Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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. published online August 7, 2008
Access the most recent version at doi:10.1101/gr.078303.108

P<P Published online August 7, 2008 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

License

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

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Abundance and Length of Simple Repeats in Vertebrate Genomes are Determined by their Structural Properties

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

[1] Institute of Biosciences and Technology, Center for Genome Research, Texas A&M University Health Science Center, 2121 West Holcombe Blvd., Houston, TX 77030

[2] Advanced Biomedical Computing Center, Advanced Technology Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD 21702

[3] Department of Molecular and Human Genetics and [4] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA.

[5] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK

[#] To whom correspondence should be addressed:

Phone (713) 677-7660, Fax (713) 677-7689, Email: abacolla@ibt.tamhsc.edu

Running title: Microsatellite repeats and DNA structure

Keywords: microsatellites, cancer, repeat polymorphisms, inherited diseases, CD, temperature-dependent absorption spectroscopy, hairpins, evolution, vertebrate genomes.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Abstract

Microsatellites are abundant in vertebrate genomes but their sequence representation and length distributions vary greatly within each family of repeats (*e.g*. tetranucleotides, etc.). Biophysical studies of 82 synthetic single-stranded oligonucleotides comprising all tetra- and tri-nucleotide repeats revealed an inverse correlation between the stability of folded-back hairpin and quadruplex structures and the sequence representation for repeats ≥30 bp in length in 9 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 overrepresented 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 tetra-nucleotide 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.

This manuscript contains Supplementary Information.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Introduction

DNA microsatellites, tandem arrays of simple repeats such as mono-, di-, tri- and tetra-nucleotides, are common in eukaryotic genomes (Lander et al. 2001; Sharma et al. 2007; Waterston et al. 2002). 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 ≥30 nts. For example, >16,000 tracts of mononucleotides comprised of ≥30 As or Ts are present in the human genome but only 7 analogous tracts of Gs or Cs are found (Bacolla et al. 2006).

$^{5'}$CpG$^{3'}$- (CpG)-containing repeats are generally rare, suggesting that cytosine methylation and subsequent deamination leading to T:A transitions from ^{5m}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 ^{5m}C deamination (Frederico et al. 1993; Lindahl and Nyberg 1974) 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 relationships remain to be clarified. Other repeats, such as ATGC$^{(1)}$, 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

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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 (reviewed in Supplementary Table 1). Also, more than twenty neurological diseases (Supplementary Table 1) are caused by the expansion of triNRs within the coding or untranslated regions of genes (reviewed in (Kovtun and McMurray 2008; Mirkin 2007; Orr and Zoghbi 2007; Wells and Ashizawa 2006)). Among these, recessive Friedreich ataxia (FA) is atypical since, contrary to all other triNR diseases, the expanded GAA:TTC tract (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) has been noted in mismatch repair-proficient cancers of the respiratory tract, skin and bladder and it involves preferentially biased purine:pyrimidine (R:Y) sequences (reviewed in Supplementary 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

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

R:Y-rich sequences, and their length distributions in the human genome correlate directly with the strength of base stacking interactions 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 factor and other regulatory genes to perform specific functions, particularly in the nervous system.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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 ≥8 identical units (Table 1). For some nt combinations the number of tracts exceeded 5000 copies, whereas for others no tracts were found (Table 1). This diversity, spanning nearly four orders of magnitude, was intriguing. Based on molecular modeling (Supplementary Information), 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 (Mirkin 2007; Wojciechowska et al. 2006). Alternatively, these structures could be bypassed by the DNA replication complex (Iyer et al. 2000; Mirkin 2007; Wells et al. 2005; 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. By contrast, those sequences unable to adopt stable folded conformations would tend to survive, giving rise to extended length distributions.

***T_a* 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

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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 to 16.7° C. The highest T_a values were observed for the self-complementary sequences, *i.e.* $d(ACGT)_9$, $d(ATCG)_9$, $d(ATGC)_9$ and $d(AGCT)_9$, with the exception of $d(CCGG)_9$ for which no cooperative transition was observed. These genomic tetraNRs were either extremely rare (≤2 occurrences) or absent. By contrast, the self-complementary $d(AATT)_9$ 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 $d(AGCG)_9$, $d(GCGT)_9$, $d(ACGG)_9$, $d(GGCT)_9$ and $d(AGGC)_9$. Hence, adjacent CG:CG and GC:GC base-pairs contributed the most stability to the tetraNR hairpins.

For the the $d(CCCT)_9$ oligonucleotide, a strong hysteresis effect was observed, with melting temperature (T_m) values of ~80.0° C and T_a values of ~30.0 °C. Additional TDAS determinations (Supplementary Information) indicated that this behavior was due to slow hairpin-formation by partially protonated cytosine pairs (C^+: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-

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

stranded coils or very stable hydrogen-bonded structures with T_a and T_m values >94° C. To resolve these ambiguities, further TDAS and CD investigations were conducted (Supplementary Information and Supplementary Fig. 1), which revealed that the $d(CCGG)_9$ and $d(XGGG)_9$ (X=C, T or A) oligonucleotides formed highly stable hairpin and quadruplex structures under physiologic K^+ and Mg^{2+} concentrations, respectively. These 4 sequences were rare in the human genomes and the numbers of quadruplex-forming repeats correlated inversely with structure stability (Rachwal et al. 2007). By 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 and see Supplementary Information for a rational of this analysis). A linear and inverse correlation was found (r = -0.607, P=0.028), a

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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 ≥8 units as if methylation-mediated and deamination-dependent C:G → T:A transitions had not occurred (Supplementary Information). The correlation remained significant (P=0.007).

Next, we reasoned that if the T_a *vs*. 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 8 additional vertebrate genomes indicated that in spite of large variations in the absolute numbers of tetraNRs of ≥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 upon DNA sequence composition. The relative tetraNR abundances for the 9 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 ≥8 units were found for CCGG, the most thermostable hairpin-forming tetraNR sequence, in any of the 9 genomes. The r^2 values decreased hyperbolically when shorter tetraNR lengths were considered (Supplementary Fig. 2); nevertheless, all regressions remained significant (P = 0.03 for tract-lengths ≥3 units). This trend is in agreement with the dependence

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

of base-paired stem-loop stability upon DNA length (Supplementary Fig. 3) and suggested that *a*) longer tracts were deleted more efficiently than shorter tracts and *b*) 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 ≥4 units in 4 non-vertebrate genomes that exhibited shorter length distributions than those noted in vertebrate genomes (Materials and 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 distribution of triNRs ≥10 units in the 9 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 – 76.7° C), and group 2, comprising the other 7 repeats, for which T_a values were low (≤27° C and Supplementary Information).

The group 1 triNRs were generally of lower abundance than the group 2 triNRs). One notable exception was the AGC sequence, which occurred ~10

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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 (reviewed in (Kovtun and McMurray 2008; Mirkin 2007; Orr and Zoghbi 2007; Wells and Ashizawa 2006)). 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 *vs.* intragenic distributions of triNRs and tetraNRs in the human genome. At lengths ≥30 bp, both the AGC and CCG triNRs were highly (~30%) overrepresented within coding regions (Supplementary Figs. 4A and 5A), indicating selection for both sequences. At lengths ≥12 bp, the group 1 triNRs and the CCCG, AGCG,

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

CCGG and ACCG tetraNRs were also overrepresented (>10%) within coding regions (Supplementary Figs. 4B and 5B). Since all these repeats are rare at lengths ≥30 bp (Tables 1 and 2) and are associated with high (>50° 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 ~2300 non-redundant genes, strong enrichment (P 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, Supplementary 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, Supplementary Fig. 6A). However, diNR tracts localized preferentially to the 3'UTR region whereas triNRs localized preferentially in the ORF (Supplementary Fig. 6B). In addition, polyQ and polyE were the most commonly encoded amino acids, whereas no poly-aromatic amino acids (F, Y and W) were found (Supplementary Fig. 6C). These results support the hypothesis that codons used for 3-C unbranched amino acids (Supplementary Fig. 6D) were preferentially recruited to perform protein functions; perhaps disordered regions involved in allosteric interactions (Friedman et al. 2008; Hilser and Thompson 2007; Liu et al. 2006; Minezaki et al. 2006; Perutz et al. 2002).

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

PolyQ and polyE runs were encoded predominantly by the first translated exon, whereas the polyA, polyL, polyG and polyP amino acids were encoded mostly by subsequent exons (Supplementary 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 3 genes (*TBP*, *MEF2A* and *POU4F2*); strong conservation (Supplementary Fig. 7) of two CAG:CTG and one GGC:GCC repeat tracts encoding polyQ and polyG, respectively, was found. In summary, simple repeats have been recruited by a large network of regulatory genes, mostly related to nervous system activity, to effect both gene and protein function.

Repeat length distributions and base stacking. In addition to the disparity in relative abundance (Tables 1 and 2), genomic tetraNRs and triNRs also manifest variable length distributions in the human genome (Supplementary Fig. 8). A compilation of tetraNR length distributions in the 9 vertebrate genomes examined revealed that only a few sequences (e.g. AGAT, ATCC, AAGG and AAAG) display extended length distributions (Supplementary Fig. 9). Inspection of the tetraNR (Supplementary Figs. 8A and 9) and triNR ((Clark et al. 2006) and Supplementary Fig. 8B) 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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Analyses of the TDAS profiles indicated a consistent and pronounced "*S*-behavior" (Tables 1 and 2) for the R-rich single-stranded oligonucleotides that corresponded to the longest tetraNR and triNR distributions in the human genome (Supplementary Fig. 8). Because the "*S*-behavior" is elicited by base stacking interactions (Applequist and Damie 1966; Cantor and Schimmel 1980; Friedman and Honig 1995; Powell et al. 1972), 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 (ΔG_v) were evaluated (Table 3). For both the tetraNR (r^2=0.76, P=0.02, not shown) and triNR oligonucleotides, (ΔG_v) correlated negatively with the slope values (Table 3 and Supplementary Information), 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 ΔG_v calculations indicated a role for base stacking in promoting long triNR and tetraNR lengths in the human genome.

To investigate quantitatively these relationships, we determined the average number of bp for the 10 longest tracts for each of the 6 genomic tetraNR and 3 triNR sequences (Table 3). With the sole exception of the AAAG

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

sequence, both the TDAS slopes and the (ΔG_V) 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 (ΔG_V) values (r^2=0.95, P=0.005 without AAAG, r^2=0.62, P=0.06 with AAAG). By 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≤0.02, P=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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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 9 vertebrate genomes. These relationships also revealed the action of positive selection in maintaining unstable repeats within coding region 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 towards 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.

Fig. 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; Mirkin 2007;

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Wang and Vasquez 2006; Wells 2007; Wells and Ashizawa 2006). 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 (Supplementary 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 (Supplementary 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 (Supplementary Table 1), revealed enrichment in development, transcriptional regulation, cell signaling and nervous system activities (Supplementary 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 ((Legendre et al. 2007) and Supplementary Table 3). 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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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 ((Ahrendt et al. 2000; Xu et al. 2001) and Supplementary Table 2).

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 target stacking strength (reviewed in (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

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

is an obligatory step in programmed V(D)J recombination, where the Artemis/DNA-PKcs 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 (reviewed in (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

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Materials and 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 9 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). *A) Standard assay conditions*. Oligonucleotides (0.6 – 0.8 OD_{260}/ml, 1 – 3 μM) were dissolved in buffer 1 (50 mM KCl, 0.5 mM $MgCl_2$, 0.4 mM Na-phosphate pH 7.0) and equilibrated overnight at 25° C. The optical absorbance 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° to 94° C and cooling (annealing curve) from 94° 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). *B) Other pH assay conditions.*

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

For the $d(CCCT)_9$ and $d(CCT)_{12}$ 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 $MgCl_2$; 0.5 mM Tris-acetate (pH 7.0), 50 mM KCl, 0.5 mM $MgCl_2$; 0.5 mM Tris-HCl (pH 8.0), 50 mM KCl, 0.5 mM $MgCl_2$ and both the T_m and T_a values were determined. *C) Variable salt concentration assay conditions* were used to distinguish between non-helical 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, pH 7.4) with or without KCl (1, 5, 10, 50, 100 and 500 mM), NaCl (1, 10, 50, 100 mM), LiCl (1, 10, 50 mM), $MgCl_2$ (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 $d(AAAT)_9$, $d(AAAG)_9$, $d(AAAC)_9$, $d(AAC)_{12}$, $d(AAT)_{12}$ and $d(AAG)_{12}$ oligonucleotides, which displayed near-linear TDAS hypochromicity curves, the slopes of the cooling curves were taken over the entire temperature range (10 – 94° C). For the $d(AAGG)_9$ and $d(AGAT)_9$ 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 $d(AAGG)_9$ oligonucleotide, two near-linear

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

segments were observed: one less steep from 94° C to ~60° C, and a second, steeper, 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 $d(XGGG)_9$ 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_{260}/ml, 1 – 3 μM) at 25° C by employing the *variable salt concentration assay conditions* (see above) with and without K^+, Na^+, Li^+ (1 – 100 mM) and Mg^{2+} (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, *etc*. (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: *A)* 9 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); *B)* 4 non-vertebrate genomes: plants [*Arabidopsis thaliana*, tair7 and rice (*Oryza*

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Acknowledgements

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 number N01-CO-12400), and financial support from BIOBASE GmbH to D.N.C. We express appreciation to 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 Dr. Jin Jen of the NCI for helpful discussions.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

[1] Abbreviations

Concerning nucleic acid nomenclature, we designated a double-stranded genomic triNR or tetraNR by its unique sequence (Tables 1 and 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. By contrast, we specify single-stranded DNA oligonucleotides and their reading frame by $d(AGC)_n$, for example. A subscript (n) indicates the number of repeating units. Hydrogen bonded nucleotides are also indicated by a colon, i.e. A:T.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure Legends

Fig. 1. Correlation between T_a and tetraNR abundance in 9 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 ≥8 units in the 9 vertebrate genomes listed in Materials and Methods. The regression (*solid line*) was calculated only for those tetraNR sequences (*solid circles*) for which either the forward or reverse oligonucleotides displayed a T_a value within the measurable temperature range (10° to 94° C). *Vertical lines*, Standard errors; b*lue*, 99% confidence interval; *open circles*, tetraNRs devoid of temperature-dependent structural transitions; CGGG, *filled square* and TGGG, *filled triangle* are tetraNRs with T_a values >94° C; *open square*, CCGG, with a T_a value >94° C but not present in any genome.

Fig. 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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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.

Applequist, 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., 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.

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 CGG.CCG sequence. *J. Biol. Chem.* **276:** 18605-18613.

Bacolla, A. and Wells, R.D. 2004. Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.* **279:** 47411-47414.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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.

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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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, 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.

Friedman, R.A. and Honig, B. 1995. A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophys. J.* **69:** 1528-1535.

Hilser, V.J. and Thompson, E.B. 2007. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **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

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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.*

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 instabilitites and neurological diseases* (eds. R.D. Wells and T. Ashizawa), pp. 277-296. Elsevier/Academic Press, San Diego, CA.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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. U. S. A.* **99:** 5596-5600.

Powell, J.T., Richards, E.G., and Gratzer, 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, J., Jr. 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.

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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, 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.

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.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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.*

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 AAAG 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 x CTG trinucleotide repeat instability in Escherichia coli. *Genetics* **176:** 27-41.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

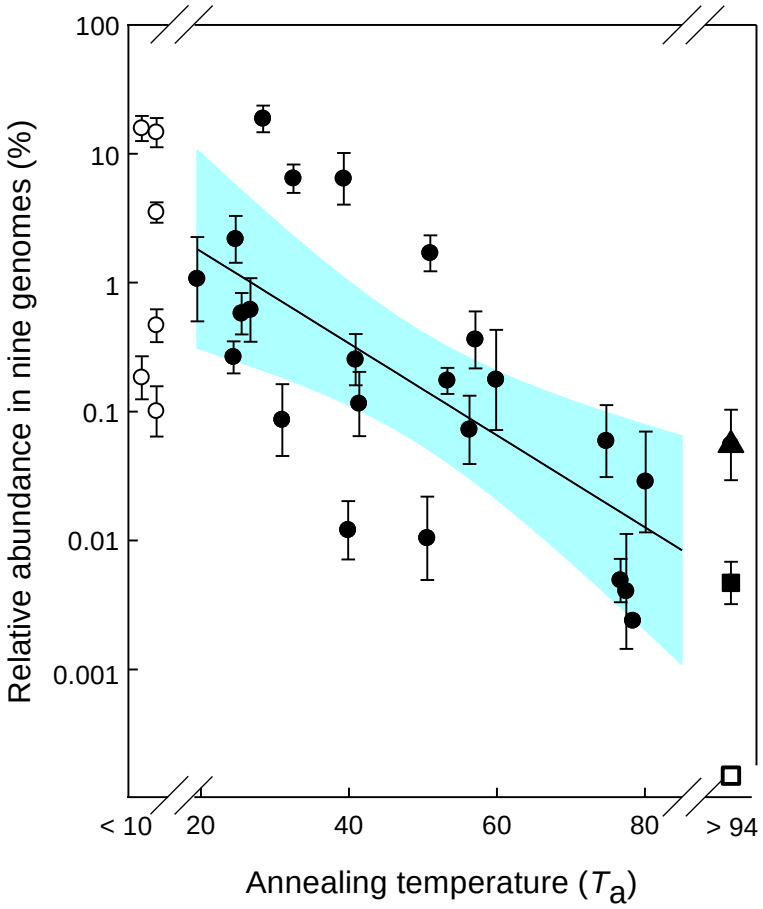

Fig. 1

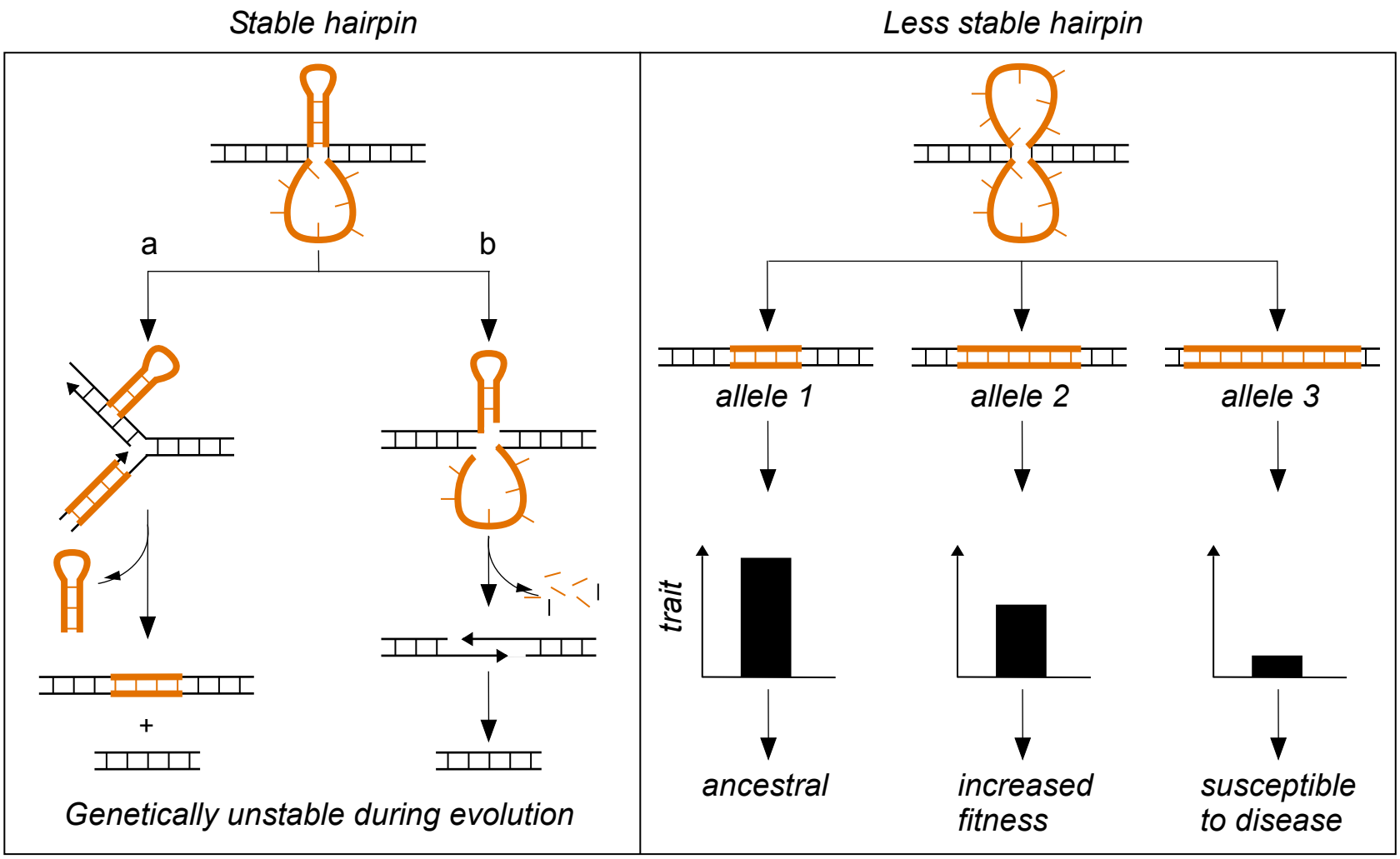
Stable hairpin
Less stable hairpin
a
b
+
Genetically unstable during evolution
allele 1
allele 2
allele 3
trait
ancestral
increased fitness
susceptible to disease

Fig. 2

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Number 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	$d(ACCG)_9$	39.4	H	$d(CGGT)_9$	50.6	H
0	ACGT	$d(ACGT)_9$	78.4	H	(*a*)		
0	AGCG	$d(AGCG)_9$	86.7	H	$d(CGCT)_9$	21.8	H
0	ATCG	$d(ATCG)_9$	77.5	H	$d(CGAT)_9$ (*b*)	76.4	H
0	ATGC	$d(ATGC)_9$	80.1	H	$d(GCAT)_9$ (*b*)	78.6	H
0	CCCG	$d(CCCG)_9$	35.8	H	$d(CGGG)_9$ (*c*)	N	Q
0	CCGG	$d(CCGG)_9$ (*c*)	N	H	(*a*)		
1	AACG	$d(AACG)_9$	39.9	H	$d(CGTT)_9$	29.0	H
2	AGCT	$d(AGCT)_9$	76.7	H	(*a*)		
3	ACGC	$d(ACGC)_9$	46.3	H	$d(GCGT)_9$	59.9	H
3	ACTG	$d(ACTG)_9$	31.0	H	$d(CAGT)_9$	N	C
4	ACCC	$d(ACCC)_9$	24.7	H	$d(GGGT)_9$ (*c*)	N	Q
6	AAGT	$d(AAGT)_9$	S	C	$d(ACTT)_9$	N	C
8	ACGG	$d(ACGG)_9$	74.8	H	$d(CCGT)_9$	22.6	H
9	AGCC	$d(AGCC)_9$	39.2	H	$d(GGCT)_9$	56.3	H
10	AACT	$d(AACT)_9$	N	C	$d(AGTT)_9$	S	C
24	AATT	$d(AATT)_9$	53.3	H	(*a*)		
34	ACTC	$d(ACTC)_9$	N	C	$d(GAGT)_9$	40.9	H
53	ACCT	$d(ACCT)_9$	N	C	$d(AGGT)_9$	24.4	H
54	AAGC	$d(AAGC)_9$	37.9	H	$d(GCTT)_9$	41.4	H
71	AATC	$d(AATC)_9$	S	C	$d(GATT)_9$	25.5	H
84	AGGC	$d(AGGC)_9$	57.1	H	$d(GCCT)_9$	N	C
86	AACC	$d(AACC)_9$	N	C	$d(GGTT)_9$	N	C
138	ACAG	$d(ACAG)_9$	19.5	H	$d(CTGT)_9$	16.7	H
539	AGGG	$d(AGGG)_9$ (*c*)	N	Q (*d*)	$d(CCCT)_9$	26.7	H
558	ACAT	$d(ACAT)_9$	N	C	$d(ATGT)_9$	24.7	H
688	AATG	$d(AATG)_9$	51.0	H	$d(CATT)_9$	N	C
1092	AAAC	$d(AAAC)_9$	S	C	$d(GTTT)_9$	N	C
1718	ATCC	$d(ATCC)_9$	N	C	$d(GGAT)_9$	21.0	H
4431	AAGG	$d(AAGG)_9$	39.3	H	$d(CCTT)_9$	N	H
5134	AGAT	$d(AGAT)_9$	28.5	H	$d(ATCT)_9$	N	C
6310	AAAG	$d(AAAG)_9$	S	C	$d(CTTT)_9$	N	C
8530	AAAT	$d(AAAT)_9$	S	C	$d(ATTT)_9$	N	C

Column 1, number of tetraNRs ≥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*, as columns 3 – 5 but for the reverse oligonucleotide sequences; (*a*), self complementary sequence with no staggered ends for which T_a measurements for the forward and reverse oligonucleotide sequences are identical; (*b*), self complementary sequence with staggered ends for which T_a measurements for the forward and reverse oligonucleotide sequences are not identical; (*c*), sequences studied by T_a analyses with *variable salt concentration assay conditions* and/or by CD; (*d*) structure in the presence of 100 mM K^+ and 10 mM Mg^{2+}, other structures are possible at different metal ion concentrations (Supplementary Information); *N*, <8% total hypochromicity during both the heating and cooling steps without a detectable first-derivative peak; *S*, >8% total hypochromicity during either the heating or cooling steps without a detectable first-derivative peak; *C*, single-stranded coil; *H*, hairpin, *Q*, quadruplex.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Table 2. TriNR abundances and T_a values

Unique genomic sequence	Number of tracts in human genome	% of all triNRs in vertebrate genomes	Forward oligo (FO)	T_a (FO)	Reverse oligo (RO)	T_a (RO)
ACG	3	0.1	$d(ACG)_{12}$	58.5	$d(CGT)_{12}$	50.6
ACT	89	3.4	$d(ACT)_{12}$	N	$d(AGT)_{12}$	27.1
CCG	118	1.1	$d(CCG)_{12}$	48.4	$d(CGG)_{12}$	76.7
AGC	152	4.5	$d(AGC)_{12}$	58.3	$d(GCT)_{12}$	59.9
ACC	188	3.4	$d(ACC)_{12}$	N	$d(GGT)_{12}$	N
AGG	229	8.5	$d(AGG)_{12}$	S	$d(CCT)_{12}$	19.2
ATC	458	6.0	$d(ATC)_{12}$	6.5	$d(GAT)_{12}$	23.0
AAG	734	8.8	$d(AAG)_{12}$	S	$d(CTT)_{12}$	N
AAC	1075	10.7	$d(AAC)_{12}$	N	$d(GTT)_{12}$	N
AAT	3668	46.4	$d(AAT)_{12}$	S	$d(ATT)_{12}$	14.2

Column 1, unique triNR sequence; *column 2*, number of triNRs ≥10 units in the human genome. *column 3*, the % of each triNR relative to all triNRs for 8 vertebrate genomes (human, chimp, mouse, rat, dog, chicken, fugu and zebrafish) was averaged. In the cow genome assembly browser the AGC tracts accounted for ~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* and *S*, as per legend to Table 1.

Downloaded from genome.cshlp.org on May 19, 2026 . Published by Cold Spring Harbor Laboratory Press

Table 3. Intrinsic base stacking and repeat tract length

Unique genomic sequence	ten longest tracts (*mean bp*)	single-stranded oligo	TDAS slope ($x10^4$)	bases X–Y $\Delta G(\nu)$	bps X–Y (*vacuum*) ΔE	bps X–Y (*water*) ΔE	bps X–Y $\Delta H°$
				kcal/mol (average value for X-Y base or bp steps)			
AAAG	221	$d(AAAG)_9$	20.4	-7.78	-13.70	-10.75	-7.80
AAGG	124	$d(AAGG)_9$	16.6	-8.03	-12.82	-10.22	-7.90
ATCC	95	$d(GGAT)_9$	12.6	-7.25	-13.12	-10.17	-7.97
AGAT	68	$d(AGAT)_9$	11.6	-7.08	-12.87	-10.27	-7.60
AAAT	61	$d(AAAT)_9$	11.2	-6.81	-13.87	-11.02	-7.40
AAAC	46	$d(AAAC)_9$	10.5	-6.63	-14.47	-10.60	-8.02
r^2			*0.96*	*0.62*	*0.03*	*0.004*	*0.02*
AAG	141	$d(AAG)_{12}$	17.8	-7.86	-13.37	-10.50	-7.87
AAT	57	$d(AAT)_{12}$	9.7	-6.58	-13.60	-10.87	-7.33
AAC	52	$d(AAC)_{12}$	7.2	-6.34	-14.40	-10.30	-8.17

Column 1, unique genomic sequence, as per Tables 1 and 2; *column 2*, average length in bp 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 (ε_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 [$\Delta E^{AB//CD}$ (ΔE^4)] of (Sponer et al. 2006)); *column 7*, as column 6 but in the hydrated state (data from Table 6 [$\Delta\Delta G_{sol}$ (water) with B3LYP/6-31G(d) data] of (Sponer et al. 2006) were added to the Table 4 $\Delta E^{AB//CD}$ (ΔE^4) 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 “ten longest tracts” column data as the *x* independent variable.