Natural Selection at an Exceptionally Long GGC Repeat in the Human RASGEF1C and Divergent Genotypes in Late-onset Neurocognitive Disorder

Z Jafarian
University of Social Welfare and Rehabilitation Sciences
S Khamse
University of Social Welfare and Rehabilitation Sciences
H Afshar
University of Social Welfare and Rehabilitation Sciences
Khorram Khorshid HR
Hope Generation Foundation
A Delbari
University of Social Welfare and Rehabilitation Sciences
M Ohadi (ohadi.mina@yahoo.com)
University of Social Welfare and Rehabilitation Sciences

Research Article

Keywords: RASGEF1C, short tandem repeat, (GGC)-repeat, natural selection, neurocognitive disorder

Posted Date: May 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-517583/v1

Abstract

Across the human protein-coding genes, the neuron-specific gene, *RASGEF1C*, contains the longest (GGC)-repeat, spanning its core promoter and 5' untranslated region (RASGEF1C-201 ENST00000361132.9). *RASGEF1C* expression dysregulation occurs in late-onset neurocognitive disorders (NCDs), such as Alzheimer's disease. Here we sequenced the GGC-repeat in a sample of human subjects (N = 269), consisting of late-onset NCDs (N = 115) and controls (N = 154). We also studied the status of this STR across vertebrates. The 6-repeat allele of this repeat was the predominant allele in the controls (frequency = 0.85) and NCD patients (frequency = 0.78). The NCD genotype compartment consisted of an excess of genotypes that lacked the 6-repeat (Mid-P exact = 0.004). We also detected divergent genotypes that were present in five NCD patients and not in the controls (Mid-P exact = 0.007). This STR expanded beyond 2-repeats specifically in primates, and was at maximum length in human. We conclude that there is natural selection for the 6-repeat allele of the *RASGEF1C* (GGC)-repeat in human, and significant divergence from that allele in late-onset NCDs. Indication of natural selection for predominantly abundant STR alleles and divergent genotypes enhance the perspective of evolutionary biology and disease pathogenesis in human complex disorders.

Introduction

Recent evidence indicates that the evolutionary pattern of a number of short tandem repeats (STRs) in human and other species may be linked to natural selection[1–5]. In human, *RASGEF1C* (RasGEF Domain Family Member 1C) is predominantly expressed in the brain (https://www.proteinatlas.org/ENSG00000146090-RASGEF1C/tissue), and aberrant regulation of this gene occurs in late-onset neurocognitive disorders (NCDs)[6] (also incorrectly known as dementias), such as Alzheimer's disease (AD). Across human protein-coding genes, *RASGEF1C* contains the longest (GGC)-repeat, spanning its core promoter and 5' untranslated region (UTR), at 13-repeats (RASGEF1C-201 ENST00000361132.9)[7]. Based on the Ensembl 103 database (ensemble.org), the transcript containing the (GGC)-repeat is at the highest support level annotated for the transcript isoforms of this gene (TSL:1).

Here we sequenced the *RASGEF1C* (GGC)-repeat in a sample of humans, consisting of late-onset NCDs and controls. We also analyzed the status of this STR across several orders of vertebrates.

Materials And Methods

Subjects

Two hundred sixty nine unrelated Iranian subjects of ≥ 60 years of age, consisting of late-onset NCD patients (n = 115) and controls (n = 154) were recruited from the provinces of Tehran, Qazvin, and Rasht. In each NCD case, the Persian version of the Abbreviated Mental Test Score (AMTS)[8,9] was implemented (AMTS < 7 was an inclusion criterion for NCD), medical records were reviewed in all participants, and CT-scans were taken where possible (approximately 40% of instances). Furthermore, in a number of subjects,

the Mini-Mental State Exam (MMSE) Test[10] was implemented in addition to the AMTS. A score of < 24 was an inclusion criterion for NCD.

The AMTS is currently one of the most accurate primary screening instruments to increase the probability of NCD[11]. The Persian version of the AMTS is a valid cognitive assessment tool for older Iranian adults, and can be used for NCD screening in Iran[8].

The control group was selected based on cognitive AMTS of > 7 and MMSE > 24, lack of major medical history, and normal 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 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 RASGEF1C (GGC)-repeat.

Genomic DNA was obtained from peripheral blood using a standard salting out method. PCR reactions for the amplification of the *RASGEF1C* (GGC)-repeat were set up with the following primers

Forward: GAGGGTGAACTGGGTTTTGG

Reverse: ACTCTAGCGGCTGAAAGAAG

PCR reactions were carried out with a GC-TEMPase 2x master mix (Amplicon) in a thermocycler (Peqlab-PEQStar) under the following conditions: touchdown PCR: 95 ◦C for 5 min, 20 cycles of denaturation at 95 ◦C for 45 s, annealing for 45 s at 67◦C (-0.5 decrease for each cycle) and extension at 72 ◦C for 1 min, and 30cycles of denaturation at 95 ◦C for 40 s, annealing at 57 ◦C for 45 s and extension at 72 ◦C for 1 min, and a final extension at 72 ◦C for 10 min. All samples included in this study were sequenced by the forward primer, using an ABI PRISM 377 DNA sequencer.

Analysis of the RASGEF1C (GGC)-repeat across vertebrates.

The interval between + 1 and + 100 of the TSS of the *RASGEF1C* was searched across all the available orders of vertebrates based on Ensembl 103. The Ensembl alignment program and CodonCode Aligner were implemented for the sequence alignments across the selected species.

Results

Predominant abundance of the RASGEF1C (GGC)6 in human.

We detected six alleles at 5, 6, 7, 8, 9, and 11-repeats, of which the predominant allele was the 6-repeat (Figs. 1 and 2). The frequency of this allele was at 0.85 and 0.78 in the controls and NCD group, respectively (Fig. 2). At significantly lower frequencies, the 8 and 11 repeats were the most abundant in the NCD group and controls, respectively.

Significant enrichment of genotypes that lacked the 6-repeat in the NCD group.

We detected significant enrichment of genotypes that lacked the 6-repeat allele in the NCD group (Mid-P exact = 0.004). Those genotypes consisted of the 7, 8, 9 and 11 repeat alleles (Figs. 3 and 4) (Table 1).

Table 1
NCD patients harboring genotypes that lacked the 6-repeat.

Patient No.	Sex	Age (years)	STR Formula	AMTS	MMSE
1	F	87	7/7*	2	-
2	F	62	7/8*	5	19
3	M	72	7/9*	2	16
4	F	70	8/8	4	-
5	F	77	8/8	4	-
6	F	67	8/8	4	-
7	F	73	8/8	5	-
8	F	63	8/8	5	-
9	M	79	8/8	1	-
10	F	79	8/11*	3	18
11	M	60	8/11*	5	21
	Sex	Age, years	Controls STR Formula	AMTS	MMSE
1	F	78	11/11	9	-
2	M	70	11/11	10	-
3	M	73	8/8	10	-
F = Female M = Male *Disease-only genotypes					

Disease-only genotypes at the RASGEF1C (GGC)n in NCD patients.

Among the genotypes that lacked the 6-repeat allele, we detected genotypes in the patients that were not detected in the control group (hence the term “disease-only”) (Mid-P exact = 0.007) (Table 1) (Fig. 4). Patients harboring those genotypes formed 4% of the NCD group, spanned a wide range of ages between 60 to 78 years, and revealed moderate to severe neurocognitive dysfunction. In line with a higher

frequency of the 8-repeat in the NCD group, we found a significant excess of the 8/8 genotype in this group in comparison to the control group.

Although not statistically significant, we detected one genotype (11/11) in the controls that was lacking in the NCD group. The frequency of the 11-repeat allele was also found to be higher in the controls vs. NCDs.

RASGEF1C (GGC)-repeat expanded specifically in primates and at maximum length in human.

Across all the species studied, the (GGC)-repeat was at maximum length in human. While in primates the minimum repeat length was 4-repeats (Fig. 5), the maximum length of (GGC)-repeat detectable in non-primates was at 2-repeats (Fig. 6), indicating that this STR expanded specifically in primates.

Discussion

GC-rich sequences are frequently subject to hypermethylation and subsequent mutations[12], and therefore identifying lengths of the range detected in the *RASGEF1C* is an exceptional event, unless where selected, such as around the + 1 TSSs[13]. The *RASGEF1C* (GGC)-repeat is the longest identified across human protein-coding genes in the + 1 to + 60 of the TSS interval[7].

We propose that there is natural selection for the 6-repeat in human. This proposition is not only based on the predominant abundance of the 6-repeat allele in the human sample studied, but also the significant enrichment of genotypes lacking this allele in the NCD compartment. Indeed, we detected divergent genotypes that lacked the 6-repeat only in the NCD group. Evidence of natural selection for an abundant allele has been previously reported by our group in the instance of the exceptionally long CA-repeat in the core promoter of the human *NHLH2* gene, and enrichment of genotypes lacking the predominantly abundant allele (21-repeat) in patients afflicted with late-onset NCD[1]. It is commonly assumed that genes influencing health in later life are not subject to natural selection. However, findings on the *APOE* alleles and several other NCD susceptibility loci[14, 15] indicate that natural selection indeed happens on such alleles.

Based on the AceView database[16], in comparison with several primates, the brain expression of *RASGEF1C* has the least quantile expression level in human (https://www.ncbi.nlm.nih.gov/IEB/Research/Acembly), which is intriguingly in line with the observed maximum length of the GGC repeat in human and possible hypermethylation/silencing mechanisms associated with this STR. Indeed, hypermethylation of the *RASGEF1C* 5' UTR has been reported in the cortical region of AD brain[6,17]. However, it should be noted that this GGC repeat is among a number of other regulatory factors which may also affect expression of *RASGEF1C*.

Despite the high prevalence and debilitating characteristics of late-onset NCDs, genetic studies on this group of disorders have resulted in a number of genes with mild to modest effect for the most part[18]. We selected the patients group based on late-onset NCD as an entity, without differentiating the NCD subtypes. The advantage of this approach was to eliminate the often-ambiguous diagnoses made for the NCD subtypes, which frequently co-occur and overlap in respect of the clinical and pathophysiological manifestations[19–21], and are associated with “probable” and “possible” conclusions for the most part (DSM-5). It is commonly assumed that natural selection is unable to hit alleles linked to late-onset NCD. However, evidence on the APOE and several other NCD alleles indicate that such alleles may indeed be subject to natural selection[22, 23].

GGC expansions are strictly linked to neurological disorders of predominant neurocognitive impairment[24–27]. The *RASGEF1C* (GGC) locus provides a prime instance of natural selection at a specific allele of a STR and deviation from this allele linking to a human disease phenotype.

For an accurate genotyping of the *RASGEF1C* (GGC)-repeat, it is warranted that all samples are sequenced by Sanger sequencing. Our data warrant sequencing the *RASGEF1C* GGC repeat in larger sample sizes and various human populations afflicted with major neurological disorders.

Conclusion

We provide a pilot study on repeat length selection at the human *RASGEF1C* exceptionally long GGC repeat, at 6-repeats, and significant enrichment of genotypes lacking this allele in patients with late-onset NCD.

Abbreviations

AD: Alzheimer’s disease

AMTS: Abbreviated Mental Test Score

MMSE: Mini-Mental State Exam

NCD: Neurocognitive disorder

RASGEF1C: RasGEF Domain Family Member 1C

STR: Short tandem repeat

TSS: Transcription start site

UTR: Untranslated region

Declarations

Acknowledgments

We wish to thank all the participants taking part in this research. This research was funded by the University of Social Welfare and Rehabilitation Sciences.

Statement of Ethics

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 responsible ethical committee 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.

Conflict of Interest Statement

The authors have no conflict of interest to declare.

Funding Sources

This research was funded by the University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.

Data Availability

Raw data are available upon request to the corresponding author.

Author Contributions

Z.J. and S.Kh. performed the laboratory experiments. H.S. collected the clinical samples and their information. H.R. KH. and A.D. contributed to data collection and co-ordination. M.O. conceived and supervised the project, and wrote the manuscript.

References

1. Afshar H, Khamse S, Alizadeh F, Delbari A, Najafipour R, Bozorgmehr A, et al. Evolving evidence on a link between the ZMYM3 exceptionally long GA-STR and human cognition. 2045-2322 Contract No: 1 (2020).
2. Alizadeh F, Moharrami T, Mousavi N, Yazarlou F, Bozorgmehr A, Shahsavand E, 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).

3. Flynn JM, Caldas I, Cristescu ME, Clark AG. Selection constrains high rates of tandem repetitive DNA mutation in Daphnia pulex. Genetics. 207(2):697-710 (2017).
4. 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(8):747-56 (2014).
5. Watts PC, Kallio ER, Koskela E, Lonn E, Mappes T, Mokkonen M. Stabilizing selection on microsatellite allele length at arginine vasopressin 1a receptor and oxytocin receptor loci. Proceedings of the Royal Society B: Biological Sciences. 284(1869):20171896 (2017).
6. Li QS, 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):1-14 (2020).
7. Namdar-Aligoodarzi P, Mohammadparast S, Zaker-Kandjani B, Kakroodi ST, Vesiehsari MJ, Ohadi M. Exceptionally long 5' UTR short tandem repeats specifically linked to primates. Gene. 569(1):88-94 (2015).
8. Foroughan M, Wahlund LO, Jafari Z, Rahgozar M, Farahani IG, Rashedi V. Validity and reliability of a bbreviated M ental T Est S core (AMTS) among older I ranian. Psychogeriatrics : the official journal of the Japanese Psychogeriatric Society. 17(6):460-5 (2017).
9. Hodkinson H. Evaluation of a mental test score for assessment of mental impairment in the elderly. Age and ageing. 1(4):233-8 (1972).
10. Carpenter CR, Banerjee J, Keyes D, Eagles D, Schnitker L, Barbic D, et al. Accuracy of Dementia Screening Instruments in Emergency Medicine: A Diagnostic Meta-analysis. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine.26(2):226-45 (2019).
11. Folstein M. A practical method for grading the cognitive state of patients for the children. J Psychiatr res. 12:189-98 (1975).
12. Kiktev DA, Sheng Z, Lobachev KS, Petes TD. GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences. 115(30):E7109-E18 (2018).
13. Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, et al. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PloS one.8(2):e54710 (2013).
14. Drenos F, Kirkwood TB. Selection on alleles affecting human longevity and late-life disease: the example of apolipoprotein E. PLoS One. 5(4):e10022 (2010).
15. Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, Bennett DA, et al. Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. The American Journal of Human Genetics. 90(4):720-6 (2012).
16. Thierry-Mieg D, Thierry-Mieg J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome biology. 7(1):1-14 (2006).

17. Smith R, Pishva E, Shireby G, Smith AR, Roubroeks JA, Hannon E, et al. Meta-analysis of epigenome-wide association studies in Alzheime's disease highlights 220 differentially methylated loci across cortex. BioRxiv. (2020).
18. de Frutos-Lucas J, Frost N, Erickson KI, Serrano JM, Maestu F, Laws SM, et al. Does APOE genotype moderate the relationship between physical activity, brain health and dementia risk? A systematic review. Ageing Research Reviews. 101173 (2020).
19. Karantzoulis S, Galvin JE. Distinguishing Alzheimer's disease from other major forms of dementia. Expert review of neurotherapeutics. 11(11):1579-91 (2011).
20. Lin Y-F, Smith AV, Aspelund T, Betensky RA, Smoller JW, Gudnason V, et al. Genetic overlap between vascular pathologies and Alzheimer's dementia and potential causal mechanisms. Alzheimer's & Dementia. 15(1):65-75 (2019).
21. Noori A, Mezlini AM, Hyman BT, Serrano-Pozo A, Das S. Systematic review and meta-analysis of human Transcriptomics reveals Neuroinflammation, deficient energy metabolism, and Proteostasis failure across Neurodegeneration. Neurobiology of Disease.105225 (2020).
22. Ajjugal Y, Kolimi N, Rathinavelan T. Secondary structural choice of DNA and RNA associated with CGG/CCG trinucleotide repeat expansion rationalizes the RNA misprocessing in FXTAS. Scientific Reports. 11(1):1-17 (2021).
23. Annear DJ, Vandeweyer G, Elinck E, Sanchis-Juan A, French CE, Raymond L, et al. Abundancy of polymorphic CGG repeats in the human genome suggest a broad involvement in neurological disease. Scientific reports. 11(1):1-11(2021).
24. Jiao B, Zhou L, Zhou Y, Weng L, Liao X, Tian Y, et al. Identification of expanded repeats in NOTCH2NLC in neurodegenerative dementias. Neurobiology of aging. 89:142. e1-. e7(2020).
25. LaCroix AJ, Stabley D, Sahraoui R, Adam MP, Mehaffey M, Kernan K, et al. GGC repeat expansion and exon 1 methylation of XYLT1 is a common pathogenic variant in Baratela-Scott syndrome. The American Journal of Human Genetics. 104(1):35-44 (2019).
26. Ma D, Tan YJ, Ng AS, Ong HL, Sim W, Lim WK, et al. Association of NOTCH2NLC repeat expansions with Parkinson disease. JAMA neurology. 77(12):1559-63 (2020).
27. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nature genetics. 51(8):1215-21 (2019).

Figures

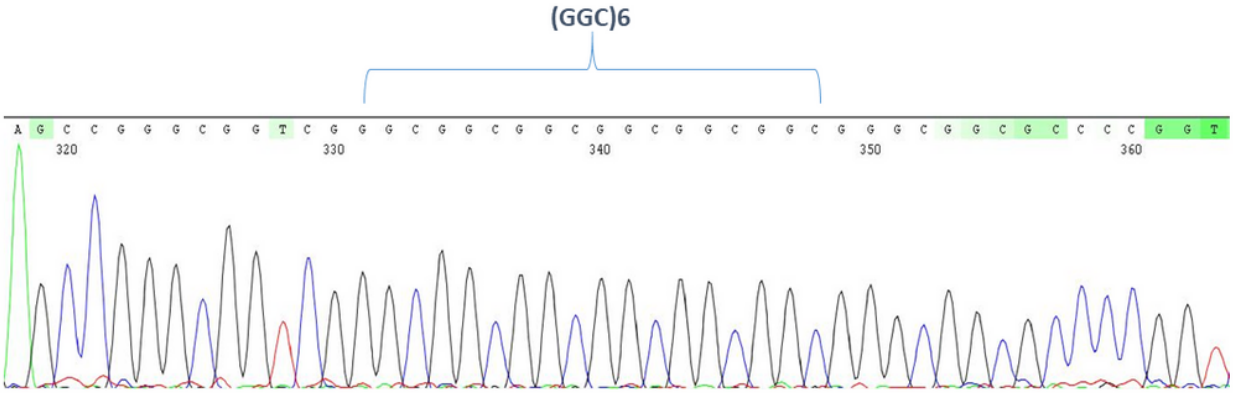

Figure 1

Electropherogram of the predominantly abundant allele at 6-repeats in the human RASGEF1C gene, represented in the 6/6 genotype.

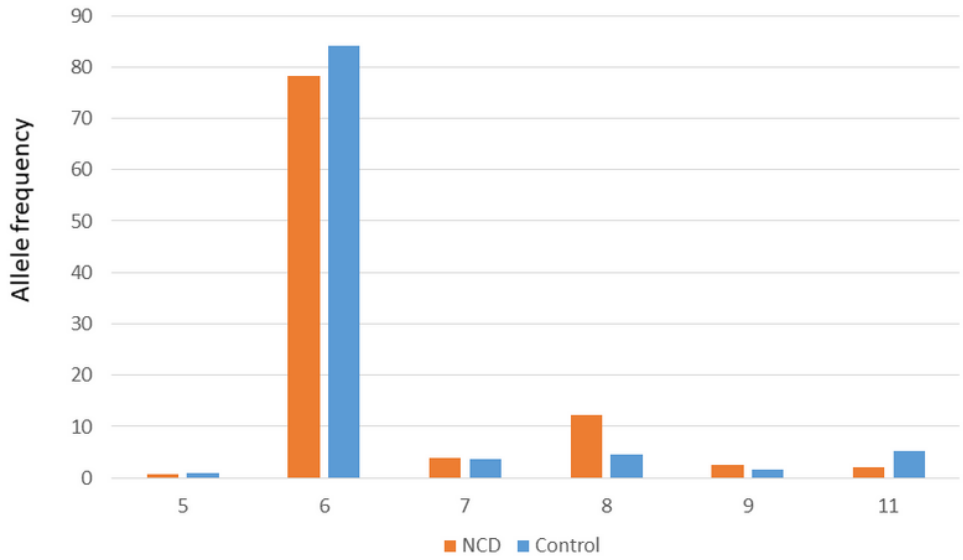

Figure 2

Allele frequency of the RASGEF1C exceptionally long (GGC) repeat in NCD patients and controls.

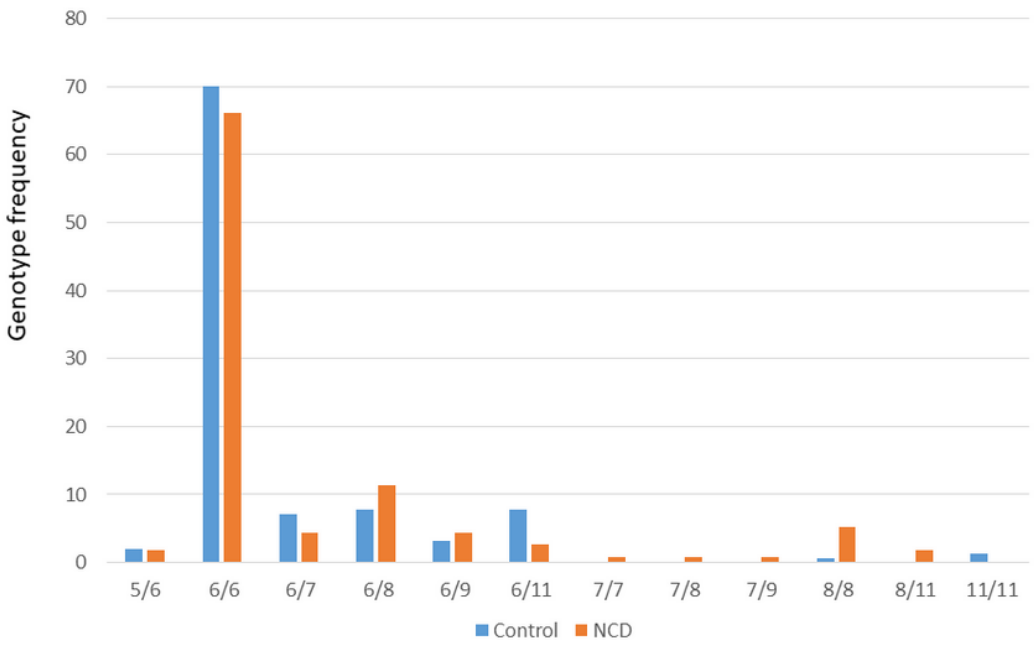

Figure 3

Genotype frequency of the RASGEF1C exceptionally long (GGC) repeat in NCD patients and controls.

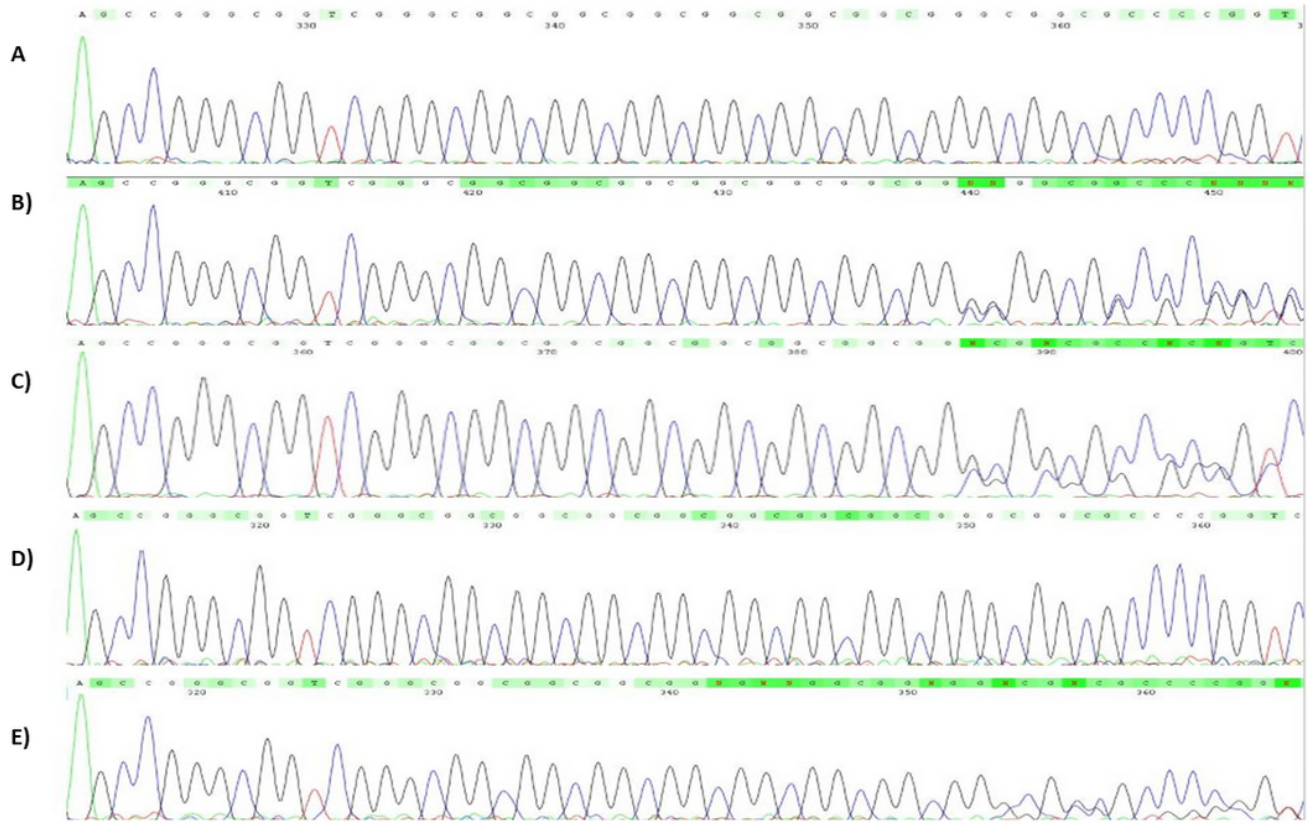

Figure 4

Electropherogram of the Non-6 genotypes at the RASGEF1C (GGC) in NCD patients. A) 7/7, B) 7/8, C) 7/9, D) 8/8, E) 8/11.

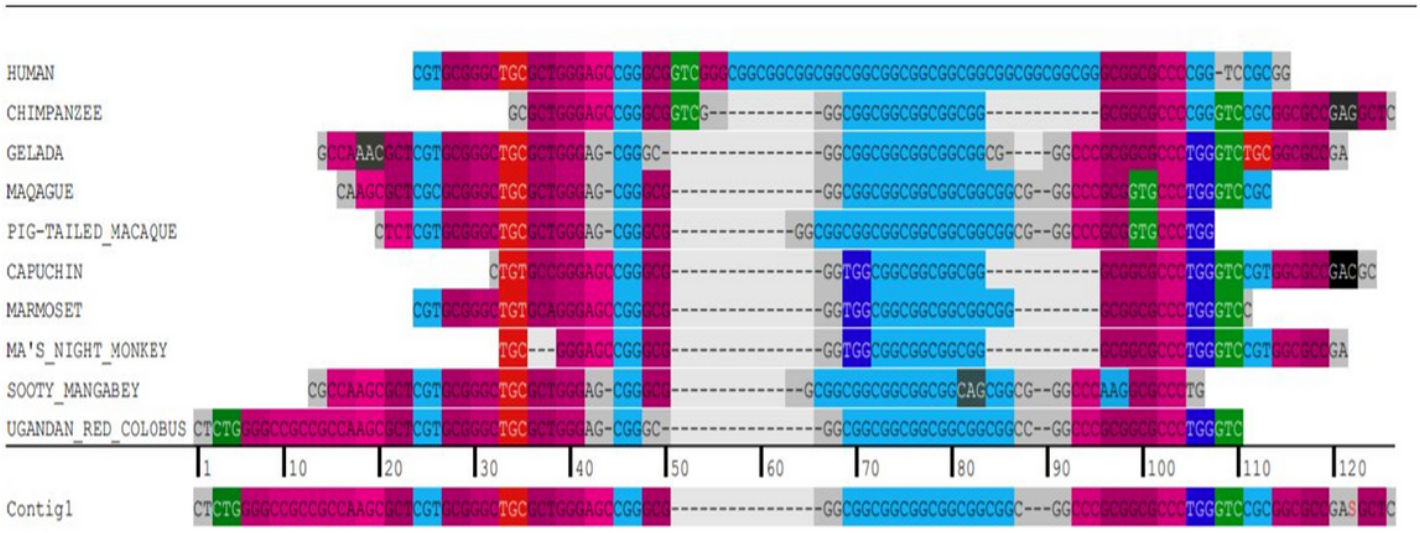

Figure 5

Sequence alignment of the RASGEF1C (GGC)-repeat across selected primate species.

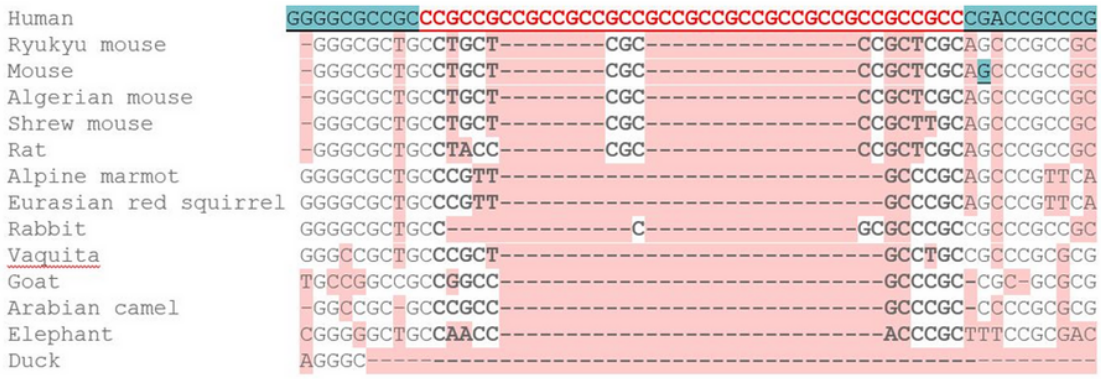

Figure 6

Sequence alignment of the RASGEF1C (GGC)-repeat across non-primates. Human is depicted as the reference sequence. Reverse strand depicted.