Genome-wide survey of tandem repeats by nanopore sequencing shows that disease-associated repeats are more polymorphic in the general population

Satomi Mitsuhashi[1], Martin C Frith[2-4], Naomichi Matsumoto[1]

1. Department of Human Genetics, Yokohama City University Graduate School of Medicine

2. Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST)

3. Graduate School of Frontier Sciences, University of Tokyo

4. Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), AIST

To whom correspondence should be addressed:

Satomi Mitsuhashi, MD, PhD
Department of Human Genetics
Yokohama City University Graduate School of Medicine
Fukuura 3-9, Kanazawa-ku, Yokohama, 236-0004, Japan
Telephone: +81-45-787-2606
Fax: +81-45-786-5219
E-mail: satomits@yokohama-cu.ac.jp

Naomichi Matsumoto, MD, PhD
Department of Human Genetics
Yokohama City University Graduate School of Medicine
Fukuura 3-9, Kanazawa-ku, Yokohama, 236-0004, Japan
Telephone: +81-45-787-2606
Fax: +81-45-786-5219
E-mail: naomat@yokohama-cu.ac.jp

Abstract

Tandem repeats are highly mutable and contribute to the development of human disease by a variety of mechanisms. However, it is difficult to predict which tandem repeats may cause a disease. We performed a genome-wide survey of the millions of human tandem repeats using long read genome sequencing data from 16 humans. We found that known Mendelian disease-causing or disease-associated repeats, especially coding CAG and 5'UTR GGC repeats, are relatively long and polymorphic in the general population. This method, especially if used in GWAS, may indicate possible new candidates of pathogenic or biologically important tandem repeats in human genomes.

Main text

There are more than 30 rare Mendelian human diseases caused by tandem repeat expansions in human genomes [1]. Genome-wide surveys of tandem repeats in individual genomes are now feasible due to the development of high-throughput sequencing technologies, which enable direct identification of large pathogenic expansions. However, it is still difficult to predict which tandem repeats cause disease, because there are thousands of tandem repeats in each individual that are different from the reference genome. We hypothesize that disease-causing tandem repeats are unstable in the general population. It is unclear whether highly polymorphic repeats cause disease, or stable repeats suddenly expand dynamically in gametes to pathogenic lengths (usually pathogenic expansions are +100 to ~10,000 repeat units) [2]. Here, we characterize disease-causing tandem repeats using long DNA reads, by measuring the variation of their length, and comparing them to other repeats. Current long read sequencing technologies have achieved reads longer than 10 kb on average, which have a high chance to cover whole tandem repeats including flanking unique sequences. Our recently developed tool, `tandem-genotypes`, can robustly detect tandem repeat changes from whole genome long read sequencing data [3]. We used this tool to genotype tandem repeats based on long reads from nanopore sequencers.

First, we identified tandem repeats in a human reference genome (hg38) using `tantan` [4] (http://cbrc3.cbrc.jp/~martin/tantan/). In total, 3,312,291 loci were identified, with the repeat units ranging from 1 to 100 bp. We used long read whole genome sequencing from 16 humans who do not have any disease or have diseases explained by chromosomal rearrangements (we suppose they do not have pathogenic tandem repeats), with average coverage of x25 (ranging x12-x39, Table S1). `tandem-genotypes` predicted lengths for more than 98% of the 3 million tandem repeats (Table S1, Figure S1), including 215,561 triplet repeats.

Until recently, most of the known disease-causing tandem repeats are CAG or GGC triplet repeats, although there are a few exceptions; quadruplet repeat (GGCT) in Myotonic Dystrophy type 2 (MIM#602668), and sextuplet repeat (GGGGCC) in Frontotemporal dementia and/or amyotrophic lateral sclerosis (MIM#614260)). CAG and GGC triplet diseases have three major disease mechanisms: poly-glutamine diseases (CAG), poly-alanine diseases (GGC), or 5'UTR GGC expansion diseases [5-7]. We investigated 20 CAG and GGC triplet repeat disease loci (Table S2). Distributions of tandem repeat length in all reads from 16 individuals show that disease-causing repeats have more variation than other repeats (the same number of repeat loci were randomly extracted from non-disease repeat loci for comparison to the disease repeat loci (CAG: n=11, GGC: n=9, AAAAT: n=6)) (Figure 1A, CAG and GGC). This supports our hypothesis that disease-causing tandem repeats are more polymorphic among the normal population. Next we plotted the variation of repeat length (interquartile range (IQR) of repeat-unit count from each read), and mean repeat length, at each exonic locus (including UTR). Shorter-unit repeats are more numerous and more variable (Figure S2). Given that different repeat sequences may have different mutation rates [8], we compared the ten kinds of triplet repeat (Figure S3). Most of the non-disease triplet repeats have little or no length polymorphism. A large fraction (>94% of all repeats) have IQR less than 2, while disease causing tandem repeats usually show more variation (always more than 2) (Figure1B, C). It is of interest that GGC and CAG repeats have more polymorphic loci than other repeat structures (Figure S3).

All 9 disease-causing CAG repeats (Table S2) have IQR >= 2 (median: 4, range: 2-11, Figure 1B). Note that all disease causing CAG repeats, except for *DMPK*, are located in protein-coding regions (Figure 1B, Table S2).

Importantly, among non-disease CAG repeats, we found a non-coding CAG repeat in *TCF4* with high IQR. This repeat was not included in our rare disease list but it has an association with Fuchs endothelial corneal dystrophy (FECD) (MIM#613267). FECD is a commonly seen disease affecting 4-5% of the population older than 40 years [9]. Initially, genome wide association studies (GWAS) showed an association of a SNP (rs613872), but later studies showed this disease has much higher association to a 4.3kb-downstream CAG repeat [10] [11]. This triplet repeat was known to be highly polymorphic [12], in agreement with our result.

Disease-causing GGC repeats are either in protein-coding or 5'-UTR regions. All known protein-coding GGC repeat diseases are caused by poly-alanine expansions. These poly-alanine loci show less variability (IQR=2) than the 5'-UTR GGC loci (IQR=5) (Figure 1C). This may reflect the difference in disease mechanisms of protein-coding and 5'-UTR GGC repeats. It is known that poly-alanine is highly toxic to cells [13] and usually fewer than 10 additional alanine residues are enough to cause disease [2]. This may explain our observation that alanine-coding GGCs are less variable in the general population. In contrast, disease-causing 5'UTR GGCs are more polymorphic. One possible pathomechanism of 5'UTR GGC repeats is gene suppression as seen in fragile X syndrome [5]. Another envisioned mechanism is repeat associated non-AUG translation, which is suspected in the neurological symptoms in patients with *FMR1* premutation (more than 55 GGC repeats). The different mechanisms may show different variation patterns of disease-causing GGC repeats.

In addition to triplet repeats, pathogenic expansions of quintuplet repeat loci (represented as AAAAT in hg38) are associated with myoclonic epilepsies [14-16]. In 2018 and 2019, five AAAAT repeat loci were reported [14-16] in addition to *BEAN1* which causes spinocerebellar ataxia 31 (MIM#117210) [17]. We also observed that disease-linked AAAAT repeat loci have variation in length (Figure 1A, AAAAT). We examined the variation and length of all intronic AAAAT repeat loci in 16 individuals, and found several highly polymorphic AAAAT repeats including disease loci (IQR=4: *SAMD12*, *BEAN1*, *TNRC6A*, Figure 1D). Quintuplet AAAAT repeat expansions are associated with newlydiscovered types of disease, and pathomechanisms of AAAAT repeat expansions are yet unclear. It may be that undiscovered pathogenic repeats for epilepsy are among the other highly polymorphic quintuplet repeats.

We repeated our analysis using repeat annotations from Tandem Repeats Finder (TRF, a.k.a. simpleRepeat.txt) [18]. The proportions of triplet repeat sequences were similar to those from `tantan` (Figure S4A). We analyzed disease-causing CAG and GGC repeats, and observed similar results to tantan-annotated repeats (Figure S4B-D), although the *TCF4* repeat was not annotated by TRF (Figure S4B).

GWAS studies have identified numerous genomic markers over the past fifteen years, however their functional relation to the diseases or traits is usually unclear. It is plausible that tandem repeats near those GWAS markers actually have functional relation to the traits. Our approach found one such example, the *TCF4* repeat for corneal disease, so there may be new candidates among the highly polymorphic repeats (Table 1, Figure 1B). We listed highly polymorphic exonic triplet repeats (IQR>=5) near GWAS signals (<10 kb) from a GWAS catalog [19] (Table 1). It is possible that polymorphic tandem repeats contribute to gene expression variation [20, 21]. A recent study showed that tandem repeats which can alter expression of near-by genes are potential drivers of published GWAS signals. Fotsing *et al.* listed such 1380 tandem repeats as eSTR (repeats associated with the expression of nearby genes) [21], although no Mendelian disease-causing repeats are included in eSTR, possibly because until today most of the known repeat diseases are not caused by altering gene expression levels but by changing protein products. However, there may be more diseases or traits caused by altering gene expression, like Fragile X syndrome. We found an interesting candidate, a 5'UTR GCA repeat in the *GLS* gene, which is highly polymorphic and also listed as an eSTR. Several lines of evidence show that an 8kb-downstream SNP is associated with reticulocyte count (Table S3). *GLS* encodes glutaminase, which catalyzes glutamine conversion to glutamate, has high activity in red blood cells, and plays a role in glutathione metabolism [22] [23]. It is an intriguing possibility that this 5'UTR repeat actually acts as a driver of the GWAS signals and affects erythrocyte maturation by altering the expression of *GLS* thus affecting glutathione metabolism. Another candidate, which does not seem to alter gene expression levels but may alter protein function, is *MMP24*. Three CAG or GGC repeats in *MMP24* are highly variable, which has not been reported previously (Table 1). These encode poly-leucine, poly-alanine (neither are annotated by TRF) and poly-proline tracts (Figure 1E, Figure S5). Two SNPs (rs2425019, rs747202389) 4.5kb and 7.5kb downstream of these repeats, respectively, are reported to be associated with height [24]. *MMP24* encodes a membrane matrix metalloprotease and has roles in embryonic development [25]. It would be interesting to investigate functional consequences of changing these repeats. These speculative examples need further association studies targeting near-by tandem repeats together with functional studies to elucidate the mechanistic relation to the phenotype.

In conclusion, our results indicate that the disease-causing coding CAG repeats, 5'UTR GGC repeats, and intronic AAAAT repeats are long and variable, but alanine-coding GGC repeats are stable (but long) among the 16 individuals. In addition, we detected highly polymorphic tandem repeats that are associated with common disease (i.e. *TCF4* repeats). This suggests that polymorphic tandem repeats may often contribute to common diseases. Our study is limited due to lack of a large number of healthy individuals from multiple ethnicities. Nevertheless, we provide a first example of applying long read sequencing to identify polymorphic tandem repeats. We believe further tandem-repeat surveys using a large number of individuals may provide more insights into human genomes and diseases.

Declarations

Ethics approval and consent to participate

All genomic DNA were examined after obtaining informed consent. Experimental protocols were approved by institutional review board of Yokohama City University under the number of A19080001.

Availability of data and materials

The sequence datasets generated and analyzed in this study are not publicly available.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by AMED under the grant numbers JP19ek0109280, JP19dm0107090, JP19ek0109301, JP19ek0109348, JP18kk020501 (to N. Matsumoto); JSPS KAKENHI under the grant numbers JP17H01539 (to N. Matsumoto) and JP19K07977 (to S. Mitsuhashi); grants from the Ministry of Health, Labor, and Welfare (to N. Matsumoto); and the Takeda Science Foundation (to N. Matsumoto).

Author contributions

SM, MCF, and NM contributed to the conception of the work and acquisition/analysis/interpretation of the data.

Figure legends

Figure1

(A) Variation of tandem repeat length (copy number). x-axis: copy number change relative to the human reference (hg38). y-axis: read count. Three different repeat types (exonic CAG, exonic GGC and intronic AAAAT) are separately shown. Disease repeats: 9 coding CAG repeats, 8 coding and 3 5'UTR GGC repeats, and 6 intronic AAAAT repeats. Other repeat: the same number of repeat loci (CAG: n=11, GGC: n=9, AAAAT: n=6) are randomly extracted from all repeat loci (CAG: n=1814, GGC: n=2907, AAAAT: n=19,665) for comparison to the disease repeats (GGC: n=9, CAG: n=11, AAAAT: n=6).

(B-D) Variation (IQR) and length of repeats with disease-associated sequences. (B) Upper: coding CAG repeats, Bottom: non-coding exonic CAG repeats. (C) Upper: coding GGC repeats, Bottom: non-coding exonic GGC repeats. (D) Intronic AAAAT repeats. Many of the disease-causing repeats have large variation and long repeat size. x-axis: IQR, y-axis: mean repeat length (bp). n provides the numbers of repeat loci. x-axis: IQR, y-axis: read count. (E) N-terminus amino acid sequence of MMP24. There are poly-leucine, poly-alanine and poly-proline tracts (red lines).

Reference

1. Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, Ramakrishnan S, Lavrenko V, Kakaradov B, Hou C, et al: **Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes.** *Am J Hum Genet* 2017, **101:**700-715.
2. Mitsuhashi S, Matsumoto N: **Long-read sequencing for rare human genetic diseases.** *J Hum Genet* 2019.
3. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, Oma Y, Kino Y, Mitsuhashi H, Matsumoto N: **Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads.** *Genome Biol* 2019, **20:**58.
4. Frith MC: **A new repeat-masking method enables specific detection of homologous sequences.** *Nucleic Acids Res* 2011, **39:**e23.
5. Feng Y, Zhang F, Lokey LK, Chastain JL, Lakkis L, Eberhart D, Warren ST: **Translational suppression by trinucleotide repeat expansion at FMR1.** *Science* 1995, **268:**731-734.
6. Amiel J, Trochet D, Clement-Ziza M, Munnich A, Lyonnet S: **Polyalanine expansions in human.** *Hum Mol Genet* 2004, **13 Spec No 2:**R235-243.
7. Adegbuyiro A, Sedighi F, Pilkington AWt, Groover S, Legleiter J: **Proteins Containing Expanded Polyglutamine Tracts and Neurodegenerative Disease.** *Biochemistry* 2017, **56:**1199-1217.
8. Ohshima K, Kang S, Wells RD: **CTG triplet repeats from human hereditary diseases are dominant genetic expansion products in Escherichia coli.** *J Biol Chem* 1996, **271:**1853-1856.
9. Baratz KH, Tosakulwong N, Ryu E, Brown WL, Branham K, Chen W, Tran KD, Schmid-Kubista KE, Heckenlively JR, Swaroop A, et al: **E2-2 protein and Fuchs's corneal dystrophy.** *N Engl J Med* 2010, **363:**1016-1024.
10. Mootha VV, Gong X, Ku HC, Xing C: **Association and familial segregation of CTG18.1 trinucleotide repeat expansion of TCF4 gene in Fuchs' endothelial corneal dystrophy.** *Invest Ophthalmol Vis Sci* 2014, **55:**33-42.
11. Wieben ED, Aleff RA, Tosakulwong N, Butz ML, Highsmith WE, Edwards

AO, Baratz KH: **A common trinucleotide repeat expansion within the transcription factor 4 (TCF4, E2-2) gene predicts Fuchs corneal dystrophy.** *PLoS One* 2012, **7:**e49083.

12. Breschel TS, McInnis MG, Margolis RL, Sirugo G, Corneliussen B, Simpson SG, McMahon FJ, MacKinnon DF, Xu JF, Pleasant N, et al: **A novel, heritable, expanding CTG repeat in an intron of the SEF2-1 gene on chromosome 18q21.1.** *Hum Mol Genet* 1997, **6:**1855-1863.
13. Toriumi K, Oma Y, Kino Y, Futai E, Sasagawa N, Ishiura S: **Expression of polyalanine stretches induces mitochondrial dysfunction.** *J Neurosci Res* 2008, **86:**1529-1537.
14. Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, Toyoshima Y, Kakita A, Takahashi H, Suzuki Y, et al: **Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy.** *Nat Genet* 2018, **50:**581-590.
15. Corbett MA, Kroes T, Veneziano L, Bennett MF, Florian R, Schneider AL, Coppola A, Licchetta L, Franceschetti S, Suppa A, et al: **Intronic ATTTC repeat expansions in STARD7 in familial adult myoclonic epilepsy linked to chromosome 2.** *Nat Commun* 2019, **10:**4920.
16. Florian RT, Kraft F, Leitao E, Kaya S, Klebe S, Magnin E, van Rootselaar AF, Buratti J, Kuhnel T, Schroder C, et al: **Unstable TTTTA/TTTCA expansions in MARCH6 are associated with Familial Adult Myoclonic Epilepsy type 3.** *Nat Commun* 2019, **10:**4919.
17. Sato N, Amino T, Kobayashi K, Asakawa S, Ishiguro T, Tsunemi T, Takahashi M, Matsuura T, Flanigan KM, Iwasaki S, et al: **Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)n.** *Am J Hum Genet* 2009, **85:**544-557.
18. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
19. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al: **The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.** *Nucleic Acids Res*

2019, **47:**D1005-D1012.

20. Bilgin Sonay T, Carvalho T, Robinson MD, Greminger MP, Krutzen M, Comas D, Highnam G, Mittelman D, Sharp A, Marques-Bonet T, Wagner A: **Tandem repeat variation in human and great ape populations and its impact on gene expression divergence.** *Genome Res* 2015, **25:**1591-1599.

21. Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, Goren A, Gymrek M: **The impact of short tandem repeat variation on gene expression.** *Nat Genet* 2019, **51:**1652-1659.

22. Whillier S, Garcia B, Chapman BE, Kuchel PW, Raftos JE: **Glutamine and alpha-ketoglutarate as glutamate sources for glutathione synthesis in human erythrocytes.** *FEBS J* 2011, **278:**3152-3163.

23. Ellory JC, Preston RL, Osotimehin B, Young JD: **Transport of amino acids for glutathione biosynthesis in human and dog red cells.** *Biomed Biochim Acta* 1983, **42:**S48-52.

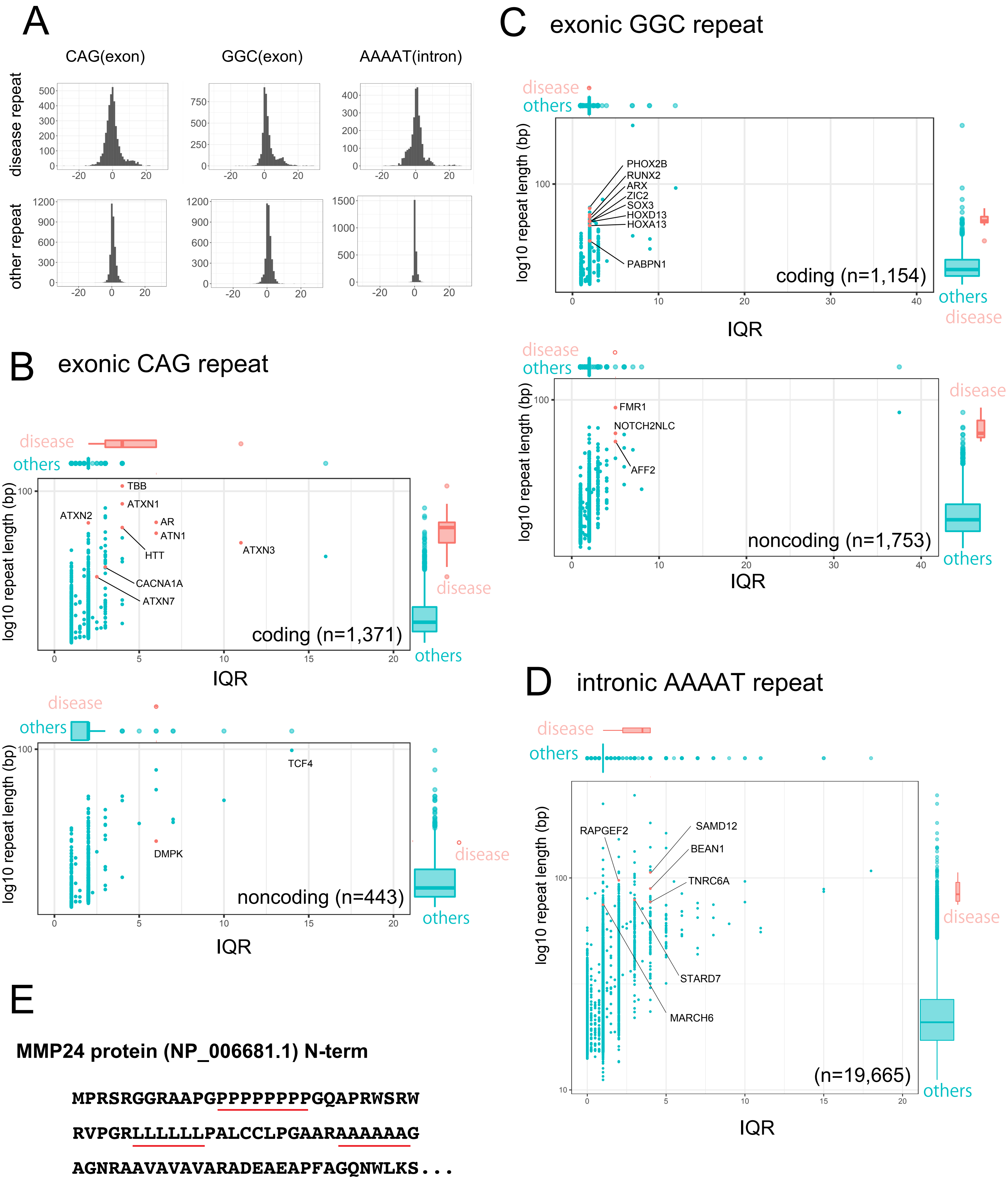
24. Lanktree MB, Guo Y, Murtaza M, Glessner JT, Bailey SD, Onland-Moret NC, Lettre G, Ongen H, Rajagopalan R, Johnson T, et al: **Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height.** *Am J Hum Genet* 2011, **88:**6-18.

25. Pei D: **Identification and characterization of the fifth membrane-type matrix metalloproteinase MT5-MMP.** *J Biol Chem* 1999, **274:**8925-8932.

chromosome	start	end	repeat	gene	annotation	IQR	mean insertion length	GWAS signal < 10kbp	GWAS disease/trait	Mendelian disease (MIM numb
chr22	45240751	45240765	GGC	*KIAA0930*	5'UTR	37.5	84.0	rs2294196,rs5766576	Hair color	
chr20	35226839	35226859	GCT	*MMP24*	coding	16	42.3	rs2425019*,rs747202389,rs11475465	Height, Blood protein levels	
chr18	55586116	55586229	AGC	*TCF4*	5'UTR	14	98.6	rs2924322,rs17598729,rs4144686, rs599550,rs12954356,rs613872*	Corneal dystrophy and others (see Table S3)	
chr20	35226773	35226798	GCC	*MMP24*	coding	12	92.4	rs2425019*,rs747202389,rs11475465		
chr1	77887912	77888014	TTC	*NEXN-AS1*	exon	12	109.6			
chr14	92071010	92071034	CTG	*ATXN3*	coding	11	44.6	rs12588287,rs10143310	Coronary artery calcification, Amyotrophic lateral sclerosis	Machado-Joseph disease (1091
chr3	149766817	149766881	AGC	*ANKUB1*	3'UTR	10	52.3			
chr21	45405396	45405408	GCG	*COL18A1*	coding	9	27.5			
chr17	40818911	40818930	GCC	*KRT10*	coding	9	33.5			
chr6	37507702	37507722	GGT	*LINC02520*	exon	8	77.1	rs3818987	Smoking status	
chr12	64144233	64144254	TTG	*SRGAP1*	3'UTR	8	36.6	rs7131691	Highest math class taken	
chr12	105236161	105236173	GCC	*APPL2*	5'UTR	8	29.0			
chr11	22193318	22193330	GAG	*ANO5*	5'UTR	8	27.4	rs76854597	Creatine kinase levels	
chr6	132758000	132758063	TTC	*VNN2*	5'UTR	7.75	58.4			
chr8	133055824	133055872	CAG	*PTCSC1*	exon	7	39.6	rs2741200	Temperament, Bone erosion in rheumatoid arthritis	
chr20	35226890	35226910	GGC	*MMP24*	coding	7	35.6	rs2425019*,rs747202389,rs11475465	Height, Blood protein levels	
chr2	235516055	235516094	CTG	*TNRC17*	exon	7	41.1			
chr19	45770204	45770266	CAG	*DMPK*	3'UTR	7	32.9			Myotonic dystrophy 1 (160
chr19	14090050	14090075	GCG	*SAMD1*	coding	7	321.2	rs34415768	Height	
chr17	32142451	32142501	CCG	*RHOT1*	5'UTR	7	50.0	rs72483203	Spherical equivalent (joint analysis main effects and education interaction)	
chr7	55887600	55887639	GCG	*ZNF713*	5'UTR	6	39.6	rs11761352	Plasma free amino acid levels	
chr5	177554489	177554531	CGC	*FAM193B*	5'UTR	6	51.4	rs335424,rs62398471,rs61142792	Heel bone mineral density, Mean corpuscular hemog Alzheimer disease and age of onset	
chr3	150703721	150703752	CTC	*ERICH6*	coding	6	40.9			
chr19	10871589	10871651	GCG	*CARM1*	5'UTR	6	61.9	rs12710258	3-hydroxypropylmercapturic acid levels in smokers	
chr16	90102279	90102334	GCA	*FAM157C*	exon	6	59.9	rs9922277	Low tan response	
chr15	23440186	23440198	TCT	*GOLGA6L2*	coding	6	36.1			
chr13	70139351	70139429	CTG	*ATXN8OS*	exon	6	76.9	rs302010	Schizophrenia	Spinocerebellar ataxia 8 (6087
chr1	98046224	98046239	GCC	*MIR137HG*	exon	6	30.9	rs1702294,rs1782810,rs1625579, rs1625579,rs4292998,rs4411173, rs2660304,rs2660302	Schizophrenia, Autism spectrum disorder, Irritable m Heel bone mineral density	
chrX	67545306	67545385	GCA	*AR*	coding	5	75.8			Spinal and bulbar muscular atrophy of Kennedy (3
chrX	148500637	148500684	GCC	*AFF2*	5'UTR	5	58.2			Mental retardation, X-linked, FRAXE type (30
chrX	147912049	147912111	GCG	*FMR1*	5'UTR	5	91.9			Fragile X syndrome (3006
chr9	35906547	35906564	CCA	*HRCT1*	coding	5	27.8	rs76452347,rs748802,rs13297831	Cardiovascular disease, Blood pressure, Resting hea Height	
chr6	24172985	24173021	AAT	*DCDC2*	3'UTR	5	43.1	rs10806984	General cognitive ability	
chr4	37891043	37891089	CCT	*TBC1D1*	5'UTR	5	54.1			
chr3	125013563	125013605	GAG	*HEG1*	coding	5	48.1			
chr2	190880868	190880920	GCA	*GLS*	5'UTR	5	39.0	rs4853525,rs11687659	Reticulocyte fraction of red cells	
chr11	119206289	119206323	CGG	*CBL*	5'UTR	5	44.5	rs4938637,rs7108857,rs36109901	Platelet count	
chr1	179840889	179840917	ATA	*TOR1AIP2*	3'UTR	5	35.5			
chr1	149390802	149390842	GGC	*NOTCH2NLC*	coding	5	62.9			Neuronal intranuclear inclusion disease (60

Table 1. Highly polymorphic triplet repeats (IQR>=5). Repeats with nearby (<10 kb) reported GWAS signals are shown. Repeats whose expansion are known to cause Mendelian disease are also shown. IQR: interquartile range. * not included in GWAS catalog.

Figure 1


A
CAG(exon)
GGC(exon)
AAAAT(intron)
disease repeat
other repeat
500
400
300
200
100
0
750
250
1200
900
600
300
1500
1000
-20
0
20
B
exonic CAG repeat
disease
others
log10 repeat length (bp)
100
TBB
ATXN1
ATXN2
AR
ATN1
HTT
ATXN3
CACNA1A
ATXN7
coding (n=1,371)
IQR
5
10
15
20
TCF4
DMPK
noncoding (n=443)
C
exonic GGC repeat
PHOX2B
RUNX2
ARX
ZIC2
SOX3
HOXD13
HOXA13
PABPN1
coding (n=1,154)
30
40
FMR1
NOTCH2NLC
AFF2
noncoding (n=1,753)
D
intronic AAAAT repeat
RAPGEF2
SAMD12
BEAN1
TNRC6A
STARD7
MARCH6
(n=19,665)
E
MMP24 protein (NP_006681.1) N-term
MPRSRGGRAAPGPPPPPPPPGQAPRWSRW
RVPGRLLLLLLLPALCCLPGAARAAAAAAG
AGNRAAVAVAVARADEAEAPFAGQNWLKS...