EHDN 2021 Remote Meeting, 9-11 September 2021

Protein coding tandem repeat in TCERG1 modifies Huntington's Disease onset

Sergey Lobanov[1], Branduff McAllister[1], Mia McDade-Kumar[1], Jong-Min Lee[2,3,4], Marcy E. MacDonald[2,3,4], James F. Gusella[2,4,5], Mina Ryten[6], Nigel Williams[1], Peter Holmans[1], Thomas Massey[1], Lesley Jones[1]

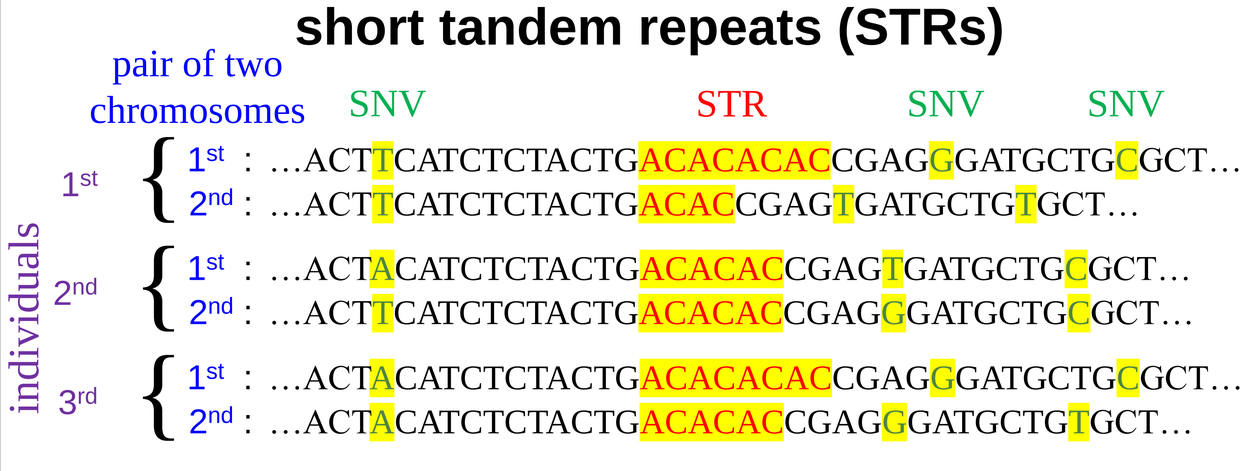
1. Medical Research Council Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, United Kingdom
2. Molecular Neurogenetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston MA 02114, USA
3. Department of Neurology, Harvard Medical School, Boston MA 02115, USA
4. Medical and Population Genetics Program, the Broad Institute of M.I.T. and Harvard, Cambridge MA 02142, USA
5. Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston MA 02115, USA
6. UCL Institute of Neurology, University College London, United Kingdom

bioRxiv 2021.07.16.452643
Email: LobanovS@cardiff.ac.uk
This work was supported by ARUK and CHDI Foundation

Introduction

Huntington's disease (HD) is an inherited neurodegenerative disorder driven by an expanded trinucleotide CAG repeat in exon 1 of the huntingtin gene (*HTT*). The length of the *HTT* CAG repeat explains around 60% of the variance in HD age at onset. The recent GWAS of HD age at onset detected a genome-wide significant association ($p=3.8 \times 10^{-10}$) with an intronic single nucleotide variant (SNV), rs79727797, in *TCERG1* (Cell 178, 887–900, 2019). Since the SNV does not modify the protein or appear to change gene expression, we explored the possibility that the SNV is in linkage disequilibrium (LD) with variants not present in the GWAS data.

Single-nucleotide variants (SNVs) and short tandem repeats (STRs)

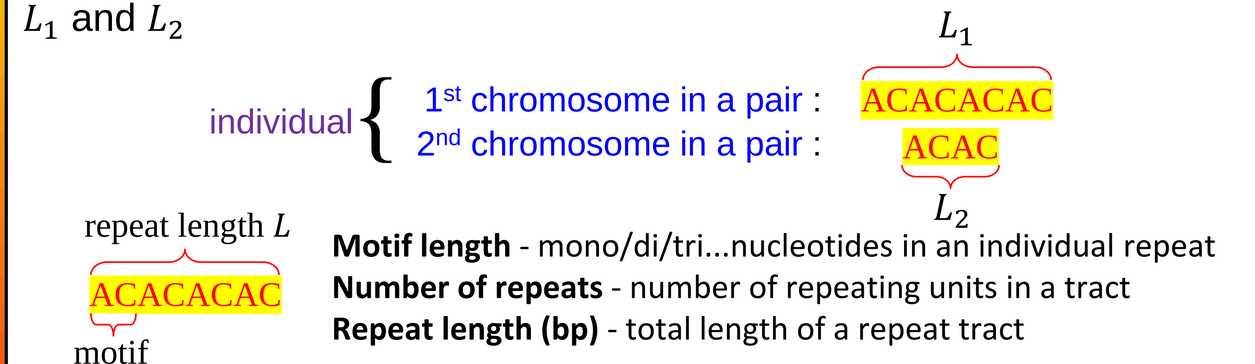


Single-nucleotide variants

Each individual's genome contains roughly 5 million SNVs, and >200 million distinct SNVs have been identified from populations around the world. Each SNV is described by one quantity – number of rare alleles which takes values 0, 1, or 2. Logistic regression analysis can be applied to this quantity in order to test SNV association with disease.

Short tandem repeats

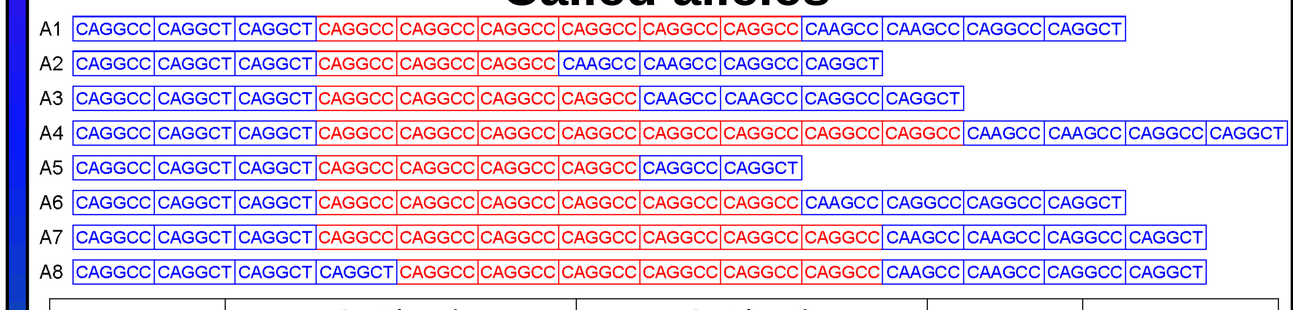
The human genome contains roughly 1.5 million STRs.
Each STR is described by two values – the repeat lengths of two alleles L_1 and L_2



Motif length - mono/di/tri...nucleotides in an individual repeat
Number of repeats - number of repeating units in a tract
Repeat length (bp) - total length of a repeat tract

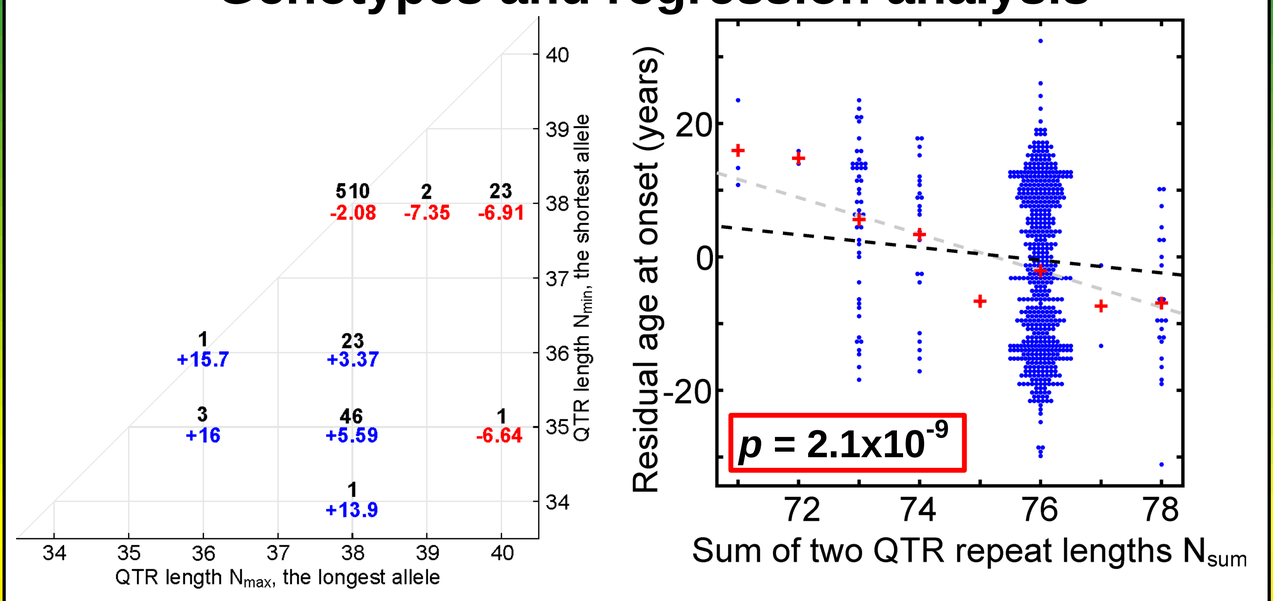
We developed a method for calling perfect and imperfect short tandem repeats from whole exome sequencing (WES) data and applied it to 610 WES samples from HD patients with age at onset discrepant from that predicted by their pure *HTT* CAG length.
To test STR association with disease, we applied linear regression analysis on sum of two repeat lengths $L_{sum} = L_1 + L_2$. We also considered longest, shortest, and difference of two alleles, but found that the sum of the two repeat lengths predicted age at onset significantly better than the other three measures.

Called alleles

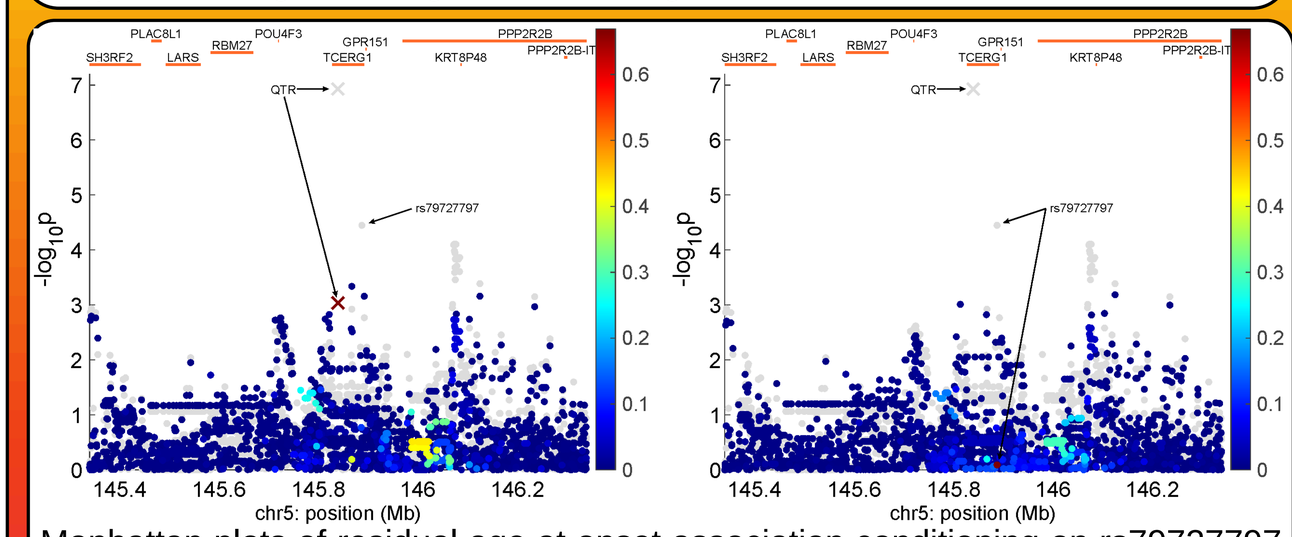


Allele	QTR length		STR length		Number of alleles	Allele frequency (%)
	N	ΔN	N	ΔN		
A1	38	0	6	0	1114	91.31
A2	35	-3	3	-3	50	4.10
A3	36	-2	4	-2	28	2.30
A4	40	2	8	2	24	1.97
A5	34	-4	4	-2	1	0.08
A6	38	0	6	0	1	0.08
A7	39	1	7	1	1	0.08
A8	39	1	6	0	1	0.08

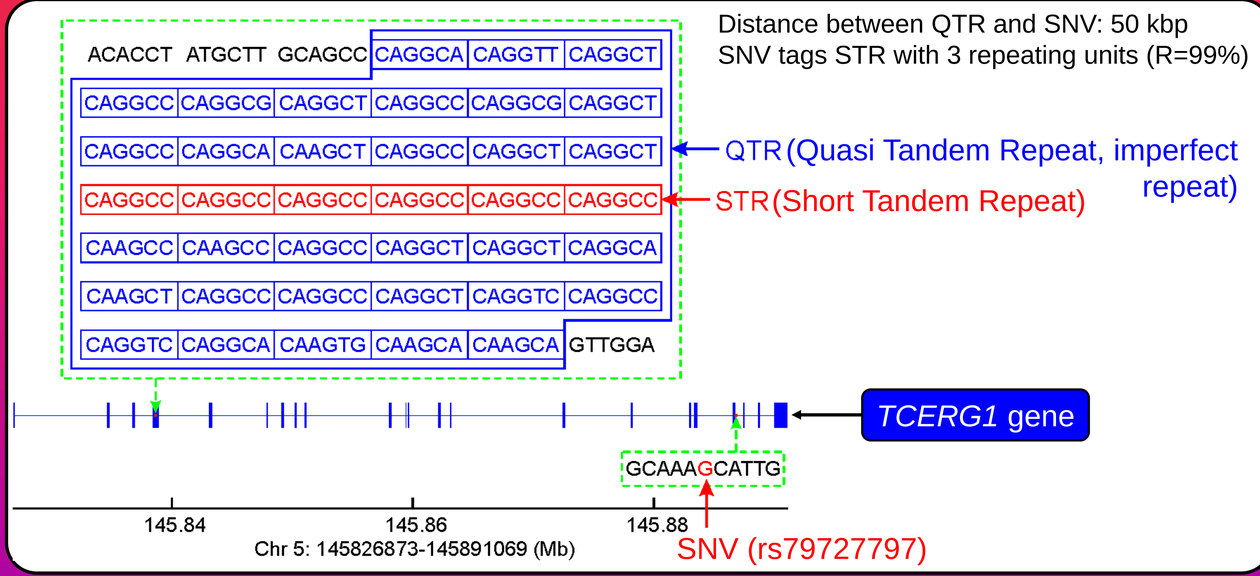
Genotypes and regression analysis



Quasi Tandem Repeat (QTR) genotypes (left panel). Black numbers mark genotype counts. Red (early) and blue (late) numbers indicate mean residual ages at onset for individual genotypes. Association of the sum of two QTR repeat lengths with the residual age at onset (right panel). Red pluses indicate mean residual age at onset for every sum of QTR repeat lengths. Grey and black lines are plotted using linear regression and regression with selection coefficients.



Manhattan plots of residual age at onset association conditioning on rs79727797 (left panel) and QTR (right panel) for 468 HD individuals with both sequencing and GWAS data. The bar on the right of the plots indicates the strength of linkage disequilibrium (r2) between each SNV/QTR and the variant being conditioned on. The grey dots mark *p*-values prior to conditioning. The variant being conditioned on necessarily disappears from the plot.



Conclusions

- Polymorphism of QTR in *TCERG1* is mainly driven by STR located in the center of QTR
- QTR length is in linkage disequilibrium (LD) with the SNV rs79727797 association of which with HD was recently reported
- The association of the sum of the QTR lengths from both alleles with residual age at onset was genome-wide significant ($p=2.1 \times 10^{-9}$)
- *p*-value for the SNV is two orders of magnitude less significant than *p*-value for QTR. *p*-value for SNV becomes non-significant if conditioning on QTR. In contrast, *p*-value for QTR remains significant if conditioning on SNV. This indicates that QTR modifies age at onset of people with Huntington's Disease and SNV only indexes QTR.