Novel biological implication of a strictly monomorphic GCC repeat in the human PRKACB gene

Safoura Khamse
Iranian research center on aging, University of social welfare and rehabilitation sciences, Tehran, Iran.
s.khamseh90@yahoo.com

Zahra Jafarian
Iranian research center on aging, University of social welfare and rehabilitation sciences, Tehran, Iran.
jafarian.mahdis@gmail.com

Ali Bozorgmehr
Iran Psychiatric Hospital, Iran University of Medical Sciences, Tehran, Iran.
bozorgmehr.a@iums.ac.ir

Mostafa Tavakoli
Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University, Mashhad, Iran.
m_tavakoli@um.ac.ir

Hossein Afshar
Iranian research center on aging, University of social welfare and rehabilitation sciences, Tehran, Iran.
drhoseinafshar@yahoo.com

Maryam Keshavarz
Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön, Germany.

keshavarz@evolbio.mpg.de

Razieh Moayedi
Department of Economics, Vrije Universiteit Amsterdam, Netherlands.
r.moayedi@student.vu.nl

Mina Ohadi*
Iranian research center on aging, University of social welfare and rehabilitation sciences, Tehran, Iran.

ohadi.mina@yahoo.com
mi.ohadi@uswr.ac.ir

1 Equal contribution
*Corresponding Author
Abbreviations

3D: Three-dimensional
NCD: Neurocognitive disorder
PRKACB: Protein Kinase CAMP-Activated Catalytic Subunit Beta
STR: Short tandem repeat
TSS Transcription start site
UTR Untranslated region
Abstract

Across human protein-coding genes, *PRKACB* (Protein Kinase CAMP-Activated Catalytic Subunit Beta) contains one of the longest GCC-repeats, and is predominantly expressed in the brain. Here we studied this STR in 300 human subjects, consisting of late-onset neurocognitive disorder (NCD) (N=150) and controls (N=150). We also studied the impact of this STR on the three-dimensional structure of DNA. While the *PRKACB* GCC-STR was strictly monomorphic at 7-repeats, we detected two 7/8 genotypes only in the NCD group. In comparison to all other lengths, (GCC)7 had the least effect on the three-dimensional structure of DNA, evidenced by minimal divergence between 0 and 7-repeats (divergence score=0.04) and significant divergence between 0 and 8 repeats (divergence score=0.50). A similar inert effect to the GCC-repeat was not detected in other classes of STRs such as GA and CA repeats. In conclusion, we report monomorphism of an exceptionally long GCC repeat in the *PRKACB* gene in human, its inert effect on DNA structure, and divergence in two cases of late-onset NCD. This is the first indication of natural selection for an exceptionally long monomorphic GCC-repeat, which probably evolved to function as an “epigenetic knob”, without changing the regional DNA structure.

**Keywords:** PRKACB; GCC repeat; monomorphism; neurocognitive disorder; short tandem repeat; epigenetic knob
**Introduction**

Short tandem repeats (STRs) are the most polymorphic genetic elements in the vertebrate genomes. Because of their polymorphic nature and plasticity, these elements provide an efficient source of variation at the inter- and intraspecies levels\textsuperscript{1-6}. Accumulating evidence indicates that certain STRs may be associated with selective advantage in human and other species\textsuperscript{7-10}. Among the exceptionally long STRs spanning the core promoter and 5′ untranslated region (UTR) of human protein-coding genes, the protein kinase cAMP-activated catalytic subunit beta (\textit{PRKACB}) gene contains one of the longest GCC-repeats, at 7-repeats\textsuperscript{5}. Across human tissues, \textit{PRKACB} has its highest level of expression in the brain\textsuperscript{11} (https://www.proteinatlas.org/ENSG00000142875-PRKACB/tissue). Furthermore, in comparison with fourteen non-human primates, the brain expression of this gene reaches highest quantity in human (https://www.ncbi.nlm.nih.gov/ieb/research/assemble/av.cgi?db=human&term=PRKACB&submit=Go). The kinase encoded by \textit{PRKACB} is involved in tau phosphorylation at Alzheimer’s disease (AD)-related sites, and regulation of this gene has been linked to neuroprotective effects against tau and Aβ-induced toxicity\textsuperscript{12,13}.

Here we studied the \textit{PRKACB} gene GCC-repeat in a sample of human subjects, consisting of late-onset neurocognitive disorder (NCD) (also incorrectly known as dementia) and controls, and analyzed its impact on DNA three-dimensional (3D) structure.

**Results**
Predominant monomorphism of the \textit{PRKACB} GCC-repeat in human, at 7-repeats, and divergence from monomorphism in two patients afflicted with late-onset NCD.

We sequenced the \textit{PRKACB} GCC-repeat in 300 human subjects, and found strict monomorphism of this STR at 7-repeats in this human sample (Fig. 1). Exceptions included two 7/8 genotypes in two patients with late-onset NCD (Fig. 2). The two patients harboring the 7/8 genotypes were females of 78 (Patient 1) and 83 (Patient 2) years of age and AMTS of 4 and 3, respectively.

In patient 1 the progression of neurocognitive symptoms was gradual, strengthening the possible diagnosis of AD. In patient 2, symptoms of neurocognitive impairment including aphonia and dementia occurred abruptly and subsequent to a cerebrovascular accident two years before the interview, strengthening the possible diagnosis of vascular dementia. Other causes of late-onset neurological disorders were ruled out in both patients by neurologists.

\textbf{Status of the \textit{PRKACB} GCC-repeat across vertebrates.}

The \textit{PRKACB} GCC-repeat was highly conserved in numerous orders of vertebrates (Table 1). This repeat probably emerged in Birds and Reptiles, as we did not detect it in Amphibians, Fish, and other eukaryotes. In human, the interval encompassing the +1 to +100 to the TSS contains two GCC-repeats of (GCC)3 and (GCC)7 formula (Table 1). We detected a long ancestral trace of GCC-repeats, identifiable by triplet nucleotides of GCC or non-GCC in various species (Fig. 3). This long trace of GCC resulted in complex and unique GCC blocks in every species.

\textbf{In silico reconstruction of the human \textit{PRKACB} 5′ UTR sequence encompassing various GCC repeats.}

DNA reconstruction of the \textit{PRKACB} 5′ UTR encompassing (GCC)7 revealed that (GCC)7 is an inert element that results in the slightest change to the structure of DNA in comparison with
other repeat lengths (Table 2) (Fig. 4). For example, 0 and 7-repeat constructs were at striking similarity (0.04 divergence), whereas the divergence scores between 7 and 8 repeat constructs, genotypes of which were observed in two cases of late-onset NCD, were among the highest scores (0.50), indicating that there was a dramatic change in the 3D structure as a result of 7 vs. 8-repeat (Table 2).

Although we detected only two alleles in the human population studied (7 and 8-repeats), we reconstructed DNAs for various lengths in order to analyze the effect of various repeat lengths on the 3D structure of the region. This analysis revealed that subtraction or addition of 7-repeats resulted in the lowest divergence in the 3D structure in comparison to other repeat lengths (Table 2).

To explore the specificity of the PRKACB GCC repeat inert effect, we performed DNA reconstruction of additional GCC/GGC-repeats in the 5′ UTR of the SMAD9 and RASGEF1C genes, and also a non-GCC STR, such as a GA-repeat in the 5′ UTR of human GPM6B. Strikingly, while the reconstructed 3D structures were significantly divergent at various lengths in the case of the GPM6B 5′ UTR GA-repeat, the SMAD9 and RASGEF1C GCC 3D structures were almost identical for all repeat lengths studied (Suppl. 1). We also studied the interactive effect of divergent non-repeat flanking sequences with the GCC repeats by studying a different species than human, such as capuchin. That analysis showed that the inert effect of the GCC block is dominant to the effect of the flanking non-repeat sequences (Suppl. 1).

**Gene network reconstruction**

Based on the available experimental evidence, the reconstructed interactive network consisted of 16 nodes and 45 edges (Fig. 5). Within this network, PRKACB interacts with genes of other
subunits of the cAMP-dependent protein kinase complex, as well as other genes such as MAPT, FOXO1, and PDE5A, which are central to maintaining cell structure and function.

Discussion

Because of their propensity to epigenetic alterations such as hypermethylation, GCC-repeats are mutation hotspots\textsuperscript{14,15}, and this may be the reason why this class of repeats do not expand beyond certain lengths unless when selected, such as in gene regulatory regions\textsuperscript{16}. Expansions of GCC repeats in the 5′ UTR of a number of genes is associated with hypermethylation and intellectual disability\textsuperscript{17}. Based on the Ensembl database, the PRKACB gene is among the top ten genes in respect of containing the longest GCC-repeat\textsuperscript{5}. Here we show that this GCC-repeat is predominantly monomorphic in human at 7-repeats, and there may be divergence from this monomorphism in late-onset NCD. We propose that there was selective advantage in human at this particular STR to reach to, and stabilize at that length. The above is in line with the observations that mutations that have negative fitness consequences tend to be eliminated from the population\textsuperscript{18}. This STR was found to be highly conserved across various mammalian species, indicating its important role in growth and development. In general, STRs near the TSSs of genes are often highly conserved, and distance from a STR to the nearest TSS is a good predictor of the STR conservation score\textsuperscript{19}.

We also found that the (GCC)\textsubscript{7} had the least effect on the 3D structure in contrast to the more significant effects that various other repeats inserted, which probably confers a modulatory effect to (GCC)\textsubscript{7} without changing the structure of DNA or RNA. Interestingly, the divergence score between the 7 and 8 repeat (the latter was an allele of the two diverged genotypes) was among
the highest. While we also detected an inert effect of GCC/GGC repeats in two additional genes, *SMAD9* and *RASGEF1C*, this effect was not observed in the GA class of STRs, where various lengths of GA-repeats resulted in significant variation in the 3D structure of DNA in the present study and in a previously reported research in the case of human *ZMYM3* (Alizadeh et al. 2019). Significant divergence has also been observed with various CA-repeat lengths in the case of the human *NHLH2* gene. The above findings indicate that exceptionally long GCC/GGC repeats might have evolved to exert epigenetic effects (epigenetic knobs) without changing the structure of the region.

GCC motifs of STRs significantly overlap with G-quadruplex (G4) non-B structures. Recent research indicates that organisms may have evolutionarily developed G4 into a novel reversible and elaborate transcriptional regulatory mechanism benefiting multiple physiological activities of higher organisms.

Future studies are warranted to sequence the *PRKACB* GCC-STR in a large number of patients with neurological disorders. Considering the pivotal role of *PRKACB* in the brain, CRISPR/Cas9 methods may also be ideal to edit this STR at the genomic level and investigate differentiation of human stem cells into neural cells.

In conclusion, we report the first instance of predominant monomorphism of an exceptionally long GCC repeat in human and divergence from this monomorphism in human disease. We also propose that GCC-STRs might have evolved in the genome as regulatory elements (such as epigenetic regulation) without dramatically changing the 3D structure of DNA.

**Methods**
**Subjects**

Three hundred unrelated Iranian subjects, consisting of late-onset NCD patients (N=150) and controls (N=150) were recruited from the provinces of Tehran, Qazvin, and Rasht. In each NCD case, the Persian version of the Abbreviated Mental Test Score (AMTS)\textsuperscript{23,24} was implemented (AMTS of <7 was an inclusion criterion for NCD), medical records were reviewed in all participants, and CT-scans were taken where possible. The AMTS is currently one of the most accurate primary screening instruments to increase the probability of NCD\textsuperscript{25}, and the Persian version of the AMTS is a valid cognitive assessment tool for older Iranian adults, which can be used for NCD screening in Iran\textsuperscript{23}. The control group was selected based on cognitive AMTS of >7, lack of major medical history, and CT-scan where possible. The cases and controls were matched based on age, gender, and residential district. The subjects’ informed consent was obtained (from their guardians where necessary) and their identities remained confidential throughout the study. The research was approved by the Ethics Committee of the University of Social Welfare and Rehabilitation Sciences, Tehran, Iran, and was consistent with the principles outlined in an internationally recognized standard for the ethical conduct of human research. All methods were performed in accordance with the relevant guidelines and regulations.

**Allele and genotype analysis of the PRKACB GCC-repeat.**

Genomic DNA was obtained from peripheral blood using a standard salting out method. PCR reactions for the amplification of the PRKACB GCC-repeat were set up with the following Primers.

**Forward:** CGCCTGCGAAGATACAGTC

**Reverse:** CAACTCACCCTCTCCAC
PCR reactions were carried out using High GC buffer (30% of the final volume per PCR reaction) in a thermocycler (PEQSTAR, model: PEQLAB) under the following conditions: 95 °C for 4 min, 35 cycles of denaturation at 94 °C for 30 s, annealing for 30 s at 58 °C and extension at 72 °C for 40 s, and a final extension of 72 °C for 5 min. All the samples included in this study were sequenced by the forward primer, using an ABI PRISM 377 DNA sequencer.

**Analysis of the PRKACB (GCC)7 across vertebrates.**

The interval between +1 and +100 of the transcription start site (TSS) of the PRKACB was searched across several orders of vertebrates based on the Ensembl Release 102 (https://asia.ensembl.org/index.html). Alignment was performed using CodonCode Aligner 9.0.1.

**In silico DNA reconstruction of various GCC-repeat lengths in human.**

The DNA structure across the PRKACB GCC-STR, SMAD9 GCC-STR, RASGEF1 GGC-STR, and GPM6B GA-STR were reconstructed according to the AA-Wedge model. For each reconstruction, the 100 nucleotide flanking sequences to the repeats were also included. Among the existing DNA predicting models, namely the Crothers, Dickerson, Jernigan, Tung-Harvey, Zhurkin, and the AA-Wedge model, the AA-Wedge model has been reported as the most consistent and accurate. Based on the twist, roll and tilt, this model predicts experimental A-tract curvature as measured by gel retardation and cyclization kinetics. After obtaining the coordinates of the nucleotides in a 3D space, the DNA structure was visualized using plot3D package in R software.

**Divergence calculation across DNA constructs of various GCC repeat lengths.**

Let \((x_i, y_i, z_i)\) for \(i = 1, 2, ..., n\) be coordinates of points for each repeat, in the first step we scaled these coordinates as follows
\[ X_i = \frac{x_i - \min(x_i)}{(x_i) - \min(x_i)} \quad i = 1, 2, ..., n \]

\[ Y_i = \frac{y_i - \min(y_i)}{(y_i) - \min(y_i)} \quad i = 1, 2, ..., n \]

\[ Z_i = \frac{z_i - \min(z_i)}{(z_i) - \min(z_i)} \quad i = 1, 2, ..., n \]

Parametric equations of a segment line that passes through \((X_i, Y_i, Z_i)\)'s were

\[ \{X(t) = \sum_{i=1}^{n-1} (X_i - (t(n - 1) - (i - 1))(X_{i+1} - X_i)l_{(n-1)\frac{i}{n-1} - 1\frac{i}{n-1}}(t) \} \]

\[ Y(t) = \sum_{i=1}^{n-1} (Y_i - (t(n - 1) - (i - 1))(Y_{i+1} - Y_i)l_{(n-1)\frac{i}{n-1} - 1\frac{i}{n-1}}(t) \} \]

\[ Z(t) = \sum_{i=1}^{n-1} (Z_i - (t(n - 1) - (i - 1))(Z_{i+1} - Z_i)l_{(n-1)\frac{i}{n-1} - 1\frac{i}{n-1}}(t) \} \]

where \(l_{[a,b]}(t) = 1 \text{ if } t \in [a, b] \) and \(l_{[a,b]}(t) = 0 \text{ if } t \notin [a, b] \). To measure similarity between two scaled lines, we defined integrated Euclidean distance (IED). Let \((X(t), Y(t), Z(t))\) and \((X'(t), Y'(t), Z'(t))\) for \(t \in [0,1]\) are two lines then IED is as follows

\[ IED = \int_0^1 \sqrt{(X(t) - X'(t))^2 + (Y(t) - Y'(t))^2 + (Z(t) - Z'(t))^2} \, dt. \]

(i,j)-th elements revealed the divergence scores obtained from the above method between the i-th and j-th diagrams.

The accuracy of data was validated by two-by-two comparison of constructs of identical lengths.

**Gene network reconstruction**

The STRING database (https://string-db.org) was used to find the interactions of PRKACB with other genes. STRING is a biological database of known molecular interactions containing information from experimental, computationally predicted, and public text collections. The minimum required interaction score set 0.7 and the maximum number of interactions to show set 20. To ensure the most important and reliable interactions, we selected only data from experimental studies as the active interaction sources. Subsequently, using Cytoscape version 3.8.2 and according to the interactions found, the interactive network was reconstructed.
Author contributions

Safoura Khamse and Zahra Jafarian did the molecular experiments..

Ali Bozorgmehr conceived and performed the in silico modeling of the DNA constructs and wrote part of the manuscript.

Mostafa Tavakoli did the mathematics analyses.

Hossein Afshar collected the human samples and their clinical information.

Maryam Keshavarz and Razieh moayedi provided critical comments and contributed to data collection.

Mina Ohadi conceived and supervised the project, and wrote the manuscript.

Conflict of interests

The authors declare that there is no conflict of interest.

Figure legends

**Fig. 1.** Strict monomorphism of the PRKACB GCC-repeat, at 7-repeats, in human. Only four samples are represented as examples.

**Fig. 2.** Two cases of late-onset NCD with 7/8 genotypes.

**Fig. 3.** Sequence alignment of the PRKACB region encompassing the GCC repeat in various primates. Each species revealed complex and unique stretches of GCC repeats. Contig1 indicates the possible ancestral sequence across the selected species. Triplet nucleotides which may be remnants of GCC sequences are detectable in the ancestral sequence.

**Fig. 4.** 3D construction of human DNA encompassing the PRKACB GCC-repeat. Striking similarity was detected in repeats where subtraction or addition of 7-repeats was included.
Fig. 5. The interactive network reconstructed based on experimental evidence. The network consists of 16 nodes, including genes of the cAMP-dependent protein kinase subunits, as well as other structurally and functionally important genes such as MAPT, FOXO1, and PDE5A. The 16 genes interact with each other through 45 edges.

References

1 Jakubosky, D. et al. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. Nature communications 11, 1-15 (2020).
1. Fotsing, S. F. et al. The impact of short tandem repeat variation on gene expression. *Nature genetics* **51**, 1652-1659 (2019).
2. Press, M. O., Hall, A. N., Morton, E. A. & Queitsch, C. Substitutions are boring: some arguments about parallel mutations and high mutation rates. *Trends in Genetics* **35**, 253-264 (2019).
3. Valipour, E. et al. Polymorphic core promoter GA-repeats alter gene expression of the early embryonic developmental genes. *Gene* **531**, 175-179 (2013).
4. Ohadi, M. et al. Core promoter short tandem repeats as evolutionary switch codes for primate speciation. *American journal of primatology* **77**, 34-43 (2015).
5. Mohammadparast, S., Bayat, H., Biglarian, A. & Ohadi, M. Exceptional expansion and conservation of a CT-repeat complex in the core promoter of PAXBP1 in primates. *American journal of primatology* **76**, 747-756 (2014).
6. Watts, P. C. et al. Stabilizing selection on microsatellite allele length at arginine vasopressin 1a receptor and oxytocin receptor loci. *Proceedings of the Royal Society B: Biological Sciences* **284**, 20171896 (2017).
7. Afshar, H. et al. Evolving evidence on a link between the ZMYM3 exceptionally long GA-STR and human cognition. *Scientific reports* **10**, 1-7 (2020).
8. Afshar, H. et al. Natural selection at the NHLH2 core promoter exceptionally long CA-repeat in human and disease-only genotypes in late-onset neurocognitive disorder. *Gerontology* **66**, 514-522 (2020).
9. Alizadeh, F. et al. Disease-only alleles at the extreme ends of the human ZMYM3 exceptionally long 5′ UTR short tandem repeat in bipolar disorder: a pilot study. *Journal of affective disorders* **251**, 86-90 (2019).
10. London, E. et al. The catalytic subunit β of PKA affects energy balance and catecholaminergic activity. *Journal of the Endocrine Society* **3**, 1062-1078 (2019).
11. Wang, L. et al. MicroRNA-200a-3p mediates neuroprotection in Alzheimer-related deficits and attenuates amyloid-beta overproduction via coregulating BACE1 and PRKACB. *Frontiers in pharmacology* **10**, 806 (2019).
12. Vázquez-Higuera, J. L. et al. Genetic variation in the tau kinases pathway may modify the risk and age at onset of Alzheimer’s disease. *Journal of Alzheimer’s Disease* **27**, 291-297 (2011).
13. Deng, J. et al. Genetic origin of sporadic cases and RNA toxicity in neuronal intranuclear inclusion disease. *Journal of Medical Genetics* (2021).
14. Li, Q. S., Sun, Y. & Wang, T. Epigenome-wide association study of Alzheimer’s disease replicates 22 differentially methylated positions and 30 differentially methylated regions. *Clinical epigenetics* **12**, 1-14 (2020).
15. Sawaya, S. et al. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PloS one* **8**, e54710 (2013).
16. Lesca, G. et al. Clinical, cytogenetic, and molecular description of a FRAXE French family. *Psychiatric genetics* **13**, 43-46 (2003).
17. Gymrek, M., Willems, T., Reich, D. & Erlich, Y. Interpreting short tandem repeat variations in humans using mutational constraint. *Nature genetics* **49**, 1495 (2017).
18. Sawaya, S. M., Lennon, D., Buschiazzo, E., Gemmell, N. & Minin, V. N. Measuring microsatellite conservation in mammalian evolution with a phylogenetic birth–death model. *Genome biology and evolution* **4**, 636-647 (2012).
19. Spiegel, J., Adhikari, S. & Balasubramanian, S. The structure and function of DNA G-quadruplexes. *Trends in Chemistry* **2**, 123-136 (2020).
20. Niu, K. et al. Identification of LARK as a novel and conserved G-quadruplex binding protein in invertebrates and vertebrates. *Nucleic acids research* **47**, 7306-7320 (2019).
Wu, F. et al. Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. *Communications biology* **4**, 1-11 (2021).

Foroughan, M. et al. Validity and reliability of a abbreviated Mental Test Score (AMTS) among older Iranians. *Psychogeriatrics: the official journal of the Japanese Psychogeriatric Society* **17**, 460-465 (2017).

Hodkinson, H. Evaluation of a mental test score for assessment of mental impairment in the elderly. *Age and ageing* **1**, 233-238 (1972).

Carpenter, C. R. et al. Accuracy of Dementia Screening Instruments in Emergency Medicine: A Diagnostic Meta-analysis. *Acad Emerg Med* **26**, 226-245, doi:10.1111/acem.13573 (2019).

Tan, R. K. Z. & Harvey, S. C. A Comparison of Six DNA Bending Models. *Journal of Biomolecular Structure and Dynamics* **5**, 497-512, doi:10.1080/07391102.1987.10506410 (1987).

Goodsell, D. S. & Dickerson, R. E. Bending and curvature calculations in B-DNA. *Nucleic Acids Research* **22**, 5497-5503, doi:10.1093/nar/22.24.5497 (1994).

Soetaert, K. plot3Drgl: Tools for plotting 3-D and 2-D data in openGL.

Szklarczyk, D. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605-D612, doi:10.1093/nar/gkaa1074 (2021).

Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).