyloid plaques. *Proc. Natl. Acad. Sci.* **99**: 5596–5600.
- Powell, J.T., Richards, E.G., and Gratzner, W.B. 1972. The nature of stacking equilibria in polynucleotides. *Biopolymers* **11**: 235–250.
- Rachwal, P.A., Brown, T., and Fox, K.R. 2007. Sequence effects of single base loops in intramolecular quadruplex DNA. *FEBS Lett.* **581**: 1657–1660.
- Sammalkorpi, H., Alhopuro, P., Lehtonen, R., Tuimala, J., Mecklin, J.P., Jarvinen, H.J., Jiricny, J., Karhu, A., and Aaltonen, L.A. 2007. Background mutation frequency in microsatellite-unstable colorectal cancer. *Cancer Res.* **67**: 5691–5698.
- SantaLucia Jr., J. and Hicks, D. 2004. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* **33**: 415–440.
- Sharan, S.K. and Kuznetsov, S.G. 2007. Resolving *RAD51C* function in late stages of homologous recombination. *Cell Div.* **2**: 15. doi: 10.1186/1747-1028-2-15.
- Sharma, P.C., Grover, A., and Kahl, G. 2007. Mining microsatellites in eukaryotic genomes. *Trends Biotechnol.* **25**: 490–498.
- Sponer, J., Jurecka, P., Marchan, I., Luque, F.J., Orozco, M., and Hobza, P. 2006. Nature of base stacking: Reference quantum-chemical stacking energies in ten unique B-DNA base-pair steps. *Chemistry* **12**: 2854–2865.
- Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeyasinghe, S., Krawczak, M., and Cooper, D.N. 2003. Human Gene Mutation Database (HGMD): 2003 Update. *Hum. Mutat.* **21**: 577–581.
- Subramanian, S., Mishra, R.K., and Singh, L. 2003. Genome-wide analysis of microsatellite repeats in humans: Their abundance and density in specific genomic regions. *Genome Biol.* **4**: R13. doi: 10.1186/gb-2003-4-2-r13.
- Tanaka, H., Cao, Y., Bergstrom, D.A., Kooperberg, C., Tapscott, S.J., and Yao, M.C. 2007. Intrastrand annealing leads to the formation of a large DNA palindrome and determines the boundaries of genomic amplification in human cancer. *Mol. Cell. Biol.* **27**: 1993–2002.
- Walsh, C.P. and Xu, G.L. 2006. Cytosine methylation and DNA repair. *Curr. Top. Microbiol. Immunol.* **301**: 283–315.
- Wang, G. and Vasquez, K.M. 2006. Non-B DNA structure-induced genetic instability. *Mutat. Res.* **598**: 103–119.
- Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.
- Wells, R.D. 2007. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci.* **32**: 271–278.
- Wells, R.D. and Ashizawa, T. 2006. *Genetic instabilities and neurological diseases*. Elsevier/Academic Press, San Diego.
- Wells, R.D., Dere, R., Hebert, M.L., Napierala, M., and Son, L.S. 2005. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res.* **33**: 3785–3798.
- Wojciechowska, M., Napierala, M., Larson, J.E., and Wells, R.D. 2006. Non-B DNA conformations formed by long repeating tracts of DM1, DM2 and FRDA genes, not the sequences per se, promote mutagenesis in flanking regions. *J. Biol. Chem.* **281**: 24531–24543.
- Xu, L., Chow, J., Bonacum, J., Eisenberger, C., Ahrendt, S.A., Spafford, M., Wu, L., Lee, S.M., Piantadosi, S., Tockman, M.S., et al. 2001. Microsatellite instability at AAG repeat sequences in respiratory tract cancers. *Int. J. Cancer* **91**: 200–204.
- Yang, W. 2006. Poor base stacking at DNA lesions may initiate recognition by many repair proteins. *DNA Repair* **5**: 654–666.
- Zahra, R., Blackwood, J.K., Sales, J., and Leach, D.R. 2007. Proofreading and secondary structure processing determine the orientation dependence of CAG-CTG trinucleotide repeat instability in *Escherichia coli*. *Genetics* **176**: 27–41.

Received March 10, 2008; accepted in revised form July 31, 2008.