Supplementary Information

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

Supplementary Text

Molecular modeling. Each of the 62 single-strands (both forward and reverse sequences) comprising the 33 unique genomic tetraNR sequences (Table 1) were modeled to evaluate their capacity to fold back into quasi-stable hairpin structures. Four slipped-frame arrangements were possible for each folded-back sequence. Hydrogen-bonding arrangements with stable base-pairs (*i.e.* C:G, A:T, G:A, G:G, A:A, G:T and G:A) that could form within either a Watson-Crick- or Hoogsteen-type (Saenger 1994) duplex DNA were recorded. From the modeling, it was evident that tetraNRs with self-complementary sequences which yielded hairpin stems with all Watson-Crick base-pairs, were rare or absent from the human genome. Alternatively, tetraNRs in which each strand was modeled into hairpins containing a doublet of CG:CG or GC:GC Watson-Crick base-pairs that alternated with non-Watson-Crick base-pairs, were of intermediate abundance, whilst tetraNRs with folded-back structures devoid of Watson-Crick base-pairs were among the most abundant. Hence, DNA structure modeling laid the foundation for the hypothesis that hairpin formation played a key role in determining tetraNR abundance in the human genome.

Complex behavior of d(CCCT)$_9$ and d(CCT)$_{12}$. As stated in the Results, d(CCCT)$_9$ and d(CCT)$_{12}$ displayed strong hysteresis effects under the *standard assay conditions* (Materials and Methods), with T_m values >20° C higher than the T_a values. Since this behavior suggested the slow formation of hairpins following limited cytosine protonation during the prolonged incubation at 25° C prior to the temperature-dependent absorption spectroscopy (TDAS) melting step, pH-dependent determinations were conducted. For the d(CCCT)$_9$ oligonucleotide, the biphasic melting curve at pH 4.5 contained two distinct transitions with T_m values of 76.3 and 83.3° C, respectively. The subsequent annealing curve showed only one transition with a T_a value of 74.8° C. For both curves, the hypochromicity ranged from ~0.2 to 0.8 from 4° to 95° C. At pH 7.0, the melting curve still displayed two transitions but with the T_m values reduced to 41.1 and 59.9° C, respectively. The annealing curve also showed a much lowered T_a at 26.7° C. At pH 8.0, no TDAS changes were observed during either the melting or annealing steps.

For the d(CCT)$_{12}$ oligonucleotide, both the melting and annealing curves at pH 4.5 displayed similar transition values at 69.2 and 67.0° C, respectively. However, at pH 7.0, the melting curve showed a T_m value of 45.7° C whereas the annealing curve showed a T_a at 19.2° C. At pH 8.0, no TDAS changes were observed during either the melting or annealing steps. Hence, the low pH increased the temperature for the midpoint of transitions and, most importantly, abrogated the hysteresis effects. We conclude that *a*) decreasing the pH from 8.0 to 4.5 induced strong hairpin formation and *b*) the kinetics of hairpin formation was slow at neutral pH but increased sharply with increasing proton concentration. Since cytosine protonation has a pK_a of ~4.5 for the isolated monomer (Saenger 1994) but has a much higher pK_a (around neutrality) for the polymeric form (Inman 1964; Jaishree and Wang 1993; Wells and Larson 1972), these data are consistent with hairpin formation at low pH being stabilized by C:C$^+$ pairs and, possibly, hairpin interactions into quadruplex structures containing i-motifs (Gueron and Leroy 2000).

Sequences with exceptionally stable or multiple DNA helices. The self-complementary d(CCGG)$_9$ and the 3 d(XGGG)$_9$ oligonucleotides (where X = A, C or T) were predicted to form very stable hairpins and quadruplex structures, respectively. However, no first-derivative peaks were detectable for these 4 oligonucleotides from 10 – 94° C. Since monovalent cations increase helix stability (Hillen et al. 1981), temperature-dependent transitions were performed in the absence of cations (buffer 2, Materials and Methods) and with or without formamide.

With d(CCGG)$_9$, no changes in absorbance were observed in the absence of cations (buffer 2). Hence, further determinations were performed in the presence of formamide. Formamide competes with hydrogen bond donor and acceptor groups, thereby destabilizing duplex DNA; 50% formamide was expected to decrease the T_m by ~30° C (Hutton 1977). A well defined cooperative transition was observed under these conditions, with T_m and T_a values of ~74.0° C.

The d(CGGG)$_9$ oligonucleotide displayed a T_a value of 54.4° C in buffer 2, indicative of a coil-to-helix transition. Addition of K^+ ions (1 – 100 mM) reduced the extent of hypochromicity from 10% to <3%, implying that the K^+ ions induced a highly thermostable structure. Given the prominent role of K^+ ions in quadruplex formation (Burge et al. 2006), CD was used to probe these DNA structures (Vorlickova et al. 2005; Xu and Sugiyama 2006). CD spectra revealed a moderate maximum at ~264 nm and a minimum at ~242 nm in the absence of cations (Supplementary Fig. 1), a finding which was to be expected from the stacked guanines in the d(CGGG)$_9$ folded conformation (Kypr and Vorlickova 2002). Addition of K^+ ions (1 – 100 mM) caused a

~3-fold enhancement of the CD spectral signatures, indicative of quadruplex formation (Hardin et al. 2001; Jin et al. 1992; Krishnan-Ghosh et al. 2004; Rachwal et al. 2007b). The 264 nm and 242 nm CD signatures were also enhanced by Na^+ ions (1 – 100 mM), although to a lesser extent, whereas they remained unchanged by the addition of similar concentrations of Li^+ (Supplementary Fig. 1, inset). Since quadruplex stabilization by monovalent cations follows the order $K^+>Na^+>Li^+$ (Hardin et al. 2001), these composite data are consistent with the conclusion that $d(CGGG)_9$ formed a quadruplex structure (Hardin et al. 1992; Kettani et al. 1998) under the *standard assay conditions*, with a T_a >94° C.

The $d(GGGT)_9$ oligonucleotide displayed no temperature-dependent changes in absorbance in buffer 2. However, addition of 50% formamide yielded a reversible coil-to-helix transition with a T_a of 74.0° C, indicating that $d(GGGT)_9$ formed a stable hydrogen-bonded structure. CD spectra in the absence of cations showed a prominent maximum (~30 mdeg) at ~264 nm and a minimum (~-10 mdeg) at ~242 nm, which were increased only moderately (~10-20%) upon addition of either K^+ or Na^+ (not shown). Hence, we conclude that $d(GGGT)_9$ formed stable quadruplex structures under all assay conditions (Krishnan-Ghosh et al. 2004).

The $d(AGGG)_9$ oligonucleotide displayed complex behavior. In the absence of cations, a cooperative coil-to-helix transition was observed with a T_a value of 49.5° C, preceded by linear decreases in absorbance at higher temperatures. Strong hysteresis was also observed, such that the melting (helix-to-coil) curve was hypochromic with respect to the annealing (coil-to-helix) curve in the linear range above the midpoint of transitions. We conclude that $d(AGGG)_9$ folded into a hairpin structure, and that in the single-stranded d(AGGG)$_9$ all bases stacked with slow kinetics to form helical coils. Addition of ≥10 mM K^+ ions abrogated the cooperative transitions but did not alter the steady change in absorbance which occurred *a*) over the entire temperature range and *b*) only during the melting step. Hence, K^+ ions both impeded hairpin formation and reduced the kinetics of base stacking. CD spectra indicated a strong maximum at ~264 nm and a minimum at ~242 nm in the absence of cations; these amplitudes increased by ~20% in the presence of ≥10 mM K^+. These CD data were consistent with the presence of stacked guanines, both in the d(AGGG)$_9$ hairpin in the absence of K^+, as well as in the single-stranded coils in the presence of K^+ (Kypr and Vorlickova 2002).

The most likely hydrogen bonds in the d(AGGG)$_9$ hairpin occurred through Hoogsteen interactions between G:G and A:A bases (Huertas and Azorin 1996; Saenger 1994). This arrangement affords a fully base-paired hairpin stem. Hoogsteen G:G and A:A interactions are typically observed in triplex DNA where they are stabilized by Mg^{2+} ions (Liquier et al. 2001). Addition of up to 100 mM Mg^{2+} ions increased the T_a value of d(AGGG)$_9$ to 54.7° C and also minimized hysteresis, indicating that the divalent cation stabilized the hairpin structure and accelerated the kinetics of base stacking. Surprisingly, the combined presence of 10 mM Mg^{2+} and ≥10 mM K^+ ions did not yield any temperature-dependent changes in absorbance, either during the melting or annealing steps. In addition, K^+ ions did not increase the CD amplitudes of the ~264 and ~242 nm peaks, which were observed in the presence of Mg^{2+} alone, strongly supporting the formation of a quadruplex structure, perhaps from the association of two d(AGGG)$_9$ hairpins. In summary, the d(AGGG)$_9$ oligonucleotide formed a quadruplex structure at K^+ and Mg^{2+} concentrations of 100 mM and 10 mM, respectively, a helical (possibly a hairpin) structure in the presence of 10 mM Mg^{2+} and <10 mM K^+ ions, and a single-stranded coil in the presence of ≤0.5 mM Mg^{2+} ions and ≥10 mM K^+ ions.

The numbers of CCGG, CGGG, TGGG and AGGG tetraNRs in the human genome were 0, 0, 4 and 539, respectively (Table 1). A comparison (Rachwal et al. 2007a) of the relative stabilities adopted by XGGG quadruplexes indicated that the $d(AAAG)_4$ sequence forms much less stable structures than the $d(CGGG)_4$ and $d(TGGG)_4$ sequences in the presence of K^+ ions alone. Taken together these biophysical investigations support the existence of an inverse relationship between the capacity of single-stranded tetraNRs to form DNA hairpin and quadruplex secondary structures and their abundance in the human genome.

Note regarding the "T_a and tetraNR abundance are inversely correlated" section. We choose to plot the highest T_a value for each pair of tetraNR sequences (Table 1) for the following reason. When double-stranded chromosomal DNA separates into single-strands during replication or transcription, a hairpin may form at a certain location on one strand but not on its complement due to their inherently different thermodynamic stabilities. If bypass of the hairpin were then to occur during DNA replication, this would result in the loss of the repeat sequence in some of the progeny molecules (Iyer et al. 2000; Zahra et al. 2007). Alternatively, if the hairpin were to induce double-strand breaks, subsequent repair could remove the repeat sequence, possibly along with flanking sequences (Wojciechowska et al. 2006). Thus, this duplex region would tend to be lost from the population over time. Moreover, this loss would depend on the most stable hairpin that would form on either one or the other strand of a given duplex DNA sequence.

***T_a* vs. tetraNR abundance correlation with projected numbers of CpG-containing tetraNRs**. To determine whether the correlation between T_a and tetraNR abundance was due solely to the very low number of CpG-containing repeats, we reassessed this correlation after projecting the expected number of CpG-containing tetraNRs as if they had not been lost through methylation-mediated C:G → T:A transitions, as follows.

First, we compared the number of CpG-containing tetraNRs of length 3 units with the number of GpC-containing tetraNRs also of length 3 units. The average number of GpC-containing tetraNRs (ATGC, AAGC, AGCC, AGCT and AGGC) was 4004±1639 (mean ± SE). The numbers of CpG-containing tetraNRs varied from 22 to 1504. Three groups were then formed. Group A contained the 3 tetraNR sequences with 50% C+G content, *i.e*. AACG, ACGT and ATCG. Group B contained the 4 sequences with 75% C+G content, *i.e*. ACCG, ACGG, AGCG and ACGC, whereas group C contained the 2 sequences with 100% C+G, *i.e*. CCCG and CCGG. The average numbers of repeats were: 45±3 (group A), 301±141 (group B) and 920±584 (group C), corresponding to 88-, 13- and 4-fold reductions, respectively, relative to the 4004 value for the GpC-containing repeats. This increase in mean number of repeats with increasing C+G content is consistent with expectation since the kinetics of CpG methylation (Bacolla et al. 2001), spontaneous deamination (Frederico et al. 1993; Lindahl and Nyberg 1974) and C:G → T:A transition rates (Elango et al. 2008) all decrease with increasing DNA stability. The CCCG and CCGG repeats were not considered further since their relative T_a values exceeded 94° C (Table 1). Hence, the "observed" numbers of CpG-containing tetraNRs were each multiplied by 88 (group A) and 13 (group B) and named the "corrected" numbers of CpG-containing tetraNRs.

Second, we projected the total numbers of tracts ≥8 units for each CpG-containing tetraNR. This was performed by reference to the rate of tetraNR loss as a function of length, determined for the GpC-containing tetraNRs. A plot of tract length (number of repeat units x, x = 3 – 8) as a function of the natural log of the number of tracts (Y_L) was described (r^2>0.98) by a 3-parameter exponential decay, $Y_L = Y_{0L} + A*e(exp)-<B>*x$. The average slope ($<B>$) value for the GpC-containing tetraNRs was 0.363±0.159 (SD). The projected numbers of CpG-containing tetraNRs with ≥8 units was therefore obtained from the relationship $lnY_{Li} = lnY_{L3} + <B>*(x_3-x_i)$, where Y_{L3} was the corrected number of each CpG-containing tetraNR with 3 units, x_3 = 3 and x_i the projected length in repeat units. The values obtained in the range of i = 8 – 15 were then added to give the total number of tracts ≥8 units for each CpG-containing tetraNRs. A second calculation was also performed by using the smallest observed B value (B_{min} = 0.199) as to give an upper limit for the projected numbers of tetraNR tracts. The total numbers of projected tetraNR tracts ranged from 12-16 using $<B>$ and 33-73 using B_{min}. A plot of T_a *vs*. numbers of tracts was then performed by using the projected numbers instead of the observed numbers of CpG-containing tetraNRs. Using $<B>$, the correlation improved somewhat as compared to the uncorrected correlation (P = 0.0068 *vs*. 0.0283), whereas it deteriorated to a marginally non-significant level (P = 0.0552) using B_{min}. In conclusion, the loss of CpG-containing tetraNRs through C:G → T:A transitions consequent to methylation-mediated deamination does not appear to be the underlying cause for the observed correlation between T_a and the abundance of tetraNR tracts in the human genome.

Note regarding the "T_a determinations in triNRs" section. The $d(CCT)_{12}$ oligonucleotide displayed hysteresis effects and pH-dependent T_a values due to cytosine protonation (see above) and therefore only the T_a value of 19.2° C at pH 7.0 was considered. Moreover, no sequences were found to adopt quadruplex structures under *standard assay conditions* (Materials and Methods).

ΔG_v values in control A-rich sequences. To determine whether or not the correlations between ΔG_v values and TDAS slopes were due solely to the A-rich nature of the tetraNR and triNR sequences, the average ΔG_v value was calculated for three 20-mer single-stranded DNA sequences containing 13/20, 14/20 and 15/20 A-residues within non-repetitive DNA. The ΔG_v values were: for sequence 1 (TAACAAAGACAATAAAATAC) -6.72 kcal/mol, for sequence 2 (AAGAACACAAATAATAAACA) -6.72 kcal/mol and for sequence 3 (ACAAACAATAAAAACAACAA) -6.62 kcal/mol. These values are less negative than those calculated for the AAAG, AAGG and AAG repeats, thereby supporting the conclusion that the correlations between the ΔG_v values and TDAS slopes were not due simply to the A-rich nature of the tetraNR and triNR sequences.

Disclaimer. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Supplementary Figure Legends and Notes

Supplementary Fig. 1. CD spectra of the d(CGGG)$_9$ oligonucleotide. A total of 0.6 OD$_{260}$/ml (2 μM) of d(CGGG)$_9$ was dissolved in buffer 2 (Materials and Methods) with or without increasing concentrations of KCl (*left inset*) to determine whether the metal ion induced quadruplex formation. The solutions were equilibrated overnight at 25° C and CD analyses were performed at 25° C. Each trace represents the average of 10 determinations. *Right Inset*, the d(CGGG)$_9$ oligonucleotide was equilibrated overnight at 25° C as before in the presence of KCl (1 – 100 mM), NaCl (1 – 100 mM) or LiCl (1 – 100 mM) and the CD spectra were recorded. The mdeg values obtained at 264 nm were then replotted as a function of increasing monovalent cation concentrations.

Supplementary Fig. 2. T_a *vs*. tetraNR abundance as a function of tract length. Six different regression coefficients (r^2) were determined. Each r^2 value was obtained from the analysis described in Fig. 1. However, the numbers of genomic tetraNRs analyzed in the 9 combined vertebrate genomes varied in length so that tracts containing ≥8, ≥7, ≥6, ≥5, ≥4 and ≥3 units, each yielded a separate r^2 value. These six r^2 values (*y*-axis) were plotted against the corresponding genomic tetraNR unit lengths (TUL) analyzed (*x*-axis) and the data were fitted by a rectangular hyperbola of the form [$(ax)/(b+x) + cx$].

Supplementary Fig. 3. T_a values as a function of d(ACTG)$_n$ oligonucleotide repeat number, *n*. Four self-complementary oligonucleotides comprising *n* = 4, 6, 9 and 12 $d(ACTG)_n$ tetraNRs were used to determine the length-dependent changes in T_a. For each length, 0.8 OD_{260} of oligonucleotide was incubated overnight at 25° C in 1 ml of buffer 1 and used to determine the T_a values. Given the $d(ACTG)_n$ self-complementarity, the oligonucleotides may form regular duplexes. Hence, the T_a values reflect the stability of the stem (not allowing for the loop) of folded-back oligonucleotides with twice as many tetraNR units.

Supplementary Fig. 4. Distribution of human triNRs in genes and intergenic regions. The triNR sequences were categorized according to their location either in intergenic regions or within RefSeq annotated genes (introns and exons, cDNAs) and expressed as percentage values. When a triNR tract overlapped an exon/intron boundary, it was classified as residing within the coding region. Similarly, when a triNR tract was located both within an intron and an exon as a result of alternative splicing, it was classified as being located within the coding region. A cut-off value of 10% was taken as a measure of significant localization within coding regions. Hence, only triNR sequences with a total number of tracts >10 were considered. The numbers of tracts for each sequence are indicated above the bars. The vertical dashed line separates the sequences with >10 tracts from those with ≤10 tracts. *Panel A*, triNR sequences with ≥10 units; *panel B*, triNR sequences with ≥4 units.

Supplementary Fig. 5. Distribution of human tetraNRs in genes and intergenic regions. All details are the same as in Supplementary Fig. 4 but for tetraNRs. *Panel*

A, tetraNR sequences with ≥8 units; *panel B*, tetraNR sequences with ≥3 units.

Supplementary Fig. 6. Micro/minisatellites in human cDNAs and their encoded amino acids. *Panel A*. Human RefSeq cDNA genes were analyzed for their content of micro/minisatellites, *i.e*. any tandem repeat sequence composed of *a*) ≥2 repeat units, *b*) a total of ≥12 nt in length (examples include $(CA)_6$, $(CTG)_4$, $(CCAT)_3$, $(GGATC)_3$, $(TTCTAC)_2$ etc.) and *c*) 2 – 11 nt per repeat unit. When a redundant genomic location was found in different cDNAs, it was counted only once. *Panel B*, the diNRs and triNRs from panel A were classified according to their location in cDNAs at the start of transcription (START), in the 5' untranslated region (5'-UTR), the open reading frame (ORF), and the 3' untranslated region (3'-UTR). Repeats at START coincided with heterogeneous transcription start sites found in several expressed sequence tags (ESTs). For the classification, the location of all diNRs and triNRs was verified manually on the UCSC Human Genome Browser. This analysis was performed on the hg17 Genome Assembly of 2004. *Panel C*, the genomic coordinates corresponding to micro/minisatellites located in the ORF regions from Panel B were used to retrieve the amino acids encoded by the repeats, which were then computed. All homopolymeric amino acid runs encoded by triNRs and hexaNR were recorded (the hexaNRs only made a minor contribution). For heteropolymeric amino acid runs, only diamino acid runs (ER, CV and TH) encoded by diNRs were found to be present in significant number. No aromatic F, Y and W (boxed) homopolymeric amino acid runs were found. The amino acids were further classified according to which translated exon (first, internal, last, and single) encoded them, and the percentage values were recorded.

Internally translated exons encoded preferentially polyQ, polyE and polyS runs, 58, 57 and 54%, respectively. By contrast, the first translated exon encoded preferentially polyA, polyL, polyG, polyP and polyH. The inset shows the results for polyA, polyL, polyG and polyP. *Panel D*, cartoon showing the side chain similarities between Q and E.

Supplementary Fig. 7. Evolutionary conservation of three human coding triNR tracts. *Panel A*, alignment of the protein sequence flanking an evolutionarily conserved polyQ amino acid run (*boxed*) in the *TBP* gene encoding the TATA-box binding protein (*upper*), followed by a proposed alignment of the conserved CAG/CAA tract (*below*) encoding the boxed polyQ run. The database sources for the protein sequences are indicated. This analysis was conducted on all available genomes at the UCSC Genome Browser (http://genome.ucsc.edu). The TATA-box binding protein is an essential component of the transcription factor IID (TFIID), a multi-peptide complex which recruits other basal transcription factors to initiate RNA polymerase II-dependent transcription of mRNAs and other small nuclear RNAs (Grob et al. 2006; Liu et al. 2007; Patikoglou et al. 1999). Expansion of the conserved CAG repeat in humans gives rise to spinocerebellar ataxia type 17 (SCA17) (Nakamura et al. 2001), a form of autosomal dominant cerebellar ataxia involving mainly the cerebral cortex, striatum, and cerebellum (reviewed in (Orr and Zoghbi 2007; Stevanin and Brice 2006)). The Figure illustrates the increase in length of the conserved polyQ run in those species with the most recent evolutionary origin, particularly human. This increase in length has mostly been caused by an expansion of clustered CAG repeats; by contrast, clustered CAA repeats have remained rather constant in number (1-3) over evolutionary time. In addition, whereas the amino acid sequence preceding the conserved polyQ run has also been highly conserved, the amino acid composition immediately after the conserved polyQ run has undergone substantial remodeling, particularly in those species with the most recent evolutionary origin, in which it has acquired an increasing number of small (A and V) and nucleophilic (S and T) amino acid residues. These are then followed by highly conserved residues.

Panel B, alignment of the protein sequence flanking an evolutionarily conserved polyQ amino acid run (*boxed*) in the *MEF2A* gene encoding the MADS box transcription enhancer factor 2, polypeptide A (myocyte enhancer factor 2A). This analysis was performed for the 9 genomes used in the present study. The MEF2A transcription factor binds specifically to A+T-rich promoter regions to activate transcription in muscle-specific genes and supports myogenic development (Huang et al. 2000; Kaushal et al. 1994; Santelli and Richmond 2000). The evolutionary comparison reveals a lengthening of the polyQ run due to the expansion of a CAG repeat in species with the most recent evolutionary origin, and increased complexity in the downstream amino acids, which are rich in Q and P residues. Remarkably, the CAG repeat has maintained the purity of its sequence composition in these genomes.

Panel C, alignment of the protein sequence flanking an evolutionarily conserved polyG amino acid run (*boxed*) in the *POU4F2* gene encoding POU domain class 4 transcription factor 2, followed by the codons encoding the boxed polyG run. POU4F2 plays a key role in retinal ganglion cell development and is essential for the establishment of the visual system (Mao et al. 2008; Mu et al. 2008). The evolutionary comparison of the *POU4F2* gene was complicated by *a*) a high degree of homology with the *POU4F1* gene of higher vertebrates and *b*) some uncertainty with respect to the evolutionary assembly of the extant *POU4F2* gene from sequences originally present in fish (fugu and zebrafish). In the chicken genome, two head-to-head DNA sequences shared homology with *POU4F1* and *POU4F2;* it remains unclear whether or not these sequences represent functional genes. This notwithstanding, a polyG run has formed upstream of a highly conserved protein-coding region; this is encoded by triplets manifesting an increasingly pure sequence composition $(GGC)_n$ in those species whose evolutionary origin has been the most recent. This polyG run was followed by another repetitive region encoding an evolutionarily stable polyS run and a less stable polyG run. *Boxed amino acids*, evolutionarily conserved amino acids encoded by a tract of at least 10 identical triNRs in the human reference genome sequence; *underlined*, putative tandem duplication events; *yellow highlight*, ≥4 GGC triNRs; l*ower case codons*, S and A amino acid interruptions.

(1) Poly(Q)-flanking region from contig11766:12,572-17,580.

(2) Alignment of NM_003194 with scaffold_200:1168122-1175114.

(3) The dinucleotide repeat interruption is not confirmed (Tomiuk et al. 2007).

(4) Alternative symbols: LOC309957, BC081907, NM_001014035, GeneID 309957.

(5) A region comprising two head-to-head segments is present in the chicken genome; these segments are orthologous to the human *POU4F1* and *POU4F2* genes.

The sequences of the triNRs were extracted from within the coordinates of following genomic regions. For the *TBP* gene: human hg18_dna range=chr6:170712710-170713273, chimpanzee panTro2_dna range=chr6:173817673-173817863, orangutan ponAbe2_dna range=chr6:174151433-174151619, rhesus, rheMac2_dna range=chr4:167567051-167567246, marmoset calJac1_dna range=Contig3625:32411-32622, mouse mm9_dna range=chr17:15641277-15641470, rat rn4_dna range=chr1:54394899-54395066, cat felCat3_dna range=scaffold_190627:97870-98147, dog canFam2_dna range=chr12:75482612-75482777, horse equCab1_dna range=chr31:301795-302042, cow bosTau4_dna range=chrUn.004.695:51682-51840, opossum monDom4_dna range=chr2:456789002-456789173, platypus ornAna1_dna range=Contig11766:13583-13767, chicken galGal3_dna range=chr3:42600469-42600630, lizard anoCar1_dna range=scaffold_200:1172491-1172684, Xenopus xenTro2_dna range=scaffold_2:6444646-6444815, zebrafish danRer5_dna range=chr13:24710556-24710700, Tetraodon tetNig1_dna range=chr5:3611822-3612063, Fugu fr2_dna range=chrUn:220346369-220346509, stickleback gasAcu1_dna range=chrII:13612758-13613012, medaka oryLat1_dna range=chr3:23352378-23352535. For the *MEF2A* gene: human hg18_dna range=chr15:98070204-98070356, chimpanzee panTro2_dna range=chr15:97711656-97711868, mouse mm9_dna range=chr7:74379877-74380130, rat DQ323505, dog canFam2_dna range=chr3:43872920-43873088, cow bosTau4_dna range=chr21:5609565-5609806, chicken galGal3_dna range=chr10:19145497-19145729, zebrafish danRer5_dna range=chr18:10777361-10777582, Fugu fr2_dna range=chrUn:260827696-260827918. For the *MEF2A* gene: human hg18_dna range=chr15:98070204-98070356, chimpanzee panTro2_dna range=chr15:97711656-97711868, mouse mm9_dna range=chr7:74379877-74380130, rat DQ323505, dog canFam2_dna range=chr3:43872920-43873088, cow bosTau4_dna range=chr21:5609565-5609806, chicken galGal3_dna range=chr10:19145497-19145729, zebrafish danRer5_dna range=chr18:10777361-10777582 and Fugu fr2_dna range=chrUn:260827696-260827918. For the *POU4F2* gene: human hg18_dna range=chr4:147779810-147780039, chimpanzee panTro2_dna range=chr4:150682649-150682913, orangutan ponAbe2_dna range=chr4:152217445-152217642, rhesus rheMac2_dna range=chr5:138768372-138768549, marmoset calJac1_dna range=Contig6718:72680-72845, mouse mm9_dna range=chr8:80959962-80960260, rat rn4_dna range=chr19:31389235-31389481, cat GeneScaffold_2960:480810:482352:1, dog canFam2_dna range=chr15:48044665-48044914, horse equCab1_dna range=chr2:71678521-71678703, cow bosTau4_dna range=chr17:12580455-12580732, opossum monDom4_dna range=chr5:132326737-132327090, platypus ornAna1_dna range=Ultra44:795734-795949.

Supplementary Fig. 8. Length-distributions of the longest tetraNR and triNR sequences in the human genome. *Panel A*, tetraNRs; *Panel B*, triNRs. The *x*-axis shows the tract lengths expressed in terms of the numbers of repeat units whereas the *y*-axis shows the total number of tracts. Only those sequences that displayed the longest length distributions are shown.

Supplementary Fig. 9. TetraNR distributions in nine vertebrate genomes. *Left*, schematic timescale for vertebrate evolution (adapted from (Matsuya et al. 2008)). *Right*, tetraNR distributions for the 10 longest repeat sequences in each species. Note that the types of sequences may not be identical in all species.

Supplementary Table Legends

Supplementary Table 1. Association of triNRs and tetraNRs with inherited human disease/traits. Representative list of studies reporting the association of polymorphic alleles with phenotypic traits and/or pathologic conditions and expanded triNRs or tetraNRs in neurological disorders.

Supplementary Table 2. Elevated microsatellite alterations at selected tetranucleotides (EMAST) in human cancer. List of studies reporting the % of microsatellite instability (MSI%) in human cancers. *Author's marker*, microsatellite designation given by the authors; *UCSC marker (hg18)*, microsatellite designation as given by the UCSC Human Genome Assembly hg18 browser at (http://genome.ucsc.edu/cgi-bin/hgGateway)

Supplementary Table 3. Enrichment analysis of human genes containing micro/minisatellitres tracts in cDNAs. All annotated human cDNAs satisfying the conditions described in the legend to Supplementary Fig. 6 were analyzed. The genes were further classified according to the micro/minisatellite location in the 5'UTR, ORF and 3'UTR. Four gene lists were drawn up, one for the composite number of genes and three for the classified locations. Each gene list was uploaded into the DAVID (http://david.abcc.ncifcrf.gov) database to conduct a gene enrichment analysis. This was performed by interrogating the four test gene lists against the Gene Ontology Biological Process (GOBP), Gene Ontology Cellular Compartment (GOCC), Gene

Ontology Molecular Function (GOMF), cell signaling pathways (KEGG Pathway) and the Swiss-Prot/Protein Informatics Resource (SP-PIR) databases. Only the most enriched terms are displayed.

Supplementary Table 4. Gene classes enriched in triNR- and tetraNR-containing genes associated with human genetic disease/phenotypic traits. The list of non-redundant genes from Supplementary Table 1 was used to conduct a gene enrichment analysis, as described in the legend to Supplementary Table 3.

Supplementary References

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.

Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K., and Neidle, S. 2006. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **34:** 5402-5415.

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.

Frederico, L.A., Kunkel, T.A., and Shaw, B.R. 1993. Cytosine deamination in mismatched base pairs. *Biochemistry* **32:** 6523-6530.

Grob, P., Cruse, M.J., Inouye, C., Peris, M., Penczek, P.A., Tjian, R., and Nogales, E. 2006. Cryo-electron microscopy studies of human TFIID: conformational breathing in the integration of gene regulatory cues. *Structure* **14:** 511-520.

Gueron, M. and Leroy, J.L. 2000. The i-motif in nucleic acids. *Curr. Opin. Struct. Biol.* **10:** 326-331.

Hardin, C.C., Perry, A.G., and White, K. 2001. Thermodynamic and kinetic characterization of the dissociation and assembly of quadruplex nucleic acids. *Biopolymers* **56:** 147-194.

Hardin, C.C., Watson, T., Corregan, M., and Bailey, C. 1992. Cation-dependent transition between the quadruplex and Watson-Crick hairpin forms of d(CGCG$_3$GCG). *Biochemistry* **31:** 833-841.

Hillen, W., Goodman, T.C., and Wells, R.D. 1981. Salt dependence and thermodynamic interpretation of the thermal denaturation of small DNA restriction fragments. *Nucleic Acids Res.* **9:** 415-436.

Huang, K., Louis, J.M., Donaldson, L., Lim, F.L., Sharrocks, A.D., and Clore, G.M. 2000. Solution structure of the MEF2A-DNA complex: structural basis for the modulation of DNA bending and specificity by MADS-box transcription factors. *EMBO J.* **19:** 2615-2628.

Huertas, D. and Azorin, F. 1996. Structural polymorphism of homopurine DNA sequences. d(GGA)n and d(GGGA)n repeats form intramolecular hairpins stabilized by different base-pairing interactions. *Biochemistry* **35:** 13125-13135.

Hutton, J.R. 1977. Renaturation kinetics and thermal stability of DNA in aqueous solutions of formamide and urea. *Nucleic Acids Res.* **4:** 3537-3555.

Inman, R.B. 1964. Transitions of DNA homopolymers. *J. Mol. Biol.* **9:** 624-637.

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.

Jaishree, T.N. and Wang, A.H. 1993. NMR studies of pH-dependent conformational polymorphism of alternating (C-T)n sequences. *Nucleic Acids Res.* **21:** 3839-3844.

Jin, R., Gaffney, B.L., Wang, C., Jones, R.A., and Breslauer, K.J. 1992. Thermodynamics and structure of a DNA tetraplex: a spectroscopic and calorimetric study of the tetramolecular complexes of d(TG_3T) and d($TG_3T_2G_3T$). *Proc. Natl. Acad. Sci. U. S. A.* **89:** 8832-8836.

Kaushal, S., Schneider, J.W., Nadal-Ginard, B., and Mahdavi, V. 1994. Activation of the myogenic lineage by MEF2A, a factor that induces and cooperates with MyoD. *Science* **266:** 1236-1240.

Kettani, A., Bouaziz, S., Gorin, A., Zhao, H., Jones, R.A., and Patel, D.J. 1998. Solution structure of a Na cation stabilized DNA quadruplex containing G.G.G.G and G.C.G.C tetrads formed by G-G-G-C repeats observed in adeno- associated viral DNA. *J. Mol. Biol.* **282:** 619-636.

Krishnan-Ghosh, Y., Liu, D., and Balasubramanian, S. 2004. Formation of an interlocked quadruplex dimer by d(GGGT). *J. Am. Chem. Soc.* **126:** 11009-11016.

Kypr, J. and Vorlickova, M. 2002. Circular dichroism spectroscopy reveals invariant conformation of guanine runs in DNA. *Biopolymers* **67:** 275-277.

Lindahl, T. and Nyberg, B. 1974. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13:** 3405-3410.

Liquier, J., Geinguenaud, F., Huynh-Dinh, T., Gouyette, C., Khomyakova, E., and Taillandier, E. 2001. Parallel and antiparallel G*G.C base triplets in pur*pur.pyr triple helices formed with (GA) third strands. *J. Biomol. Struct. Dyn.* **19:** 527-534.

Liu, Q.X., Nakashima-Kamimura, N., Ikeo, K., Hirose, S., and Gojobori, T. 2007. Compensatory change of interacting amino acids in the coevolution of transcriptional coactivator MBF1 and TATA-box-binding protein. *Mol. Biol. Evol.* **24:** 1458-1463.

Mao, C.A., Kiyama, T., Pan, P., Furuta, Y., Hadjantonakis, A.K., and Klein, W.H. 2008. Eomesodermin, a target gene of Pou4f2, is required for retinal ganglion cell and optic nerve development in the mouse. *Development* **135:** 271-280.

Matsuya, A., Sakate, R., Kawahara, Y., Koyanagi, K.O., Sato, Y., Fujii, Y., Yamasaki, C., Habara, T., Nakaoka, H., Todokoro, F. et al. 2008. Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.* **36:** D787-792.

Mu, X., Fu, X., Beremand, P.D., Thomas, T.L., and Klein, W.H. 2008. Gene regulation logic in retinal ganglion cell development: Isl1 defines a critical branch distinct from but overlapping with Pou4f2. *Proc. Natl. Acad. Sci. U. S. A.* **105:** 6942-6947.

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

Orr, H.T. and Zoghbi, H.Y. 2007. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* **30:** 575-621.

Patikoglou, G.A., Kim, J.L., Sun, L., Yang, S.H., Kodadek, T., and Burley, S.K. 1999. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.* **13:** 3217-3230.

Rachwal, P.A., Brown, T., and Fox, K.R. 2007a. Sequence effects of single base loops in intramolecular quadruplex DNA. *FEBS Lett.* **581:** 1657-1660.

Rachwal, P.A., Findlow, I.S., Werner, J.M., Brown, T., and Fox, K.R. 2007b. Intramolecular DNA quadruplexes with different arrangements of short and long loops. *Nucleic Acids Res.* **35:** 4214-4222.

Saenger, W. 1994. *Principles of nucleic acid structure*. Springer-Verlag, Berlin.

Santelli, E. and Richmond, T.J. 2000. Crystal structure of MEF2A core bound to DNA at 1.5 A resolution. *J. Mol. Biol.* **297:** 437-449.

Stevanin, G. and Brice, A. 2006. Spinocerebellar ataxia 17 and Huntington's disease-like 4. In *Genetic instabilities and neurological diseases* (eds. R.D. Wells and T. Ashizawa), pp. 475-483. Elsevier/Academic Press, Amsterdam.

Tomiuk, J., Bachmann, L., Bauer, C., Rolfs, A., Schols, L., Roos, C., Zischler, H., Schuler, M.M., Bruntner, S., Riess, O. et al. 2007. Repeat expansion in spinocerebellar ataxia type 17 alleles of the TATA-box binding protein gene: an evolutionary approach. *Eur. J. Hum. Genet.* **15:** 81-87.

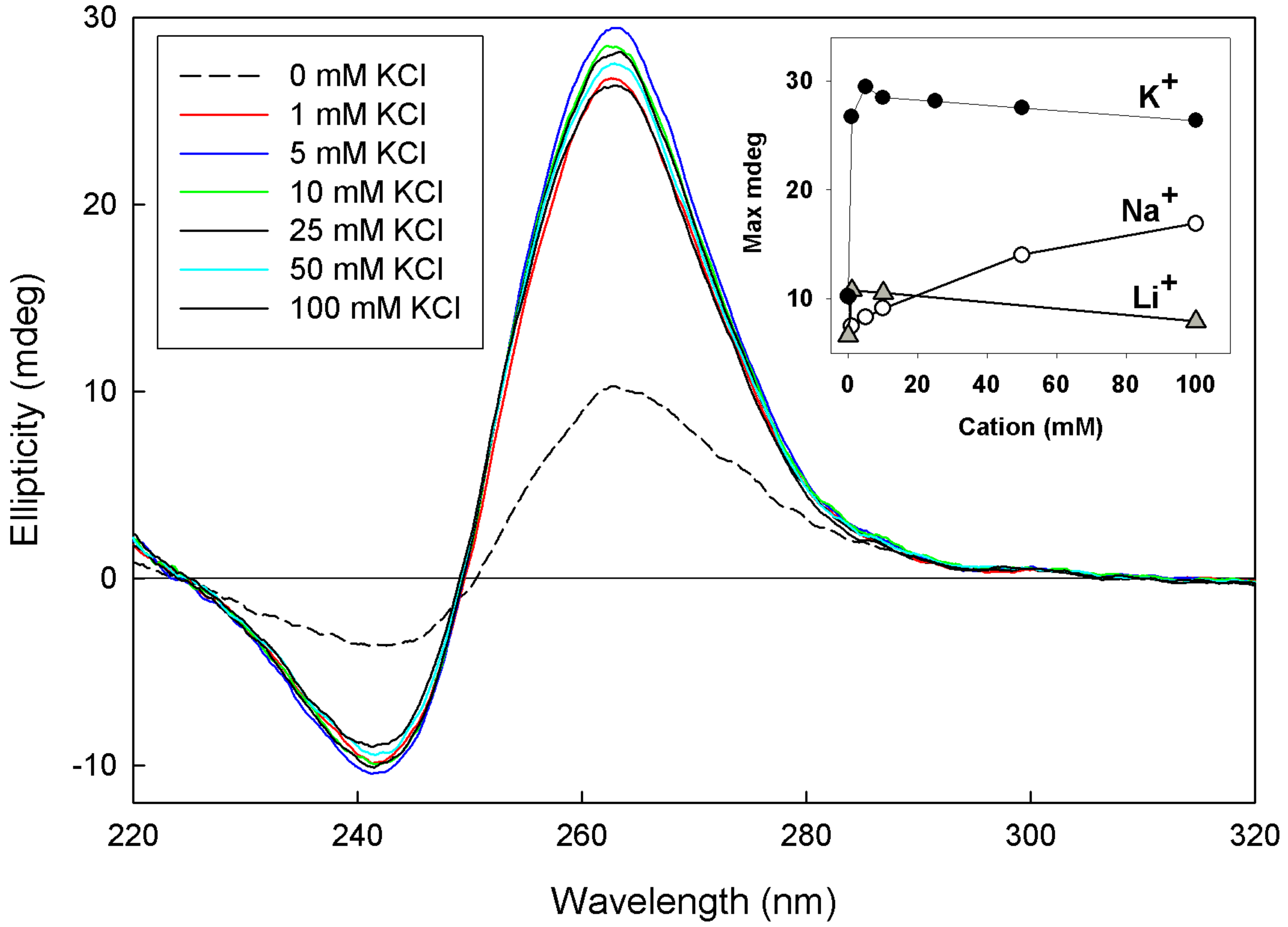
Vorlickova, M., Chladkova, J., Kejnovska, I., Fialova, M., and Kypr, J. 2005. Guanine tetraplex topology of human telomere DNA is governed by the number of (TTAGGG) repeats. *Nucleic Acids Res.* **33:** 5851-5860.

Wells, R.D. and Larson, J.E. 1972. Buoyant density studies on natural and synthetic deoxyribonucleic acids in neutral and alkaline solutions. *J. Biol. Chem.* **247:** 3405-3409.

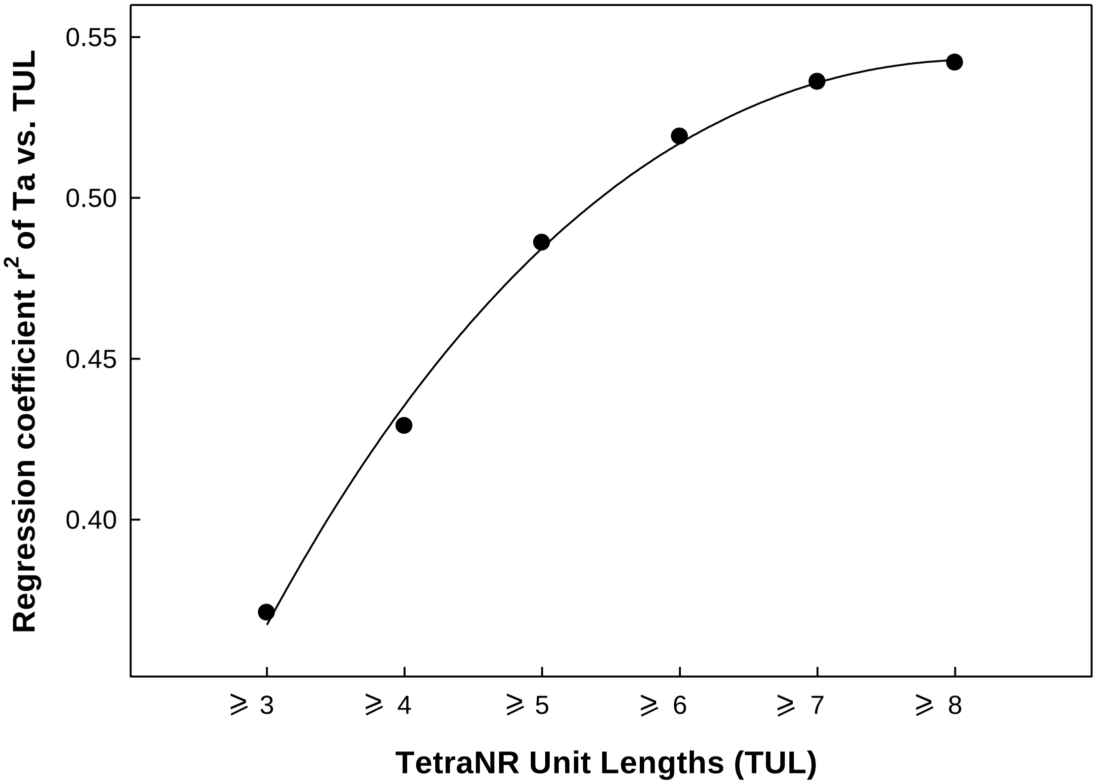
Wojciechowska, M., Napierala, M., Larson, J.E., and Wells, R.D. 2006. Non-B DNA conformations formed by long repeating tracts of DM1, DM2 and FRDA genes, not the sequences per se, promote mutagenesis in flanking regions. *J. Biol. Chem.*

Xu, Y. and Sugiyama, H. 2006. Formation of the G-quadruplex and i-motif structures in retinoblastoma susceptibility genes (Rb). *Nucleic Acids Res.* **34:** 949-954.

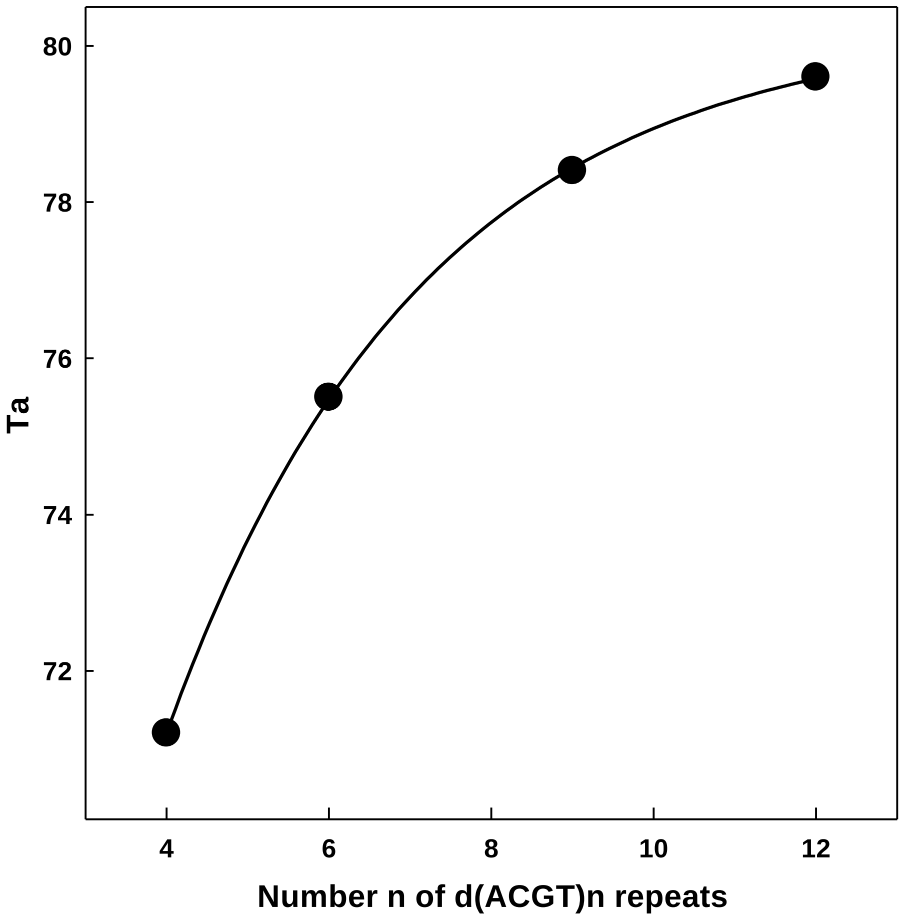
Zahra, R., Blackwood, J.K., Sales, J., and Leach, D.R. 2007. Proofreading and secondary structure processing determine the orientation dependence of CAG x CTG trinucleotide repeat instability in Escherichia coli. *Genetics* **176:** 27-41.



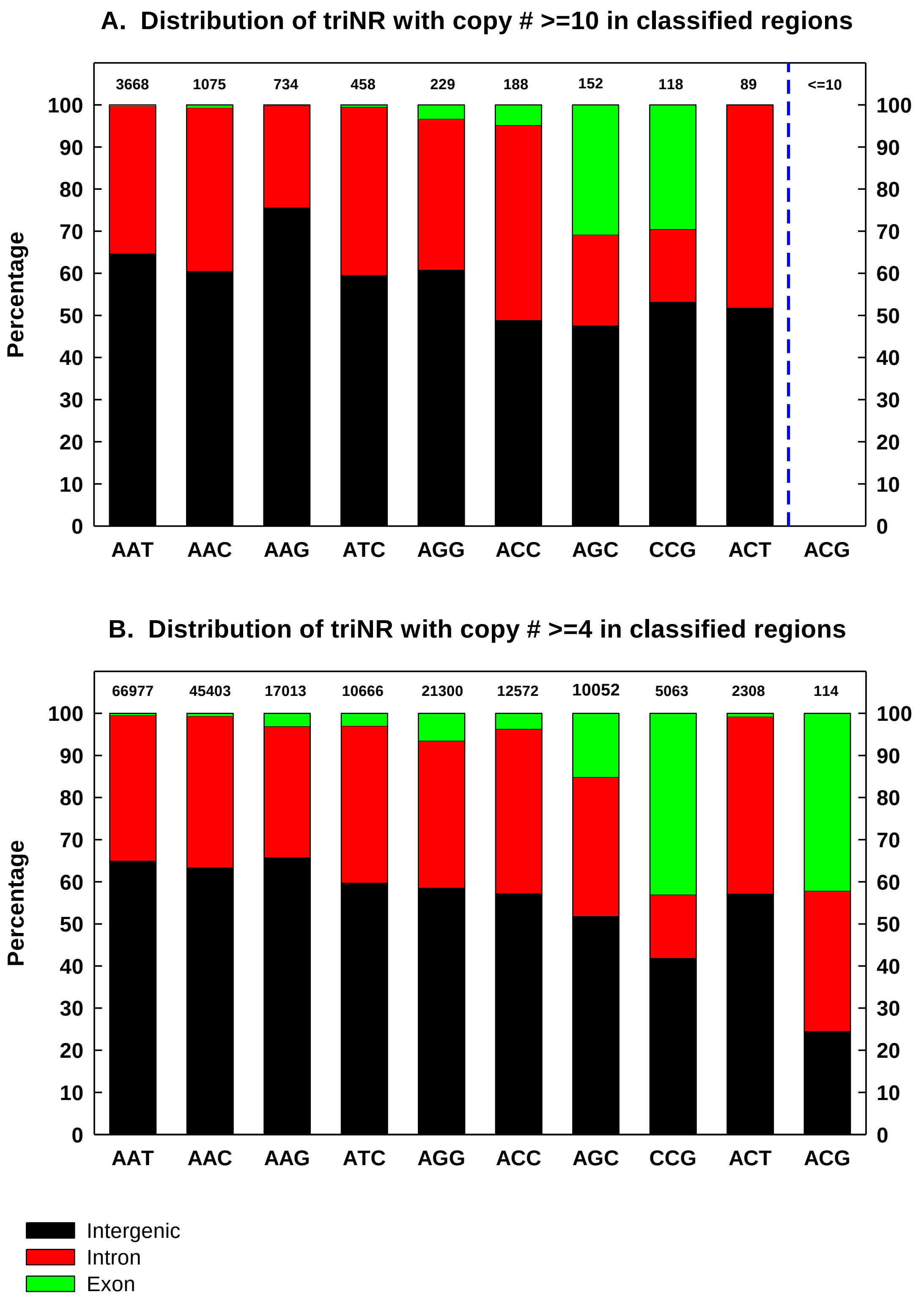
Supplementary Fig. 1



Supplementary Fig. 2



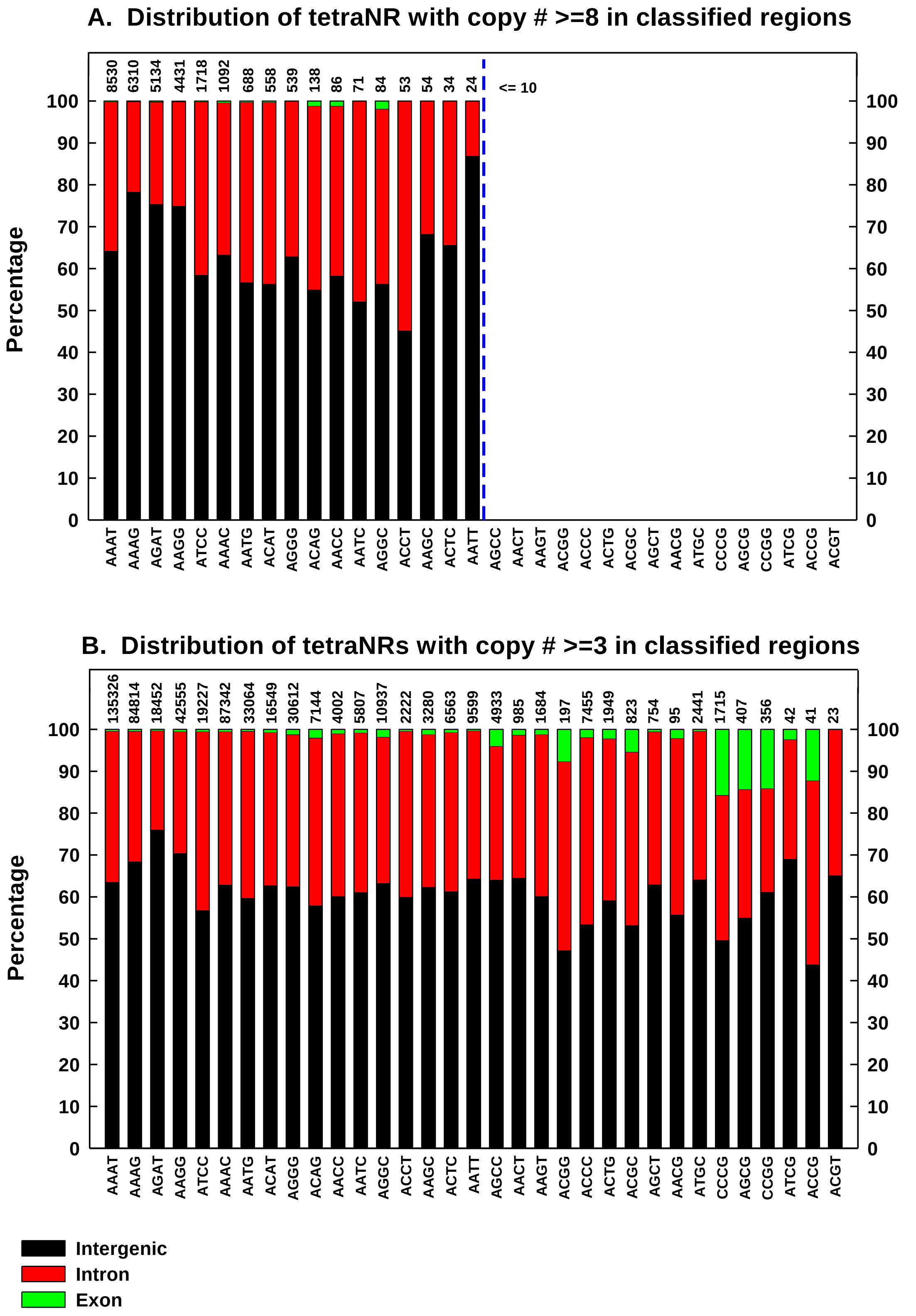
Supplementary Fig. 3



A. Distribution of triNR with copy # >=10 in classified regions

B. Distribution of triNR with copy # >=4 in classified regions

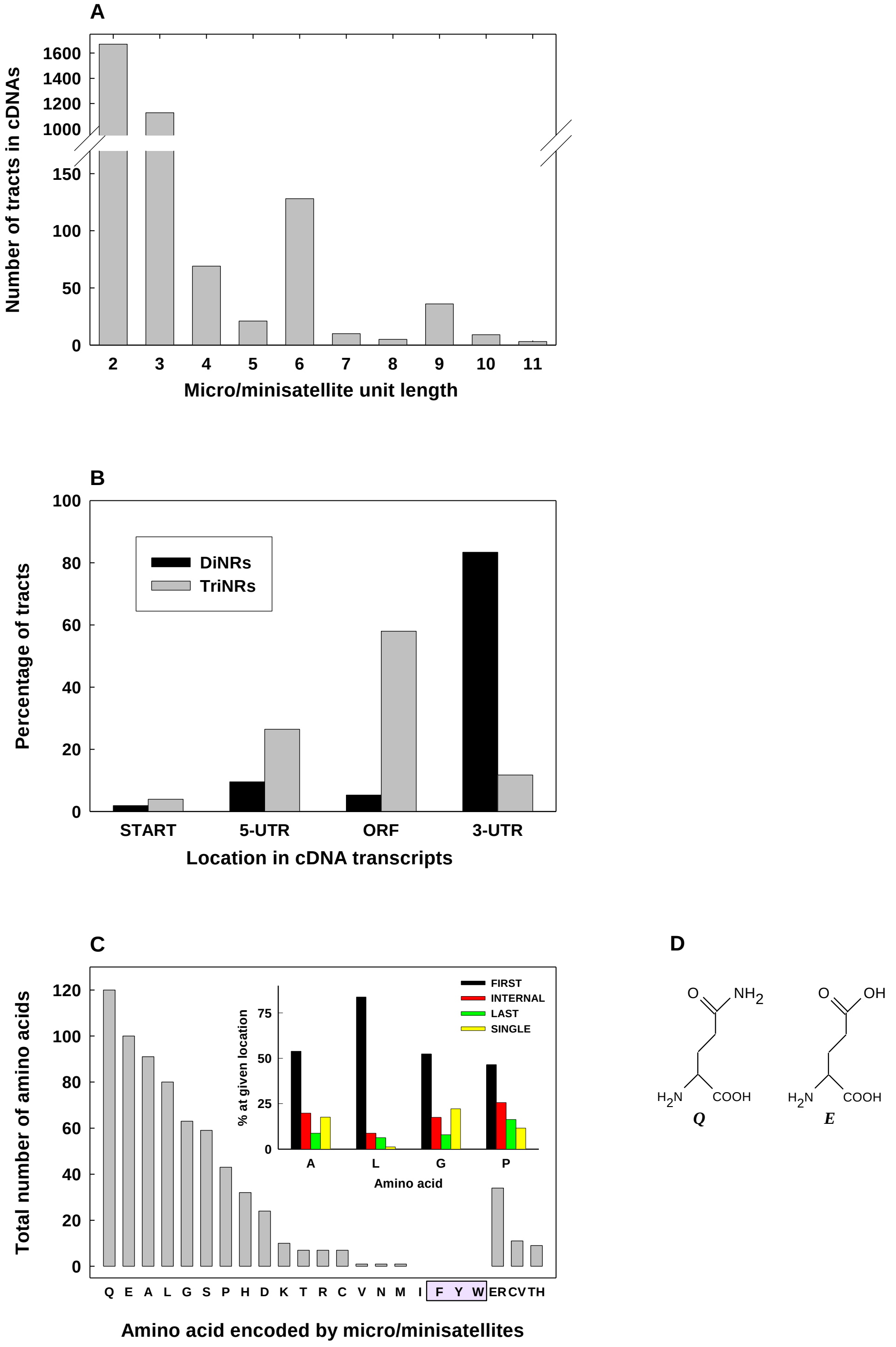
Supplementary Fig. 4



A. Distribution of tetraNR with copy # >=8 in classified regions

B. Distribution of tetraNRs with copy # >=3 in classified regions

Supplementary Fig. 5



Supplementary Figure 6

Supplementary Figure 7. Evolutionary conservation of three human coding triNR tracts containing >10 repeats in the reference human genome assembly hg18

A. *TBP* gene encoding the TATA box binding protein (SCA17)

Protein sequence flanking the poly(Q) region (boxed)

```
Human         ENSP00000375942     TPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQR QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ AVAAAAVQQSTSQQATQ------------GTSGQAPQ
Chimpanzee    ENSPTRP00000032146  TPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQR QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ------ AVAAAAVQQSTSQQATQ------------GTSGQAPQ
Orangutan     Chr6.1611.1         TPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQR QQQQQQQQQQQQQQQQQQQQQQQQQQ------------ AVAAAAVQQSASQQATQ------------GTAGQAPQ
Rhesus        ENSMMUP00000020642  TPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQR QQQQQQQQQQQQQQQQQQQQQQQQQ------------- AVAAAAVQQSTSQQTTQ------------GTSGQAPQ
Marmoset      Contig3625.002.a              MPYGTGLTPQPIQNTNSLSILEEQQR QQQQQQQQQQQQQQQQQQQQQQQQQQQ----------- AVAAAAVQQSTSQQATSQQATQ-------GTSGQAPQ
Mouse         ENSMUSP00000014911  TPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQR QQQQQQQQQQQQQ------------------------- AVATAAASVQQSTSQQPTQ----------GASGQTPQ
Rat           ENSRNOP00000002038  TPGIPIFSPMMPYGTGLTPQPVQNTNSLSILEEQQR QQQQQQQQQQQQQQQ----------------------- AVATAAASVQQSTSQQPTQ----------GASGQTPQ
Cat           ENSFCAP00000006527  TPGIPIFSPMMPYGTGLTPQPIQSTNSLSILEEQQR QQQQQQQQQQAQQQQQQQQQQAQQQQQQQ--------- AVAAVQQSASQQATQ--------------AASGQTPQ
Dog           Chr12.76.035.a      TPGIPIFSPMMPYGTGLTPQPIQNTNSLSLLEEQQR QQQQQQQAQQQQAQQQQQQQQQAQQQ------------- AVTAVQQSTSQQATQ--------------GASGQTPQ
Horse         ENSECAP00000006680  TPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQR QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ-------- AAAAVQQSTSQQATQ--------------GASGQTPQ
Cow           ENSBTAP00000010109  TPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQR QQQQQQQQQQQQQQQQQQQ------------------- AAVAAVQQSTSQQATQ-------------GPSGQTPQ
Opossum       Chr2.46.028.a       TPGIPLFSPMMPYGTGLTPQPVQSTSSLSILEEQQR QQQQQQQQQQQQ-------------------------- AAQQAATQQATQ-----------------GTSGQTPQ
Platypus (1)  ENSOANP00000020337  TPGIPIFSPMMPYGTGLTPQPVQSSNSLSLLEEQRR QQQQQQQQQQQQQQQ----------------------- AATAQQAPA--------------------GASGQTPQ
Chicken       ENSGALP00000036926  TPGIPIFSPMMPYGTGLTPQPVQSTNSLSILEEQQR QQQQQQ-------------------------------- AAQQSTSQQATQ-----------------GTSGQTPQ
Lizard   (2)                      TPGIPIFSPMMPYGTGLTPQPVQTTNSLSILEEQQR QQQQ---------------------------------- AAAQQSTSQPTQ-----------------GSSGQTPQ
Xenopus       ENSXETP00000018725  TPGINIFSPLMPYGTGLTPQPVQTTNSLSILEEQQR QQQQ---------------------------------- AQQSTSQQGNQ------------------GS-GQTPQ
Zebrafish     ENSDARP00000010211  TPGLPIFSPMMPYGTGLTPQPVQNSNSLSLLEEQQR QQQQQQ-------------------------------- AASQQQGGMV-------------------GGSGQTPQ
Tetraodon     GSTENP00032802001   TPGMPIFSPMMPYGSGLTPQPVQNTNSLSILEEQQR QQQQQQQQQQQQQ------------------------- AQQAQQAGTGIPGT-----------------------
Fugu          ENSTRUP00000017711  TPSMPIFSPMMPYGSGLTPQPVQNTNSLSILEEQQR QQQQQQ-------------------------------- AQQANTESEIQEFYFLVCVWVENYSLGIPGTSGTTPQ
Stickleback   ENSGACP00000021304  TPGMSMFSPMMPYGSGLTPQPVQNTNSLSILEEQQR QQQQQQ-------------------------------- AQQANAGIPLTRLP---------------GTSGQTPQ
Medaka        ENSORLP00000013489  TPTMPVFSPMMPYGSGLTPQPVQNTNSLSILEEQQR QQQQQQQQQQ---------------------------- TQQLNTGLP--------------------GASGQTPQ
```

Alignment of the CAG/CAA repeat tract in the boxed poly(Q)-coding region above

```
Human            (CAG)3  (CAA)3  (CAG)8   CAA     CAG     CAA    (CAG)19  CAA   CAG
Chimpanzee (3)   (CAG)3  (CAA)3  (CAG)7   CA      ---     ---    (CAG)17  CAA   CAG
Orangutan        (CAG)3  (CAA)2  (CAG)8   CAA     CAG     CAA    (CAG)8   CAA   CAG
Rhesus           (CAG)3  (CAA)2  (CAG)5  (CAA)2   CAG    (CAA)2  (CAG)8   CAA   CAG
Marmoset         (CAG)4   CAA    (CAG)14 (CAA)2  (CAG)2   CAA    (CAG)3
Mouse            (CAG)3   CAA     CAG     CAA    (CAG)3  (CAA)2  (CAG)2
Rat              (CAG)8   CAA    (CAG)3   CAA    (CAG)2
Cat              (CAG)5   CAA    (CAG)2   CAA     CAG     gcg    (CAG)6   CAA  (CAG)3  gct  (CAG)7
Dog               ---     CAA    (CAG)6   gcc    (CAG)4   gcc    (CAG)5   CAA  (CAG)2  gcc  (CAG)3
Horse            (CAG)3   CAA    (CAG)11  CAA    (CAG)8   CAA    (CAG)5
Cow              (CAG)5   CAA    (CAG)13
Opossum          (CAG)8   CAA    (CAG)2   CAA
Platypus         (CAG)7  (CAA)2  (CAG)3   CAA    (CAG)2
Chicken          (CAG)4   CAA     CAG
Lizard            ---     CAA     CAG     CAA     CAG
Xenopus           ---    (CAA)3   CAG
Zebrafish         ---    (CAA)2  (CAG)4
Tetraodon         CAG    (CAA)2   CAG    (CAA)2  (CAG)5   CAA     CAG
Fugu              CAG     CAA     CAG    (CAA)3
Stickleback       ---    (CAA)2  (CAG)2   CAA     CAG
Medaka            ---     CAA     CAG    (CAA)3  (CAG)3  (CAA)2
```

B. *MEF2A* gene encoding the MADS box transcription enhancer factor 2, polypeptide A (myocyte enhancer factor 2A)

Protein sequence flanking the poly(Q) region (boxed)

```
Human       ENSP00000284368      PPRDRMTPSGF-- QQQQQQQQQQQ PPPPPQPQPQ-PPQPQPRQE-MGRSPVDSLSSSSSSYDGSDREDPRGD
Chimpanzee  ENSPTRP00000012805   PPRDRMTPSGF-- QQQQQQQ---- PPPPPQPQPQ-PPQPQPRQE-MGRSPVDSLSSSSSSYDGSDREDPRGD
Mouse       ENSMUSP00000072281   PPRDRMTPSGF-- QQQQQQ----- PQQQPPPQ---PPQPQPRQE-MGRSPVDSLSSSSSSYDGSDREDPRGD
Rat         DQ323505 (4)         PPRDRMTPSGF-- QQQQQQQ---- PQQQPPPQPP-QPQPQPRQE-MGRSPVDSLSSSSSSYDGSDREDPRGD
Dog         Chr3.44.008.a        PPRDRMTPSGF-- QQQQ------- PQQQQPPPPQPSQPPQPRQE-MGRSPVDSLSSSSSSYDGSDREDPRGD
Cow         ENSBTAP00000014074   PPRDRMTPSGF-- QQQQQ------ PQPPPPPPQ--APQPQPRQE-VGRSPVDSLSSSSSSYDGSDREDPRGD
Chicken     ENSGALP00000033241   PPRDRVTPSGFP- QQQ-------- PPQQPQPPQP-PQQPPQRQE-MGRSPVDSLSSSSSSYDGSDREDPRSD
Zebrafish   ENSDARP00000074652   PPRERVTPSGFPP QQQQ------- PPSGRPD-------------MGRSPVDSLSSSCSSYDGSDREDHRPD
Fugu        ENSTRUP00000017406   PPRERVTPSGFPP QQ--------- PQQGSSR-----------QEVLGRSPADSLSSSCSSYDGSDREDHRPD
```

All boxed Q resides above are encoded by CAG repeats except the first boxed Q of the mouse genome, which is encoded by a CAA repeat

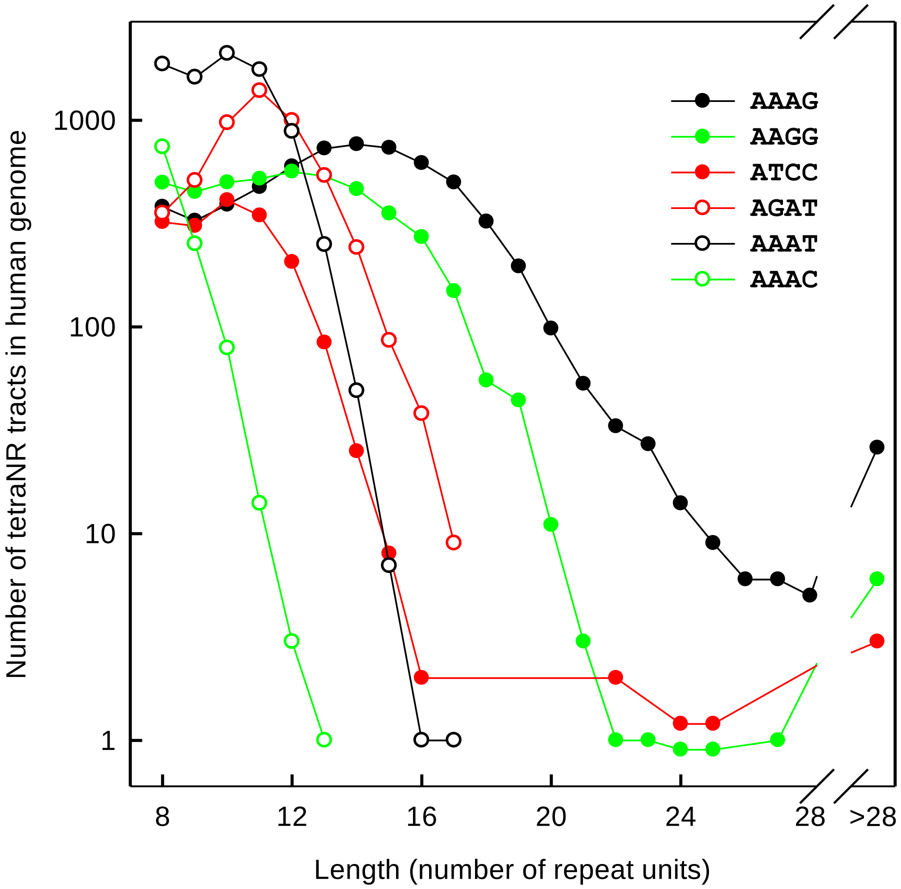
C. *POU4F2* gene encoding the POU domain class 4 transcription factor 2

Protein sequence flanking the conserved poly(G) region (boxed)

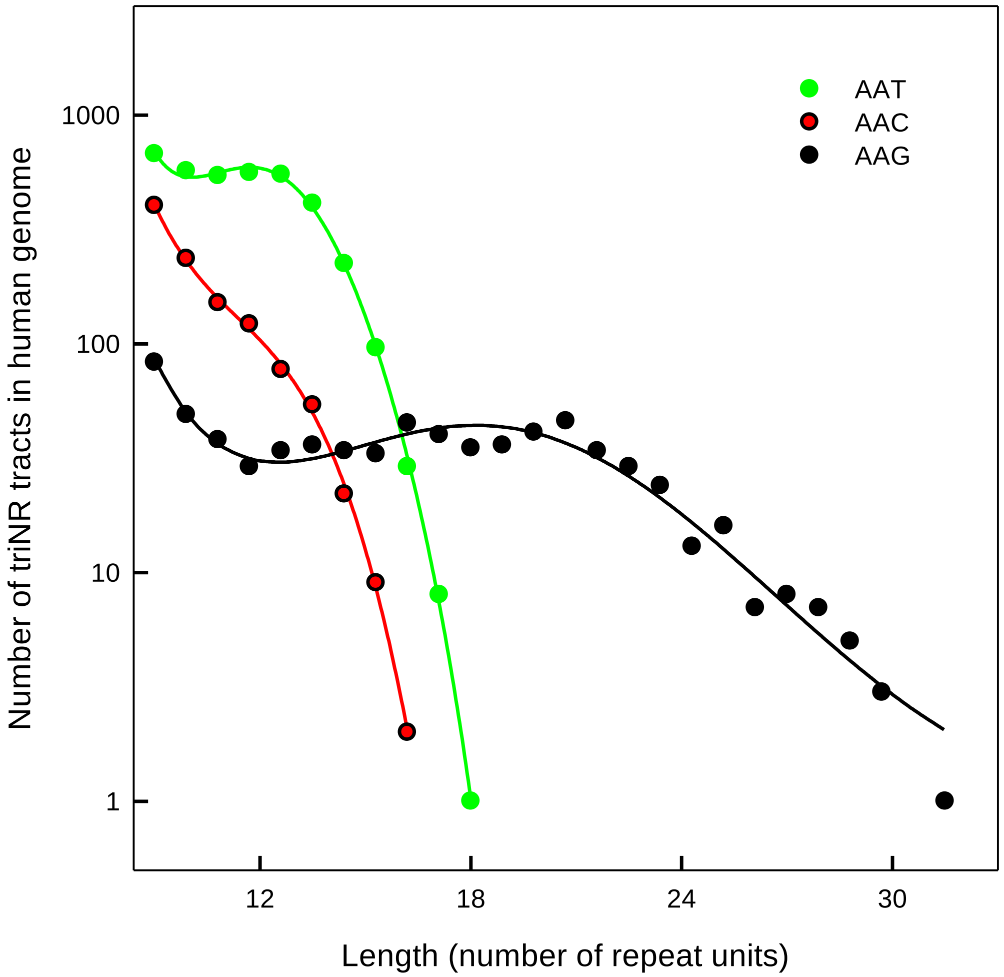
```
Human.     ENSP00000281321        PIAPSASSPSSSSNA---- GGGGGGGGGGGGGGG- RSSSSSSSGSSGGGG--------------SEAMRRACLPTPPSNIFGGLDESLLA
Chimpanzee ENSPTRP00000028307     PTAPSASSPSSSSNA---- GGGGGGGGGGGG---- RSSSSSSSGSSGGGG--------------SEAMRRACLPTPPSNIFGGLDESLLA
Orangutan  Chr4.1034.1            PTAPSASSPSSSSNA---- GGGGGGGGGGG----- RSSSSSSSGSSGSGGGG------------SEAMRRACLPTPPSNIFGGLDESLLA
Rhesus     ENSMMUP00000030857     PTAPSASSPSSSSNA---- GGGGGGGGGGGG---- RNSSSSSSGSSGSGGGG------------SEAMRRACLPTPPSNIFGGLDESLLA
Marmoset   Contig6718.002.a       PTAPSASSPSSSSNA---- GGGGGGGGGGGG---- RSSSSSSSGSSGSGGGG------------SEALRRACLPTPPSNIFGGLDESLLA
Mouse:     ENSMUSP00000034115     PAAPSASSPSSSSNA---- GGGGGGGGGGGGGG-- RSSSSSSSGSGGSGGGGG-----------SEAMRRACLPTPPSNIFGGLDESLLA
Rat:       ENSRNOP00000016422     PAAPSASSPSSSSNA---- GSGGGGGGGGGGGGGG RSSSSSSSGSGGGGGGG------------SEAMRRACLPTPPSNIFGGLDESLLA
Cat        ENSFCAP00000002806     -----------SSNA---- GGGGGSGGGCGG---- RSSSSSSSGSSGSSGGG------------SEAMRRACLPTPPSNIFGGLDESLLA
Dog:       Chr15.49.002.a         PAAPSASSPSSSSSA---- GGGGGSGGGGGG---- RSSSSSSSGSSGSSGGG------------SEAMRRACLPTPPSNIFGGLDESLLA
Horse      ENSECAP00000018181     PAAPSASSPGSSSNA---- GGGGSSGGGGGG---- RSSSSSSSGSSGGGG--------------SEAMRRACLPTPPSNIFGGLDESLLA
Cow:       ENSBTAP00000007421     PAAPSASSPSSSSNA---- GGGGGSGGGGGG---- RSSSSSSSGSSGGSGGG------------SEAMRRACLPTPPSNIFGGLDESLLA
Opossum    Chr5.14.008.a          CTSSSAAPSSSSPSNTSS- GGGGGGSGG------- RSSNSGSSGSSGGGGGGGGSGGGGG----SEAMRRACLPTPPSNIFGGLDESLLA
Platypus   ENSOANP00000010988     CTSAPAAPSSSSPSSSSSP GGGGAGAGGAGGG--- SAGSSSSGG--------------------PEAMRRACLPTPPSNIFGGLDESLLA
Chicken:   Chr13_random:8436-8657 (5)   POU4F1<                             >POU4F2                 GNIFAGFDETLLR
Zebraf.:   AY196176               SLHSSSSSSTLTSNAPSSSCSSS------------ RHSSTISSSGGGS----------------SEAMRRACLPTPPSNIFGGLDESLLA
Fugu:      ENSTRUP00000008461     -LHSSSSST-LTSNAPSS-CSSS------------ RHSSTIISSSGGSS---------------SEAMRRACLPTPPSNIFGGLDESLLA
```

Highlight of (four or more consecutive) GGC repeats in the boxed poly(G)-coding region abve

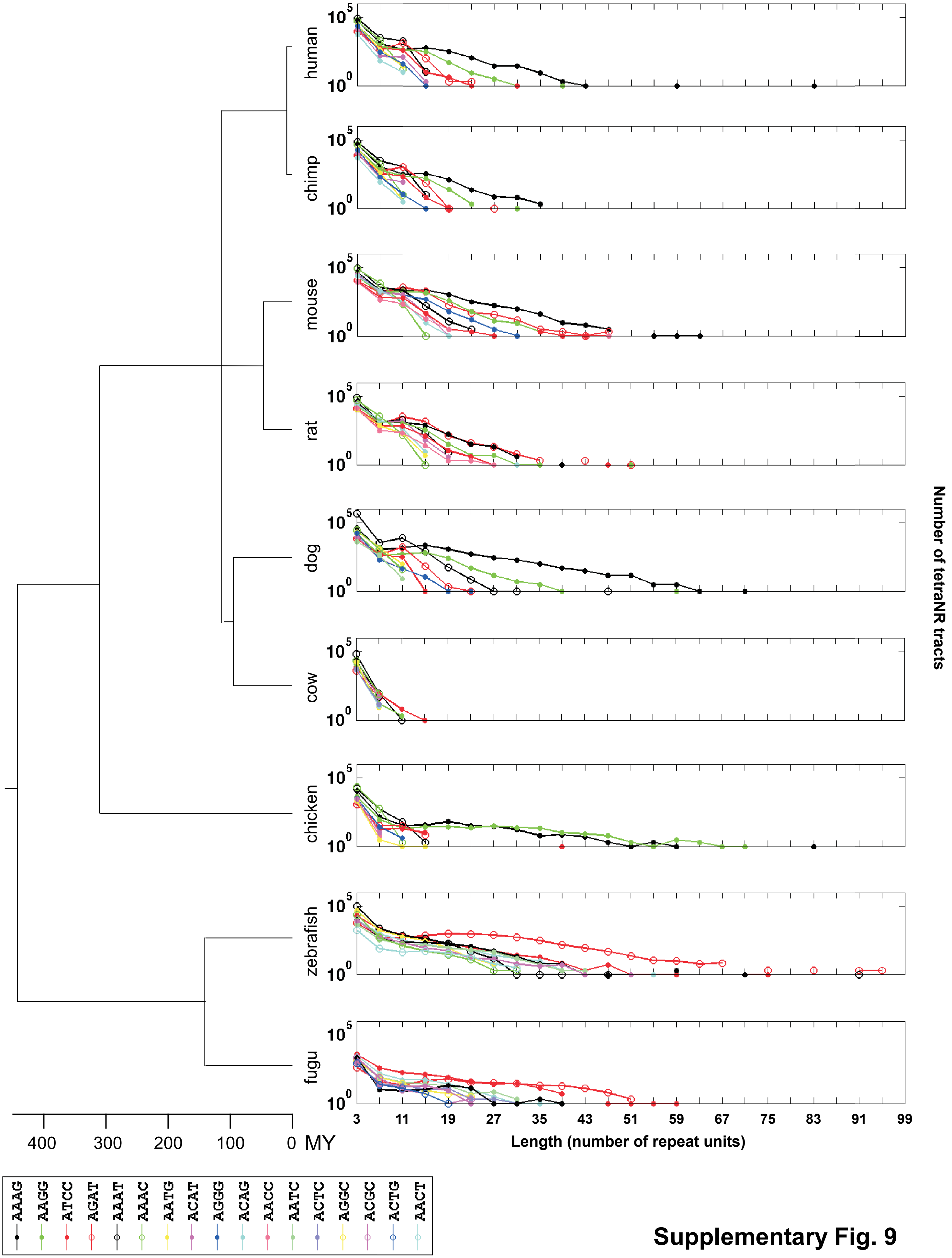
```
Human      GGT GGT GGC GGC GGC GGC GGC GGC GGC GGC GGC GGC GGC GGA GGC
Chimpanzee GGT GGC GGC GGC GGG GGC GGC GGC GGC GGC GGA GGC
Orangutan  GGT GGC GGT GGC GGC GGC GGC GGC GGC GGA GGC
Rhesus     GGT GGC GGT GGA GGC GGC GGC GGC GGC GGC GGA GGC
Marmoset   GGT GGC GGC GGT GGC GGC GGC GGC GGA GGA GGA GGA
Mouse      GGC GGC GGC GGC GGT GGC GGC GGA GGC GGA GGC GGC GGC GGC
Rat        GGC agc GGC GGC GGC GGC GGC GGC GGC GGC GGC GGT GGT GGC GGA GGC
Cat        GGC GGC GGT GGT GGC agt GGC GGG GGC tgt GGA GGC
Dog        GGC GGC GGC GGC GGC agt GGC GGG GGC GGC GGA GGC
Horse      GGC GGC GGT GGC agt agt GGC GGA GGC GGT GGA GGC
Cow        GGC GGC GGC GGC GGG agt GGC GGG GGC GGT GGA GGC
Opossum    GGG GGC GGT GGT GGA GGC agt GGG GGC
Platypus   GGG GGC GGC GGA gca GGA gcg GGA GGA gca GGA GGA GGC
```



Supplementary Fig. 8A



Supplementary Fig. 8B



Supplementary Fig. 9

Supplementary Table 1. Association of triNRs and tetraNRs with inherited human disease/traits

A. TetraNRs

Disease - phenotypic trait	Gene	Chromosomal location	Amplet	Normal copy length or number	Pathological copy length or number	Location	Author	Journal	Vol.	Page	Year
Panic disorder, association	*CCK*	3p21-pter	Mixed TRs	363-399 bp	379-383 bp	5' UTR	Hattori	Mol Psychiatry	6	465	2001
Atherogenic lipoprotein phenotype	*CETP*	16q13	GAAA/GGAA (R-tract)	324-400 bp	404-464	promoter	Talmud	Circulation	101	2461	2000
Myotonic dystrophy	*CNBP*	3q21	CCTG	104-176 bp	75-11000	IVS1	Liquori	Science	293	864	2001
Hyperandrogeny in women, association	*CYP19A1*	15q21	TTTA	7-13	7	IVS4	Baghaei	Obesity Res	11	578	2003
Obesity in women, association	*CYP19A1*	15q21	TTTA	7-13	7	IVS4	Baghaei	Obesity Res	10	115	2002
Breast cancer, increased risk association	*CYP19A1*	15q21	TTTA	7,8,9,11,12	12	IVS4	Kristensen	Pharmaco genetics	8	43	1998
Prostate cancer, association	*CYP19A1*	15q21	TTTA	167-191 bp	171, 187 bp	IVS4	Latil	Cancer	92	1130	2001
Osteoporosis, association	*CYP19A1*	15q21	TTTA	7-14	Long alleles = protective	IVS4	Masi	J Clin Endocrinol Metab	86	2263	2001
Osteoporosis in late menopause, association	*CYP19A1*	15q21	TTTA	8,10-13	7	IVS4	Riancho	Bone	36	917	2005
Breast cancer, increased risk, association	*CYP19A1*	15q21	TTTA	6-9,11-13	10	IVS4	Ribeiro	Toxicol Lett	164	90	2006
Breast cancer, increased risk, association	*CYP19A1*	15q21	TTTA	168, 175-191bp	171 bp	IVS4	Siegelmann-Danieli	Br J Cancer	79	456	1999
Low bone mineral density in elder men, association	*CYP19A1*	15q21	TTTA	7-13	7	IVS4	Van Pottelbergh	J Clin Endocrinol Metab	88	3075	2003
Graves disease, susceptibility, association	*GC*	4q12	TAAA	6,8,10	8	IVS8	Pani	J Clin Endocrinol Metab	87	2564	2002
Juvenile absence epilepsy, association	*GRIK1*	21q22	AGAT	7-8, 10-12	9	non-coding	Sander	Am J Med Genet	74	416	1997

Schizophrenia, increased risk	*JARID2*	6p23-p24	TGAA	11-13	8,10	3' end	Pedrosa	Am J Med Genet B	144B	45	2007
Obesity, association	*LEP*	7q31	TTTC	Allele I - short	Allele II - long	3' UTR	McGarvey	Int J Obes	26	783	2002
Preeclampsia, increased risk	*LEP*	7q31	TTTC	Alleles I/I II/II I=short, II = long	Alleles I/II	3' UTR	Muy-Rivera	Physiol Res	54	167	2005
Hypertension, association	*LEP*	7q31	TTTC	Allele II - long	Allele I - short	3' UTR	Shintani	J Clin Endocrinol Metab	87	2909	2002
Inflammatory polyarthritis, susceptibility	*MIF*	22q11	CATT	5-6	7	promoter	Barton	Genes Immun	4	487	2003
Severity of rheumatoid arthritis, association	*MIF*	22q11	CATT	5	6-8	promoter	Baugh	Genes Immun	3	170	2002
Celiac disease, susceptibility	*MIF*	22q11	CATT	5-6	7 (8 is rare)	promoter	Nunez	Genes Immun	8	168	2007
Infection in cystic fibrosis, susceptibility	*MIF*	22q11	CAAT	5	6-8	promoter	Plant	Am J Respir Crit Care Med	172	1412	2005
Obesity, increased risk	*MIF*	22q11	CATT	5	6-8	promoter	Sakaue	Int J Obes	30	238	2006
Obsessive-compulsive disorder, association	*MOG*	6p21.3-p22	Mixed TRs	Alleles 1, 3-7	Allele 2	3' UTR	Zai	Am J Med Genet	129B	64	2004
Recurrent early onset major depressive disorder, association	*None*		CTAT	96-120,128 bp	124 bp		Philibert	Am J Med Genet	121B	39	2003
Increased height, association	*PTHR1*	3p21-p22	AAAG	3,5-8	6	promoter	Minagawa	J Clin Endocrinol Metab	87	1791	2002
Increased height, association	*PTHR1*	3p21-p22	AAAG	2,5-9	6	promoter	Scillitani	Hum Genet	119	416	2006
Diabetes, type 2, association	*TCF7L2*	10q25	CTTT	6	5, 7-11	IVS3	Grant	Nat Genet	38	320	2006
Diabetes, type 2, association	*TCF7L2*	10q25	CTTT			IVS3	Hayashi	Diabetologica	50	980	2007
Personality traits	*TFAP2B*	6p12	CAAA	4-5	4-5	IVS2	Damberg	Mol Psychiatry	5	220	2000
Functional study	*TH*	11p15	TCAT	5-10		IVS1	Albanese	Hum Mol Genet	10	1785	2001
Tobacco dependence, protection, association	*TH*	11p15	TCAT	6-10	7 (protective)	IVS1	Anney	Pharmaco genetics	14	73	2004
Schizophrenia, susceptibility	*TH*	11p15	TCAT	6-10	7	IVS1	Jacewicz	Forensic Sci Int	162	24	2006
Tobacco dependence, protection, association	*TH*	11p15	TCAT	6-10	7 (protective)	IVS1	Olsson	Behav Genet	34	85	2004

Physiology of catecholamine secretion, blood pressure and heart rate	*TH*	11p15	TCAT	5-10		IVS1	Zhang	Physiol Genomics	19	277	2004
Functional Study	*TNFRSF8*	1p36	ATCC	4, 11		promoter	Croager	Am J Pathol	156	1723	2000
Hodgkin's lymphoma - anaplastic large cell lymphoma	*TNFRSF8*	1p36	ATCC (ACCC, ATGC and GTCC)	1050 - 1612 bp		promoter	Durkop	Biochim Biophys Acta	1519	185	2001

B. TriNRs

Fragile site, FRAXE	*AFF2*	Xq28	CCG	4-39	>200	5' UTR	Knight	Cell	74	127	1993
Spino-bulbar muscular atrophy (Kennedy disease)	*AR*	Xq12	CAG	17-26	40-52	coding region	LaSpada	Nature	352	77	1991
Androgen insensitivity syndrome	*AR*	Xq12	CAG	17-26	8	coding region	Kooy	Am J Med Genet	85	209	1999
Prostate cancer, increased risk, association	*AR*	Xq12	CAG	9-29	<22	exon 1	Stanford	Cancer Res	57	1194	1997
Prostate cancer, increased risk, association	*AR*	Xq12	GGN	8-17	≤16	exon 1	Stanford	Cancer Res	57	1194	1997
Mental retardation and epilepsy	*ARX*	Xp21.3	GCG	10	17	exon 2	Stromme	Nat Genet	30	441	2002
Central hypoventilation syndrome ?	*ASCL1*	12q23.2	CAG	14	15, 17	coding region	Sasaki	Hum Genet	114	22	2003
Osteoarthritis, susceptibility, association	*ASPN*	9q22.31	GAY	13	14	coding region	Kizawa	Nat Genet	37	138	2005
Dentatorubro-pallidoluysian atrophy (Haw river)	*ATN1*	12p13.31	CAG	3-36	48-93	coding region	Brusco	Arch Neurol	61	727	2004
Dentatorubro-pallidoluysian atrophy (Haw river)	*ATN1*	12p13.31	CAG	9-23	40-100	coding region	Koide	Nat Genet	6	9	1994
Schizophrenia, association	*ATXN1*	6p22.3	CAG	25-36	31	coding region	Joo	Psychiatr Genet	9	7	1999
Spinocerebellar ataxia 1	*ATXN1*	6p22.3	CAG	25-36	40-100	coding region	Orr	Nat Genet	4	221	1993
Spinocerebellar ataxia 1	*ATXN1*	6p22.3	CAG	6-44	37-91	coding region	Brusco	Arch Neurol	61	727	2004

Spinocerebellar ataxia 2	*ATXN2*	12q24.12	CAG	14-31	32-500	coding region	Brusco	Arch Neurol	61	727	2004
Multiple sclerosis, susceptibility, association	*ATXN2*	12q24.12	CAG	15-29	22	coding region	Chataway	Neurogenetics	2	91	1999
Spinocerebellar ataxia 2	*ATXN2*	12q24.12	CAG	15-29	35-59	coding region	Sanpei	Nat Genet	14	277	1996
Machado-Joseph disease	*ATXN3*	14q32.12	CAG	12-44	45	coding region	Padiath	Am J Med Genet	133B	124	2005
Machado-Joseph disease	*ATXN3*	14q32.12	CAG	13-36	68-79	coding region	Kawaguchi	Nat Genet	8	221	1994
Machado-Joseph disease	*ATXN3*	14q32.12	CAG	13-47	53-86	coding region	Brusco	Arch Neurol	61	727	2004
Spinocerebellar ataxia 7	*ATXN7*	3p14.1	CAG	7-17	38-130	coding region	David	Nat Genet	17	65	1997
Spinocerebellar ataxia 7	*ATXN7*	3p14.1	CAG	7-35	36- >300	coding region	Brusco	Arch Neurol	61	727	2004
Spinocerebellar ataxia 8	*ATXN8OS*	13q21	CTG	15-50	110-130	3' UTR	Brusco	Arch Neurol	61	727	2004
Spinocerebellar ataxia 8	*ATXN8OS*	13q21	CTG	16-92	107-127	3' UTR	Koob	Nat Genet	21	379	1999
Spinocerebellar ataxia 6	*CACNA1A*	19p13.13	CAG	4-16	21-28	coding region	Zhuchenko	Nat Genet	15	62	1997
Spinocerebellar ataxia 6	*CACNA1A*	19p13.13	CAG	4-18	19-33	coding region	Brusco	Arch Neurol	61	727	2004
Jacobsen syndrome	*CBL*	11q23.3	CCG	11	700-800	5' gene	Jones	Nature	376	145	1995
Diabetic nephropathy, protection , association	*CNDP1*	18q22.3	CTG	5-7	5	exon 2	Janssen	Diabetes	54	2320	2005
Anorexia nervosa, binging/purging, association	*CNR1*	6q15	AAT	12-20	14	?	Siegfried	Am J Med Genet B Neuropsychiatr Genet	125B	126	2004
IV drug dependence, susceptibility, association	*CNR1*	6q15	AAT	12-20	16-20	?	Comings	Mol Psychiatry	2	161	1997
Mental retardation	*DIP2B*	12q13.13	CGG	6-23	250-285	promoter	Winnepenninckx	Am J Hum Genet	80	221	2007
Myotonic dystrophy, increased risk, association	*DMPK*	19q13.32	CTG	5-37	>19	3' UTR	Gennarelli	Hum Genet	105	165	1999
Myotonic dystrophy	*DMPK*	19q13.32	CTG	<30	50- >2000	3' UTR	Brook	Cell	68	799	1992
Increased transcription, association	*FMR1*	Xq27.3	CGG	5-40	41-60	5' UTR	Loesch	J Med Genet	44	200	2007

Fragile X mental retardation syndrome	*FMR1*	Xq27.3	CGG	5-52 and 60-200 (pm)	>200	5' UTR	Fu	Cell	67	1047	1991
Premature ovarian failure, association with	*FMR1*	Xq27.3	CGG	<55	55-200	5' UTR	Allingham-Hawkins	Am J Med Genet	83	322	1999
Congen. anom. of kidney and urinary tract (CAKUT)?	*FOXC1*	6p25.3	GGC	6	7	coding region	Nakano	Tokai J Exp Clin Med	28	121	2003
Fragile site, FRA10A	*FRA10A*	10q23.33	CGG	8-13	>200	5' UTR	Sarafidou	Genomics	84	69	2004
Friedreich ataxia	*FXN*	9q21.11	GAA	7-20	200-900	IVS1	Campuzano	Science	271	1423	1996
Schizophrenia, association	*GCLC*	6p12.1	GAG	7, 9	8	5' UTR	Gysin	Proc Natl Acad Sci USA	104	16621	2007
Warfarin sensitivity, association	*GGCX*	2p11.2	CAA	10	11, 13	IVS6	Shikata	Blood	103	2630	2004
Cancer suceptibility, association	*GSPT1*	16p13.13	GGN	8-12	12	exon 1	Brito	Carcinogenesis	26	2046	2005
Huntington disease	*HD*	4p16.3 - 4p16.2	CAG	<35	36-39	coding region	Quarrell	J Med Genet	44	e68	2007
Huntington disease	*HD*	4p16.3 - 4p16.2	CAG	<35	40-400	coding region	HDCRG	Cell	72	971	1993
Hand-foot-genital syndrome	*HOXA13*	7p15.2	GCN	14	22	coding region	Innis	Hum Mol Genet	13	2841	2004
Hand-foot-genital syndrome	*HOXA13*	7p15.2	GCN	18	29	coding region	Innis	Hum Mol Genet	13	2841	2004
Hand-foot-genital syndrome	*HOXA13*	7p15.2	GCN	18	30	coding region	Innis	Hum Mol Genet	13	2841	2004
Brachydactyly-syndactyly syndrome	*HOXD13*	2q31.1	GCN	15	8	exon 1	Zhao	Am J Hum Genet	80	361	2007
Huntington disease-like 2	*JPH3*	16q24.2	CTG	6-27	44-57	exon 2A	Holmes	Nat Genet	29	377	2001
Juvenile myoclonic epilepsy, decreased risk, association	*KCNN3*	1q21.3	CAG	13-18, 20-21	19	3' gene	Vijai	J Med Genet	42	439	2005
Ataxia, association	*KCNN3*	1q21.3	CAG	18-20	≥22	3' gene	Figueroa	Arch Neurol	58	1649	2001
Mental retardation ?	*MECP2B*	Xq28	GCN	7	10	coding region	Harvey	Am J Med Genet B Neuropsychiatr Genet	144	355	2007
Mental retardation ?	*MECP2B*	Xq28	GCN	7	5	coding region	Harvey	Am J Med Genet B Neuropsychiatr Genet	144	355	2007
Behcet disease, association	*MICA*	ND	GCT	4,5,6,9	6	exon 5	Mizuki	Proc Natl Acad Sci USA	94	1298	1997
Breast cancer, risk, association	*NCOA3*	20q13.12	Gln	20-27	28-29	coding region	Rebbeck	Cancer Res	61	5420	2001

Lower FENO in asthmatics, association	*NOS1*	12q24.22	AAT	8-17	≥12	IVS20	Wechsler	Am J Respir Crit Care Med	162	2043	2000
Schizophrenia, association	*NUMBL*	19q13.2	CAG	14-20	18	coding region	Passos Gregorio	Schizophr Res	88	275	2006
Oculopharyngeal muscular dystrophy	*PABPN1*	14q11.2	GCG	6	7-13	coding region	Brais	Nat Genet	18	164	1998
Increased transcription, association	*PAX7*	1p36.13	CCT	8-11	11	promoter	Syagailo	Gene	294	259	2002
Low LDL cholesterol, association	*PCSK9*	1p32.3	CTG	9	10	coding region	Yue	Hum Mutat	27	460	2006
Schizophrenia, association	*PHOX2B*	4p13	GCN	20	13, 15, 22	coding region	Toyota	Hum Mol Genet	13	551	2004
Central hypoventilation syndrome	*PHOX2B*	4p13	GCN	20	25-29	coding region	Amiel	Nat Genet	33	459	2003
Central hypoventilation syndrome	*PHOX2B*	4p13	GCN	20	25-30	coding region	Sasaki	Hum Genet	114	22	2003
Central hypoventilation syndrome	*PHOX2B*	4p13	GCN	20	25-33	coding region	Matera	J Med Genet	41	373	2004
Central hypoventilation syndrome	*PHOX2B*	4p13	GCN	20	26	coding region	Chen	J Formos Med Assoc	106	69	2007
Male subfertility, association	*POLG*	15q26.1	CAG	5-13	10	coding region	Harris	Int J Androl	29	421	2006
Parkinson disease, association	*POLG*	15q26.1	CAG	6-14	6, 8, 9, 12, 13	coding region	Luoma	Neurology	69	1152	2007
Spinocerebellar ataxia 12	*PPP2R2B*	5q32	CAG	7-28	66-78	5' flanking	Holmes	Nat Genet	23	391	1999
Spinocerebellar ataxia 12	*PPP2R2B*	5q32	CAG	7-31	55-78	promoter	Brusco	Arch Neurol	61	727	2004
Reduced expression, association	*RELN*	7q22.1	GGC	4-13	13	5' UTR	Persico	J Neural Transm	113	1373	2006
Increased insulin sensitivity, association	*RETN*	19p13.2	ATG	7	6	3' UTR	Pizzuti	J Clin Endocrinol Metab	87	4403	2002
Cleidocranial dysplasia	*RUNX2*	6p12.3	GCK	17	27	coding region	Mundlos	Cell	89	773	1997
Cataract formation in myotonic dystrophy	*SIX5*	19q13.32	CTG	5-37	≥50	promoter	Winchester	Hum Mol Genet	8	481	1999
Spinocerebellar ataxia 17	*TBP*	6q27	CAG	25-42	45-63	coding region	Brusco	Arch Neurol	61	727	2004
Spinocerebellar ataxia 17	*TBP*	6q27	CAG	27-44	50-55	coding region	Zuhlke	Eur J Hum Genet	9	160	2001
Spinocerebellar ataxia 17	*TBP*	6q27	CAG	31-42	63	coding region	Koide	Hum Mol Genet	8	2047	1999

Spinocerebellar ataxia 17	*TBP*	6q27	CAG	49	52-53	coding region	Zuhlke	BMC Med Genet	6	27	2005
Colorectal cancer, association	*TGFBR1*	9q22.33	GCG	9	6	exon 1	Bian	J Clin Oncol	23	3074	2005
Bone mineral density, association	*VLDLR*	9p24.2	CGG	5-10	≥8	5' UTR	Yamada	Genomics	86	76	2005
Dementia, increased risk, association	*VLDLR*	9p24.2	CGG	5-11	5	5' UTR	Helbecque	Neurology	56	1183	2001

pm = premutation

K = G or T

N = A, C, G or T

Y = C or T

ND = not determined

HDCRG = Huntington's Disease Collaborative Research Group

Note: the number of repeats in normal and disease-associated states does not differentiate between homo- and heterozygosity and the presence/absence of repeat interruptions.

Supplementary Table 2. EMAST in human cancer

Author's marker	UCSC marker (hg18)	Chromosomal location	Gene	Gene location	mRNA	Protein	Reference Genome Sequence	Sequence designation (this work)	Tumor	MSI% (1)	Author	Journal	Vol.	Page	Year
D8S311	D8S392	8p11.21	*ANK1*	Intron 1	NM_000037	ankyrin 1 isoform 4	(GAAA)17	AAAG	Non-small cell lung cancer	21	Woenckhaus	Int J Oncol	23	1357	2003
D8S1179	CHLC.GATA7G07	8q24.13	*DA879437*	Intron	Spliced EST		(TCTA)11	AGAT	Gastro-intestinal	17	Vauhkonen	Forensic Sci Int	139	159	2004
D19S433	GGAA2A03	19q12	*C19orf2*	Intron 1	NM_003796	RPB5-mediating protein	(AGGA)13	AAGG	Gastro-intestinal	17	Vauhkonen	Forensic Sci Int	139	159	2004
D21S11	D21S11	21q21.1	None	None			(TCTA)11	AGAT	Gastro-intestinal	15	Vauhkonen	Forensic Sci Int	139	159	2004
D18S51	UT574	18q21.33	*BCL2*	Intron 1	NM_000633	B-cell lymphoma protein 2	(GAAA)18	AAAG	Gastro-intestinal	15	Vauhkonen	Forensic Sci Int	139	159	2004
L17686	D7S1482	7q31.32	None	None			R230 (2) including (GAAA)18 (GAAG)10	AAAG - AAGG	Non-small cell lung cancer	14	Ahrendt	Cancer Res	60	2488	2000
VWA	VWF	12p13.31	*VWF*	Intron	NM_000552	von Willebrand factor preproprotein	(TCTG)5 (TCTA)11	ACAG - AGAT	Gastro-intestinal	12	Vauhkonen	Forensic Sci Int	139	159	2004
FGA	FGA	4q32.1	*FGA*	Intron	NM_021871 NM_000508	fibrinogen, alpha polypeptide isoform alpha fibrinogen, alpha polypeptide isoform alpha-E	(TTTC)14	AAAG	Gastro-intestinal	12	Vauhkonen	Forensic Sci Int	139	159	2004
D13S317	CHLC.GATA7G10	13q31.1	None	None			(TATC)11	AGAT	Gastro-intestinal	12	Vauhkonen	Forensic Sci Int	139	159	2004
L17686	D7S1482	7q31.32	None	None			R230 (2) including (GAAA)18 (GAAG)10	AAAG - AAGG	Non-small cell lung cancer	11	Xu	Int J Cancer	91	200	2001

CSF1PO	GDB:212649	5q33.1	*CSF1R*	Intron	NM_005211	colony stimulating factor 1 receptor precursor	(ATAG)13	AGAT	Gastro-intestinal	10	Vauhkonen	Forensic Sci Int	139	159	2004
D7S820	GATA3F01	7q21.11	*SEMA3A*	Intron	NM_006080	semaphorin 3A precursor	(GATA)13	AGAT	Gastro-intestinal	10	Vauhkonen	Forensic Sci Int	139	159	2004
D8S321	D8S321	8q24.21	*CD518126*	Intron	Spliced EST		R189 (1) including (GAAA)11	AAAG	Non-small cell lung cancer	9	Xu	Int J Cancer	91	200	2001
UT5307	No Matches	8 (Author)					(AAAG)19 (Author)	AAAG	Non-small cell lung cancer	9	Xu	Int J Cancer	91	200	2001
D20S82	UT250	20p12.3	*BC038533*	Intron	Spliced EST		(GAAA)17	AAAG	Non-small cell lung cancer	8	Ahrendt	Cancer Res	60	2488	2000
UT5320	UT5320	8q24.13	*CD676340*	Intron	Spliced EST		R171 (2) including (GAAA)14 and (AAGG)8	AAAG - AAGG	Non-small cell lung cancer	7	Ahrendt	Cancer Res	60	2488	2000
UT5320	UT5320	8q24.13	*CD676340*	Intron	Spliced EST		R171 (2) including (GAAA)14 and (AAGG)8	AAAG - AAGG	Non-small cell lung cancer	7	Xu	Int J Cancer	91	200	2001
D20S82	UT250	20p12.3	*BC038533*	Intron	Spliced EST		(GAAA)17	AAAG	Non-small cell lung cancer	7	Xu	Int J Cancer	91	200	2001
D5S818	CHLC.GATA3F03	5q23.2	None	None			(AGAT)11	AGAT	Gastro-intestinal	7	Vauhkonen	Forensic Sci Int	139	159	2004
D16S539	SHGC-17627	16q24.1	None	None			(GATA)11	AGAT	Gastro-intestinal	7	Vauhkonen	Forensic Sci Int	139	159	2004
D2S1338	GGAA3A09	2q35	None	None			(CCTT)13	AAGG	Gastro-intestinal	7	Vauhkonen	Forensic Sci Int	139	159	2004
CSF1R	CSF1R	5q33.1	*PDGFRB*	Exon	NM_002609 + ESTs by transcription initiation and termination	platelet-derived growth factor receptor beta	(TAGA) (Author)	AGAT	Non-small cell lung cancer	6	Xu	Int J Cancer	91	200	2001
ACTBP2	No Matches	5 (Author)					(AAAG)11 (AAAG)15 Author	AAAG	Non-small cell lung cancer	5	Xu	Int J Cancer	91	200	2001
D11S488	D11S488	11q24.1	*8 ESTs*	Intron	Spliced ESTs		R171 (2) including (AAAG)7 - (AAGG)10	AAAG - AAGG	Non-small cell lung cancer	5	Xu	Int J Cancer	91	200	2001

D9S242	UT914	9q33.3	None	None			(AGAA)15	AAAG	Non-small cell lung cancer	5	Xu	Int J Cancer	91	200	2001
D20S85	UT236	20q12	None	None			(TTTC)15	AAAG	Non-small cell lung cancer	5	Xu	Int J Cancer	91	200	2001
D3S1358	D3S1358	3p21.31	*LARS2*	Intron	NM_015340	leucyl-tRNA synthetase 2, mitochondrial	(ATAG)14	AGAT	Gastro-intestinal	5	Vauhkonen	Forensic Sci Int	139	159	2004
D8S321	D8S321	8q24.21	*CD518126*	Intron	Spliced EST		R189 (2) including (GAAA)11	AAAG	Non-small cell lung cancer	4	Ahrendt	Cancer Res.	60	2488	2000
D9S753	UT8063	9q22.32	None	None			(GAAA)9	AAAG	Non-small cell lung cancer	4	Xu	Int J Cancer	91	200	2001
D20S77	UT235	20p13	*AW614549*	Intron	Spliced EST		R273 (2) including (GAAA)15	AAAG	Non-small cell lung cancer	4	Xu	Int J Cancer	91	200	2001
L17835	UT1496	7p11.2	None	None			(AAGG)16	AAGG	Non-small cell lung cancer	3	Ahrendt	Cancer Res	60	2488	2000
D20S85	UT236	20q12	None	None			(TTTC)15	AAAG	Non-small cell lung cancer	3	Ahrendt	Cancer Res	60	2488	2000
UT5307	No Matches	8 (Author)					(AAAG)19 (Author)	AAAG	Non-small cell lung cancer	2	Ahrendt	Cancer Res	60	2488	2000
D9S242	UT914	9q33.3	None	None			(AGAA)15	AAAG	Non-small cell lung cancer	2	Ahrendt	Cancer Res	60	2488	2000
TPOX	GDB:196336	2p25.3	*TPO*	Intron	NM_000547 NM_175719-22	thyroid peroxidase isoform a thyroid peroxidase isoform b-e	(TGAA)8	AATG	Gastro-intestinal	2	Vauhkonen	Forensic Sci Int	139	159	2004
G29028	C19-11H65	19p13.11	*CPAMD8*	Intron	NM_015692	alpha-2-macroglobulin domain	R218 (2) including (AAGG)7 (AAAG)19	AAAG - AAGG	Non-small cell lung cancer	1	Ahrendt	Cancer Res	60	2488	2000
D11S488	D11S488	11q24.1	8 ESTs	Intron	Spliced ESTs		R171 (2) including (AAAG)7 - (AAGG)10	AAAG - AAGG	Non-small cell lung cancer	1	Ahrendt	Cancer Res	60	2488	2000
ACTbeta2	No Matches	5 (Author)					(AAAG)11 (AAAG)15 Author	AAAG	Non-small cell lung cancer	1	Ahrendt	Cancer Res	60	2488	2000

G08460	CHLC.GATA52G02	5p14.1	*BE728951*	Intron 1	Spliced EST	(TCTA)11	AGAT	Non-small cell lung cancer	1	Ahrendt	Cancer Res	60	2488	2000
D7S1482	D7S1482	7q31.32	None	None		R230 (2) including (GAAA)18 (GAAG)10	AAAG - AAGG	Non melanoma skin cancer	EMAST in 75% of tumors (Note 3)	Danaee	Oncogene	21	4894	2002
D7S1482	D7S1482	7q31.32	None	None		R230 (2) including (GAAA)18 (GAAG)10	AAAG - AAGG	Bladder	EMAST in 44% of tumors (Note 3)	Danaee	Oncogene	21	4894	2002
D9S303	CHLC.GATA3D04	9q21.32	*BG674167*	Intron	Spliced EST	(GATA)12	AGAT		Note (3)	Danaee	Oncogene	21	4894	2002
D20S82	UT250	20p12.3	*BC038533*	Intron	Spliced EST	(GAAA)17	AAAG		Note (3)	Danaee	Oncogene	21	4894	2002
D7S1485	UT1496	7p11.2	None	None		(AAGG)16	AAGG		Note (3)	Danaee	Oncogene	21	4894	2002
D9S242	UT914	9q33.3	None	None		(AGAA)15	AAAG		Note (3)	Danaee	Oncogene	21	4894	2002
D5S1502	CHLC.GATA52G02	5p14.1	*BE728951*	Intron 1	Spliced EST	(TCTA)11	AGAT		Note (3)	Danaee	Oncogene	21	4894	2002
D9S252	UT2103	9q21.33	*BX089714* *AA868955* *AA861911*	Intron Transcription	Spliced ESTs	(GATA)10	AGAT		Note (3)	Danaee	Oncogene	21	4894	2002

(1) %MSI = percent of tumor samples showing microsatellite instability.

(2) Rx = Homopurine tract containing x number of purine bases.

(3) EMAST = Elevated microsatellite instability at selected tetraNRs. % EMAST in both non melanoma skin cancer and bladder tumors is for all 7 (D7S1482, D9S303, D20S82, D7S1485, D9S242, D5S1502, D9S252) markers combined.

Supplementary Table 3. Enrichment of human genes containing micro/minisatellites tracts in cDNAs

A. ALL LOCATIONS - 2315 non-redundant genes (311 not in databases)

Database	Enriched Term	Number of test genes	% of test genes	P-Value
GOBP	regulation of cellular process	628	0.27	2.31E-38
	regulation of biological process	659	0.28	3.86E-38
	regulation of cellular physiological process	598	0.26	9.16E-36
	regulation of metabolism	484	0.21	9.86E-35
	regulation of physiological process	606	0.26	3.89E-34
	regulation of cellular metabolism	469	0.20	4.53E-33
	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metab.	449	0.19	1.09E-32
	regulation of transcription	444	0.19	1.60E-32
	development	374	0.16	7.20E-32
	regulation of transcription, DNA-dependent	418	0.18	4.44E-31
	transcription	452	0.20	3.63E-30
	transcription, DNA-dependent	424	0.18	6.14E-30
GOCC	nucleus	643	0.28	9.28E-22
	intracellular membrane-bound organelle	798	0.34	1.57E-09
	membrane-bound organelle	798	0.34	1.64E-09
	integral to plasma membrane	200	0.09	2.20E-07
	intrinsic to plasma membrane	201	0.09	2.26E-07
	plasma membrane	257	0.11	2.62E-05
	synapse	32	0.01	4.79E-05
GOMF	transcription regulator activity	311	0.13	2.01E-40
	protein binding	710	0.31	2.01E-33
	transcription factor activity	226	0.10	2.47E-29
	binding	1447	0.63	1.12E-28
	DNA binding	397	0.17	7.57E-28
	sequence-specific DNA binding	122	0.05	3.83E-23
	nucleic acid binding	525	0.23	4.41E-19
	RNA polymerase II transcription factor activity	67	0.03	1.25E-16

KEGG Pathway	HSA04360:AXON GUIDANCE	37	0.02	2.27E-05
	HSA04010:MAPK SIGNALING PATHWAY	57	0.02	2.10E-04
	HSA04310:WNT SIGNALING PATHWAY	36	0.02	2.45E-04
	HSA04810:REGULATION OF ACTIN CYTOSKELETON	45	0.02	5.70E-04
	HSA04020:CALCIUM SIGNALING PATHWAY	38	0.02	1.68E-03
	HSA04350:TGF-BETA SIGNALING PATHWAY	22	0.01	2.14E-03
	HSA04520:ADHERENS JUNCTION	20	0.01	4.77E-03
SP - PIR	nuclear protein	552	0.24	1.51E-59
	transcription	334	0.14	8.55E-58
	transcription regulation	334	0.14	1.03E-55
	dna-binding	319	0.14	1.86E-47
	phosphorylation	352	0.15	1.89E-40
	activator	101	0.04	3.31E-26
	developmental protein	111	0.05	2.53E-20
	membrane	498	0.22	9.62E-20
	repressor	65	0.03	1.02E-16
	signal	348	0.15	3.82E-16
	triplet repeat expansion	19	0.01	2.47E-15
	chromosomal translocation	53	0.02	8.41E-15
	metal-binding	313	0.14	7.44E-14

B. 5'UTR - 537 non-redundant genes (56 not in databases)

Database	Enriched Term	Number of test genes	% of test genes	P-Value
GOBP	regulation of cellular process	152	0.28	1.23E-09
	regulation of biological process	158	0.29	3.46E-09
	development	94	0.18	3.69E-09
	regulation of cellular physiological process	145	0.27	4.03E-09
	regulation of physiological process	146	0.27	1.64E-08
	regulation of metabolism	113	0.21	1.44E-07
	regulation of transcription, DNA-dependent	97	0.18	9.97E-07
	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metab.	103	0.19	1.07E-06

	nervous system development	34	0.06	1.11E-06
	system development	34	0.06	1.34E-06
	regulation of cellular metabolism	107	0.20	1.36E-06
	regulation of transcription	101	0.19	1.89E-06
GOCC	nucleus	150	0.28	1.73E-05
	integral to plasma membrane	51	0.10	2.29E-03
	intrinsic to plasma membrane	51	0.10	2.64E-03
	synapse	10	0.02	8.31E-03
	intracellular membrane-bound organelle	181	0.34	2.32E-02
	membrane-bound organelle	181	0.34	2.35E-02
	postsynaptic membrane	7	0.01	2.42E-02
GOMF	protein binding	176	0.33	1.28E-10
	transcription regulator activity	73	0.14	1.92E-09
	RNA polymerase II transcription factor activity	22	0.04	7.60E-08
	binding	341	0.64	2.48E-07
	transcription factor activity	51	0.10	2.43E-06
	DNA binding	90	0.17	5.06E-06
	protein kinase activity	37	0.07	1.35E-05
	phosphotransferase activity, alcohol group as acceptor	40	0.07	4.57E-05
KEGG Pathway	HSA04020:CALCIUM SIGNALING PATHWAY	13	0.02	1.55E-02
	HSA04720:LONG-TERM POTENTIATION	7	0.01	1.90E-02
SP-PIR	nuclear protein	137	0.26	2.32E-17
	transcription	78	0.15	2.61E-13
	transcription regulation	78	0.15	6.96E-13
	phosphorylation	86	0.16	3.62E-11
	dna-binding	74	0.14	6.00E-11
	membrane	133	0.25	1.33E-09
	repressor	20	0.04	4.49E-07
	transferase	57	0.11	8.38E-07
	serine/threonine-protein kinase	25	0.05	1.56E-06
	kinase	38	0.07	1.54E-05
	serine/threonine-specific protein kinase	10	0.02	1.90E-05
	activator	22	0.04	1.90E-05
	nucleotide-binding	54	0.10	3.82E-05

C. ORF - 661 non-redundant genes (68 not in databases)				
Database	Enriched Term	Number of test genes	% of test genes	P-Value
GOBP	transcription	177	0.27	7.67E-27
	regulation of transcription	172	0.26	7.85E-27
	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	173	0.26	1.28E-26
	regulation of cellular metabolism	179	0.27	1.31E-26
	regulation of metabolism	182	0.28	1.50E-26
	regulation of cellular physiological process	211	0.32	3.58E-24
	regulation of biological process	227	0.34	3.93E-24
	regulation of transcription, DNA-dependent	159	0.24	5.98E-24
	transcription, DNA-dependent	162	0.25	6.45E-24
	regulation of physiological process	214	0.32	1.02E-23
	regulation of cellular process	216	0.33	1.59E-23
	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	207	0.31	3.07E-19
GOCC	nucleus	268	0.41	1.20E-29
	intracellular membrane-bound organelle	304	0.46	2.77E-16
	membrane-bound organelle	304	0.46	2.86E-16
	intracellular organelle	336	0.51	1.42E-13
	organelle	336	0.51	1.49E-13
	intracellular	369	0.56	1.66E-09
	nuclear lumen	33	0.05	6.78E-06
GOMF	DNA binding	170	0.26	4.12E-30
	transcription regulator activity	126	0.19	1.24E-28
	nucleic acid binding	217	0.33	1.08E-26
	transcription factor activity	88	0.13	1.02E-18
	sequence-specific DNA binding	56	0.08	1.41E-18
	binding	444	0.67	4.86E-15
	protein binding	219	0.33	5.17E-13
	RNA polymerase II transcription factor activity	29	0.04	8.72E-11
KEGG Pathway	HSA04010:MAPK SIGNALING PATHWAY	19	0.03	2.29E-03

	HSA04520:ADHERENS JUNCTION	8	0.01	1.22E-02
	HSA04930:TYPE II DIABETES MELLITUS	6	0.01	1.39E-02
	HSA04350:TGF-BETA SIGNALING PATHWAY	8	0.01	1.77E-02
	HSA00790:FOLATE BIOSYNTHESIS	5	0.01	2.76E-02
	HSA03030:DNA POLYMERASE	4	0.01	4.22E-02
SP-PIR	nuclear protein	225	0.34	6.56E-49
	dna-binding	138	0.21	4.74E-38
	transcription regulation	122	0.18	1.12E-28
	transcription	120	0.18	2.75E-28
	triplet repeat expansion	17	0.03	6.54E-21
	activator	44	0.07	3.41E-17
	phosphorylation	111	0.17	4.10E-15
	homeobox	32	0.05	5.38E-12
	zinc-finger	81	0.12	1.46E-09
	zinc	91	0.14	8.13E-09
	developmental protein	37	0.06	2.13E-08
	coiled coil	56	0.08	2.87E-08
	DNA binding	34	0.05	1.01E-07

D. 3'UTR - 1289 non-redundant genes (171 not in databases)

Database	Enriched Term	Number of test genes	% of test genes	P-Value
GOBP	regulation of cellular process	326	0.25	2.55E-15
	regulation of biological process	341	0.26	8.10E-15
	development	198	0.15	1.03E-14
	regulation of cellular physiological process	306	0.24	2.95E-13
	regulation of metabolism	246	0.19	5.24E-13
	regulation of physiological process	310	0.24	1.63E-12
	regulation of transcription, DNA-dependent	214	0.17	3.01E-12
	regulation of transcription	224	0.17	6.93E-12
	regulation of cellular metabolism	236	0.18	8.30E-12
	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metab.	226	0.18	8.58E-12

	transcription, DNA-dependent	217	0.17	1.03E-11
	system development	70	0.05	6.56E-11
GOCC	intrinsic to plasma membrane	124	0.10	7.59E-08
	integral to plasma membrane	123	0.10	9.31E-08
	plasma membrane	162	0.13	1.70E-07
	nucleus	298	0.23	1.04E-03
	synaptic vesicle	11	0.01	1.07E-03
	synapse	18	0.01	2.59E-03
	membrane-bound vesicle	21	0.02	3.92E-03
GOMF	protein binding	378	0.29	2.83E-15
	transcription regulator activity	154	0.12	2.94E-15
	transcription factor activity	116	0.09	1.03E-12
	binding	779	0.60	2.08E-12
	sequence-specific DNA binding	61	0.05	1.05E-09
	DNA binding	184	0.14	6.73E-07
	metal ion binding	280	0.22	1.08E-06
	ion binding	280	0.22	1.08E-06
KEGG Pathway	HSA04360:AXON GUIDANCE	27	0.02	5.54E-06
	HSA04510:FOCAL ADHESION	29	0.02	1.55E-03
	HSA04810:REGULATION OF ACTIN CYTOSKELETON	28	0.02	2.27E-03
	HSA04310:WNT SIGNALING PATHWAY	22	0.02	2.40E-03
	HSA04630:JAK-STAT SIGNALING PATHWAY	21	0.02	6.66E-03
	HSA04350:TGF-BETA SIGNALING PATHWAY	13	0.01	1.93E-02
	HSA04010:MAPK SIGNALING PATHWAY	30	0.02	2.43E-02
SP-PIR	transcription	177	0.14	2.27E-27
	transcription regulation	179	0.14	2.84E-27
	phosphorylation	184	0.14	3.12E-18
	membrane	299	0.23	1.59E-16
	dna-binding	147	0.11	4.22E-14
	nuclear protein	250	0.19	4.68E-14
	developmental protein	60	0.05	1.85E-10
	chromosomal translocation	33	0.03	2.23E-10
	transmembrane	268	0.21	2.24E-10
	activator	49	0.04	2.35E-10

	repressor	38	0.03	2.93E-10
	calcium	71	0.06	4.09E-09
	transport	108	0.08	6.33E-08

Supplementary Table 4. Gene classes enriched in triNR- and tetraNR-containing genes associated with human genetic disease/phenotypic traits

Database	Enriched Term	Number of genes	% of genes	P-Value	Fold Enrichment
GOBP	development	25	42	5.64E-09	3.65
	system development	14	24	1.17E-08	7.88
	nervous system development	13	22	1.02E-07	7.38
	synaptic transmission	8	14	2.27E-05	9.11
	transmission of nerve impulse	8	14	3.11E-05	8.67
	cell-cell signaling	10	17	1.28E-04	5.00
	central nervous system development	5	8	9.42E-04	11.24
GOMF	transcription regulator activity	16	27	1.50E-05	3.62
	transcription factor activity	13	22	4.65E-05	4.09
	sequence-specific DNA binding	8	14	4.44E-04	5.65
	protein binding	27	46	5.21E-04	1.85