Genome-wide Survey of Tandem Repeats by Nanopore Sequencing Shows that Disease-associated Repeats are More Polymorphic in the General Population

Satomi Mitsuhashi (satomits.gfd@mri.tmd.ac.jp)
Medical Research Institute, Tokyo Medical and Dental University https://orcid.org/0000-0002-5036-6858

Martin C Frith
University of Tokyo

Naomichi Matsumoto
Yokohama City University

Research article

Keywords: Nanopore long read sequencing, Tandem repeats, Triplet repeat disease, Genome-wide analysis

Posted Date: October 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-79348/v1

Version of Record: A version of this preprint was published on January 7th, 2021. See the published version at https://doi.org/10.1186/s12920-020-00853-3.

Abstract

Background: Tandem repeats are highly mutable and contribute to the development of human disease by a variety of mechanisms. It is difficult to predict which tandem repeats may cause a disease. One hypothesis is that changeable tandem repeats are the source of genetic diseases, because disease-causing repeats are polymorphic in healthy individuals. However, it is not clear whether disease-causing repeats are more polymorphic than other repeats.

Methods: We performed a genome-wide survey of the millions of human tandem repeats using publicly available long read genome sequencing data from 21 humans. We measured tandem repeat copy number changes using tandem-genotypes. Length variation of known disease-associated repeats was compared to other repeat loci.

Results: We found that known Mendelian disease-causing or disease-associated repeats, especially CAG and 5'UTR GGC repeats, are relatively long and polymorphic in the general population. We also show that repeat lengths of two disease-causing tandem repeats, in *ATXN3* and *GLS*, are correlated with near-by GWAS SNP genotypes.

Conclusions: We provide a catalog of polymorphic tandem repeats across a variety of repeat unit lengths and sequences, from long read sequencing data. This method especially if used in genome wide association study (GWAS), may indicate possible new candidates of pathogenic or biologically important tandem repeats in human genomes.

Background

There are more than 30 rare Mendelian 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 [2–4]. 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. Usually pathogenic expansions are + 100 to ~ 10,000 base-pairs, and the risk cutoff is beyond ~ 100 base-pairs [1, 2]. Some disease-causing repeats are polymorphic even in healthy individuals [5]. If disease-causing tandem repeats have distinct variation in the general population, compared to other repeats, that would help identify novel disease-causing repeat candidates.

Although tandem repeats are highly mutable and can affect phenotype, they are rarely considered in genome-wide association studies (GWAS). GWAS has found many polymorphisms that have significant but weak association with phenotypes, so far failing usually to give satisfying genetic explanations of the phenotypes. As tandem repeats' rapid evolution causes them to have weak association with nearby polymorphisms, we may hypothesize that repeats explain these phenotypes, as represented in previous studies [6, 7].

Current genome-wide studies of tandem repeats using short read sequencers are mainly focusing on short repeats (repeat unit range: 1–6 bp) [8] due to the limitation of detecting long repeats. Current long read sequencing technologies (PacBio and Nanopore) have achieved reads longer than 10 kb on average, which have a high chance to cover whole tandem repeats including flanking unique sequences [9, 10]. However, to the best of our knowledge, there has been no study that characterizes the genotypic variation of disease-causing and other tandem repeats using only long reads.

Until recently, most of the known disease-causing tandem repeats are CAG or GGC triplet repeats [1], although there are a few exceptions; quadruplet repeat (CCTG) in Myotonic Dystrophy type 2 (MIM#602668), and sextuplet repeat (GGGGCC) in Frontotemporal dementia and/or amyotrophic lateral sclerosis (ALS) (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 [11–13]. In addition to triplet repeats, pathogenic expansions of quintuplet repeat loci (represented as AAAAT in hg38) are associated with myoclonic epilepsies. In 2018 and 2019, six AAAAT repeat loci were reported [4, 14–16] in addition to *BEAN1* which causes spinocerebellar ataxia 31 (MIM#117210) [17]. We focus on these triplet and quintuplet repeats so that we can test several disease loci.

Our recently developed tool, tandem-genotypes, can robustly detect tandem repeat changes from whole genome long read sequencing data [18]. Here, we used this tool to measure tandem repeats in publicly available nanopore long read whole genome sequencing data. We show that certain types of disease-causing tandem repeats have greater length variation than other repeats.

Methods

Long read sequencing and mapping to the reference genome

We used 21 long read whole genome sequencing datasets, from 21 humans (Table S1). Fifteen of these are from previous studies [10, 19, 20]. The other six were sequenced by our group, using Nanopore PromethION as previously described [3], with DNA obtained from lymphoblastoid cell lines. Reads were mapped to the human reference genome GRCh38 using LAST according to the instructions (https://github.com/mcfrith/last-rna/blob/master/last-long-reads.md), with repeat-masked reference genome.

```
last-train GRCh38 data.fa > train-out

lastal -p train-out GRCh38 data.fa | last-split > alignment.maf
```

Tandem repeat detection

Tandem repeats in the human reference genome GRCh38 were detected using tantan (http://cbrc3.cbrc.jp/~martin/tantan/) [21], with this command:

tantan -f4 -w2000 GRCh38.fa > tantan-out

Prediction of tandem repeat copy number changes relative to the reference

Tandem-repeat copy number changes relative to the reference were predicted using tandem-genotypes. We used one non-default parameter, n=10 instead of n=60, to make it more specific but less sensitive. This is because the precise boundaries of (inexact) repeats are ambiguous: n=10 makes it less likely to regard an insertion near a repeat as an expansion of the repeat, but more likely to miss expansions of repeats with fuzzy boundaries [11]. Disease-associated tandem repeats were analyzed separately, using the repeat annotations in Table 1.

tandem-genotypes -n10 -g refFlat.txt tantan-out alignnment.maf > out

All tandem-genotypes output files from 21 datasets were merged like this:

tandem-genotypes-join file1 file2 file3... > merged-file

IQR and mean length were calculated from tandem-genotypes output using GNU datamash (https://www.gnu.org/software/datamash/).

Repeat disease selection

We selected triplet-repeat and quintuplet-repeat diseases, because several diseases are known in this category. We took these repeats from a previously published article [1], and recently discovered repeat diseases were added by manual literature search.

Phasing the repeat and near-by GWAS SNP

Phasing of a disease-associated (*ATXN3* or *GLS*) tandem-repeat and nearby GWAS SNP (<10 kb) [22] was done from consensus sequences of the DNA reads. Briefly, a repeat's copy number in each of the two alleles was estimated by tandem-genotypes, then the reads from the two alleles were merged into two consensus sequences, and re-aligned to the reference genome. tandem-genotypes-merge merges these reads using lamassemble [16]:

tandem-genotypes -o2 -v repeat-locus alignnment.maf > out

tandem-genotypes-merge reads.fa train-out out > merged.fa

Results

We identified tandem repeats in a human reference genome (GRCh38) using tantan [23] (http://cbrc3.cbrc.jp/~martin/tantan/). In total, 3,347,418 loci were identified, with the repeat units ranging from 1 to 2000 bp. We used 21 publicly available long read whole genome sequencing datasets (we suppose they do not have pathogenic tandem repeat expansions), with average coverage of 27x (ranging 8x-48x, Table S1). tandem-genotypes predicted lengths for more than 98% of the 3 million tandem repeats (Table S1), including 215,561 triplet repeats.

We investigated 12 CAG and 14 GGC triplet repeat and 7 AAATA quintuplet repeat disease loci (Table 1), and plotted the distribution of copy number changes from the reference in all the reads. We found that disease-causing repeats show different distribution from other non-disease repeats (Supplementary Fig S1A-C). We randomly extracted the same number of non-disease repeat loci for comparison to the disease repeat loci (CAG: n = 12, GGC: n = 14, AAAAT: n = 7)) (Supplementary Figure S1). This supports our hypothesis that disease-causing tandem repeats are more polymorphic among the normal population than other loci.

Given that different repeat sequences may have different mutation rates [24], we compared the ten kinds of non-disease triplet repeats (All triplet repeats can be categorized into 10 kinds. Note that AAC repeats includes AAC, ACA, CAA, GTT, TGT, TTG repeats) (Supplementary Figure S2). 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). Most of the non-disease triplet repeats have little or no length polymorphism. A large fraction (> 94% of all repeats) have IQR 2 or less, while disease causing tandem repeats usually show more variation (always more than 2) (Table 1). It is of interest that GGC and CAG repeats have more polymorphic loci than other repeat structures (Supplementary Figure S2). In addition, shorter-unit repeats are more numerous and more variable (Supplementary Figure S3). Therefore, we analyzed the variation (IQR) and repeat length for disease causing repeats in comparison to other repeats considering the repeat unit and repeat location.

Disease-associating CAG repeats are longer and more variable than most other CAG repeats (Fig. 1A, B, Table 1). We showed coding and non-coding repeats separately (A: coding, B: non-coding). All disease-causing CAG repeats are located in protein-coding regions except for *DMPK*, *GLS*, and *TCF4* which are in 5'UTR (Table 1). Next we tested GGC repeats. Disease-causing 5'-UTR GGC loci are long and variable (Fig. 2B) but protein-coding regions are long but show less variability (Fig. 2A). Gene names were used to indicate the disease-causing repeats because the pathogenic repeats are present only once in each gene. All known protein-coding GGC repeat diseases are located at poly-alanine tracts. This may reflect the difference in disease mechanisms of protein-coding versus 5'-UTR GGC repeats or protein-coding GGC versus CAG repeats. Next, we examined the variation and length of all intronic AAAAT repeat loci in 21 individuals, and found several highly polymorphic AAAAT repeats including disease loci (Fig. 3, Table 1).

We repeated our analysis using repeat annotations from Tandem Repeats Finder (TRF, a.k.a. simpleRepeat.txt) [25]. TRF annotates fewer repeats than tantan (Supplementary Figure S4A), however,

the proportion of triplet repeat sequences is similar (Supplementary Figure S4B). Numbers of intersections between these annotations were calculated using bedtools v2.27.1 (Supplementary Table S2). We analyzed disease-associated CAG and GGC repeats, and observed similar results to tantan-annotated repeats (Supplementary Figure S5: CAG, S6: GGC, S7: AAAAT).

Next, we tested if polymorphic disease-associated tandem repeats are correlated with reported GWAS SNPs. We tested *ATXN3* and *GLS* disease-associated repeats because they are highly polymorphic among disease-associated CAG repeats. These repeats have two (rs12588287: coronary artery calcification [26], rs10143310: ALS [27]) and one (rs4853525: reticulocyte count [28]) near-by GWAS SNPs (< 10 kb) [22], respectively. Due to the limited coverage and read length, we could obtain genotypes in most but not all of the 21 cases (Supplementary Table S3). In each case, one of the two SNP alleles is significantly ($p < 0.05$, unpaired t-test) associated with longer repeats (Supplementary Figure S8). Risk alleles tend to occur with shorter repeats for two SNPs: rs4853525-C and rs12588287-T. Risk allele for rs10143310 is not available [27]. This merits further investigation by genotyping a larger number of individuals.

Finally, we listed highly polymorphic repeats (IQR > = 5) which have very near GWAS signals (< 100 bp) from a GWAS catalog [22] (Table S4). We found an interesting candidate, an intronic repeat in the *CLN8* gene: a SNP within this repeat (rs11986414) and a near-by SNP (rs4875960) are reported to be associated with severity of Gaucher syndrome [29]. It is an intriguing possibility that this repeat genuinely acts as a driver of the GWAS signals and affects the disease severity. We found that the A genotypes of these two SNPs are correlated with shorter repeat (Supplementary Figure S9). 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.

Discussion

We showed that CAG, non-coding GGC and intronic AAAAT disease-associated tandem-repeats are polymorphic and long compared to other repeats using whole genome long read sequencing data. However, coding GGC repeats did not show such variability, although the repeat lengths were longer than other repeats. It is known that poly-alanine is toxic to cells [30] 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-associated 5'UTR GGCs are more polymorphic. One possible pathomechanism of 5'UTR GGC repeats is gene suppression as seen in fragile X syndrome [11]. 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) [31]. The different mechanisms may reflect different variation patterns of disease-causing GGC repeats. Quintuplet AAAAT repeat loci are associated with newly-discovered types of disease, and pathomechanisms of AAAAT repeat expansions are yet unclear [15]. We also showed that there are several highly polymorphic AAAAT repeats which may be undiscovered pathogenic repeats for epilepsy.

GWAS 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. Interestingly, some repeat expansion disease loci may be associated with multiple diseases or traits, even when the repeat length is within the normal range [32, 33]. It is reported that polymorphic tandem repeats contribute to gene expression variation [34]. A recent study showed that tandem repeats which can alter expression of near-by genes are potential drivers of published GWAS signals [35]. Fotsing *et al.* listed 1380 such tandem repeats as eSTR (repeats associated with the expression of nearby genes), although no Mendelian disease-causing repeats are included in eSTR, possibly because most of the known repeat diseases may not be caused by altering gene expression levels but by changing protein products. However, there may be other diseases or traits caused by altering gene expression, like Fragile X syndrome.

Importantly, among disease associated CAG repeats, the noncoding repeat in *TCF4* has high IQR. This triplet repeat was known to be highly polymorphic [36], in agreement with our result. This repeat has an association with Fuchs endothelial corneal dystrophy (FECD) (MIM#613267) [37]. Initially, GWAS showed an association of a SNP (rs613872), but later studies showed this disease has much higher association to a 43 kb-downstream CAG repeat which is in linkage disequilibrium with the GWAS SNP [6, 7]. It is intriguing to consider that further studies on polymorphic repeats may lead to the discovery of true pathogenic variants from GWAS SNPs. However, it is reported that tandem repeats with multiple genotypes are poorly tagged with SNPs [38]. Nevertheless, some repeat expansion diseases are known to be linked to certain haplotypes [39, 40], although there are repeat expansions that do not share haplotype or occur *de novo* [41]. We showed some examples in this study. The first example is a 5'UTR GCA repeat in the *GLS* gene, which is highly polymorphic and also listed as an eSTR [35]. Expansions (> ~ 680 repeats) are known to cause deficiency of GLS and linked to neurological disease [42]. Several lines of evidence show that an 8 kb-downstream SNP is associated with reticulocyte count (Supplementary Table S3). We showed that this SNP is correlated with repeat length. *GLS* encodes glutaminase, which catalyzes glutamine conversion to glutamate, has high activity in red blood cells (erythrocytes), and plays a role in glutathione metabolism [43] [44]. There is an intriguing possibility that this 5'UTR repeat actually acts as a driver of the GWAS signal and affects reticulocyte-erythrocyte maturation by altering the expression of *GLS* thus affecting glutathione metabolism. The next example is *ATXN3*. We found two near-by GWAS SNPs, including one associated with ALS, are significantly correlated with repeat length. Since another spinocerebellar ataxia repeat in *ATXN2* is associated with ALS, this locus is of interest. A final example is the Gaucher disease severity associated SNPs in and near the polymorphic repeat in an intron of *CLN8*. These speculative examples need further association studies targeting near-by tandem repeats together with functional studies to elucidate the mechanistic relation to the phenotype.

Conclusion

In conclusion, our results indicate that known disease-associated 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 21 individuals. 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.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by AMED under the grant numbers JP20ek0109486, JP20dm0107090, JP20ek0109301, JP20ek0109348, JP20kk0205012 (to N. Matsumoto); JSPS KAKENHI under the grant numbers JP17H01539 (to N. Matsumoto) and JP19K07977 and 16H06279 (PAGS) (to S. Mitsuhashi); intramural grants of NCNP from the Ministry of Health, Labor, and Welfare (30-6 and 30-7) (to N. Matsumoto); and the Takeda Science Foundation (to N. Matsumoto).

Availability of data and materials

PromethION WGS sequence data is available from DDBJ (DRA009852). Other public data were downloaded from NCBI or Human PanGenome Project (https://github.com/human-pangenomics/hpgp-data) under accession numbers described in Table S1.

Author contributions

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

Acknowledgement

Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Web resources

tantan (http://cbrc3.cbrc.jp/~martin/tantan/)

tandem-genotypes (https://github.com/mcfrith/tandem-genotypes)

lamassemble (https://gitlab.com/mcfrith/lamassemble/blob/master/lamassemble)

bedtools (https://bedtools.readthedocs.io/en/latest/)

Extended Data

IQR data for all the repeat loci except homopolymers.

References

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* 2020, **65:**11-19.
3. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, Koike H, Hashiguchi A, Takashima H, Sugiyama H, et al: **Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease.** *Nat Genet* 2019, **51:**1215-1221.
4. 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.
5. McMurray CT: **Mechanisms of trinucleotide repeat instability during human development.** *Nat Rev Genet* 2010, **11:**786-799.
6. 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.
7. 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.

8. Gymrek M: **A genomic view of short tandem repeats.** *Curr Opin Genet Dev* 2017, **44:**9-16.
9. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al: **Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.** *Nat Biotechnol* 2019, **37:**1155-1162.
10. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al: **Nanopore sequencing and assembly of a human genome with ultra-long reads.** *Nat Biotechnol* 2018, **36:**338-345.
11. 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.
12. 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.
13. Adegbuyiro A, Sedighi F, Pilkington AWt, Groover S, Legleiter J: **Proteins Containing Expanded Polyglutamine Tracts and Neurodegenerative Disease.** *Biochemistry* 2017, **56:**1199-1217.
14. 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.
15. 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.
16. Yeetong P, Pongpanich M, Srichomthong C, Assawapitaksakul A, Shotelersuk V, Tantirukdham N, Chunharas C, Suphapeetiporn K, Shotelersuk V: **TTTCA repeat insertions in an intron of YEATS2 in benign adult familial myoclonic epilepsy type 4.** *Brain* 2019, **142:**3360-3366.
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. 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.
19. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al: **Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes.** *Nature Biotechnology* 2020.
20. De Coster W, De Rijk P, De Roeck A, De Pooter T, D'Hert S, Strazisar M, Sleegers K, Van Broeckhoven C: **Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome.** *Genome Res* 2019, **29:**1178-1187.
21. Frith MC: **Gentle masking of low-complexity sequences improves homology search.** *PLoS One* 2011, **6:**e28819.
22. 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.

23. Frith MC: **A new repeat-masking method enables specific detection of homologous sequences.** *Nucleic Acids Res* 2011, **39:**e23.
24. 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.
25. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
26. Wojczynski MK, Li M, Bielak LF, Kerr KF, Reiner AP, Wong ND, Yanek LR, Qu L, White CC, Lange LA, et al: **Genetics of coronary artery calcification among African Americans, a meta-analysis.** *BMC Med Genet* 2013, **14:**75.
27. Nicolas A, Kenna KP, Renton AE, Ticozzi N, Faghri F, Chia R, Dominov JA, Kenna BJ, Nalls MA, Keagle P, et al: **Genome-wide Analyses Identify KIF5A as a Novel ALS Gene.** *Neuron* 2018, **97:**1268-1283 e1266.
28. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-Mckay F, Kostadima MA, et al: **The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease.** *Cell* 2016, **167:**1415-1429 e1419.
29. Zhang CK, Stein PB, Liu J, Wang Z, Yang R, Cho JH, Gregersen PK, Aerts JM, Zhao H, Pastores GM, Mistry PK: **Genome-wide association study of N370S homozygous Gaucher disease reveals the candidacy of CLN8 gene as a genetic modifier contributing to extreme phenotypic variation.** *Am J Hematol* 2012, **87:**377-383.
30. 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.
31. Hagerman PJ, Hagerman RJ: **Fragile X-associated tremor/ataxia syndrome.** *Ann N Y Acad Sci* 2015, **1338:**58-70.
32. Lee JK, Conrad A, Epping E, Mathews K, Magnotta V, Dawson JD, Nopoulos P: **Effect of Trinucleotide Repeats in the Huntington's Gene on Intelligence.** *EBioMedicine* 2018, **31:**47-53.
33. Neuenschwander AG, Thai KK, Figueroa KP, Pulst SM: **Amyotrophic lateral sclerosis risk for spinocerebellar ataxia type 2 ATXN2 CAG repeat alleles: a meta-analysis.** *JAMA Neurol* 2014, **71:**1529-1534.
34. 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.
35. 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.
36. 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.

37. 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.
38. Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, Joshi RS, Mittelman D, Sharp AJ: **Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans.** *Nucleic Acids Res* 2016, **44:**3750-3762.
39. Majounie E, Renton AE, Mok K, Dopper EG, Waite A, Rollinson S, Chio A, Restagno G, Nicolaou N, Simon-Sanchez J, et al: **Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study.** *Lancet Neurol* 2012, **11:**323-330.
40. Lee JM, Kim KH, Shin A, Chao MJ, Abu Elneel K, Gillis T, Mysore JS, Kaye JA, Zahed H, Kratter IH, et al: **Sequence-Level Analysis of the Major European Huntington Disease Haplotype.** *Am J Hum Genet* 2015, **97:**435-444.
41. Doi H, Okubo M, Fukai R, Fujita A, Mitsuhashi S, Takahashi K, Kunii M, Tada M, Fukuda H, Mizuguchi T, et al: **Reply to "GGC Repeat Expansion of NOTCH2NLC is Rare in European Leukoencephalopathy".** *Ann Neurol* 2020.
42. Rumping L, Jans JJ, van Hasselt PM: **Glutaminase Deficiency Caused by Short Tandem Repeat Expansion in GLS.** *N Engl J Med* 2019, **381:**1185.
43. 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.
44. 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.

Table

Due to technical limitations, table 1 is only available as a download in the Supplemental Files section.

Figures

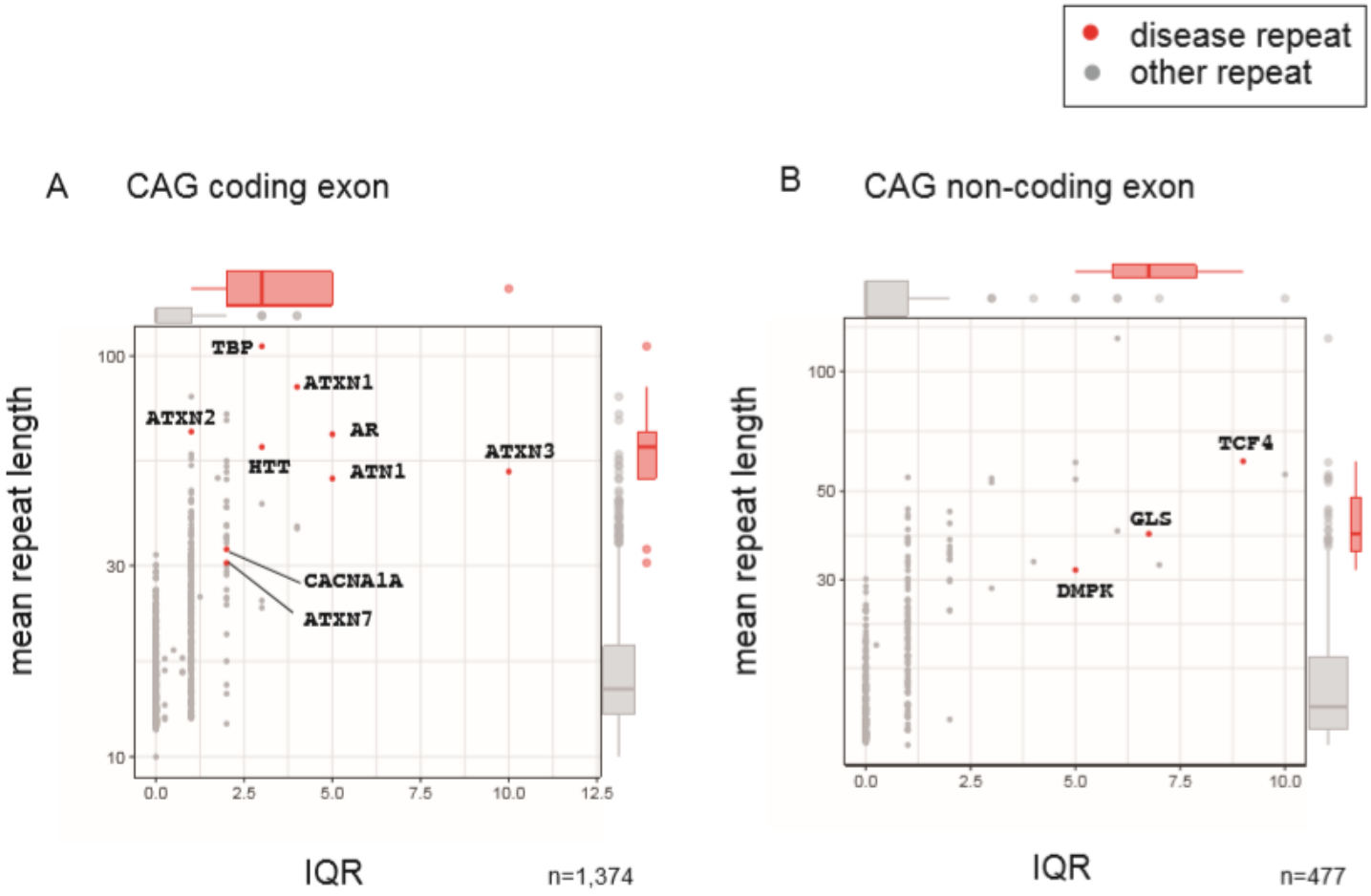

Figure 1

Variation (IQR, interquartile range) and length of repeats with disease-associated sequences. coding CAG repeats (A), and non-coding exonic CAG repeats (B). x-axis: IQR, y-axis: mean repeat length (bp). n provides the numbers of repeat loci. In merged boxplots on the right and upper, ranges are the 25th and 75th percentiles, dots are outliers and lines in boxes are median.

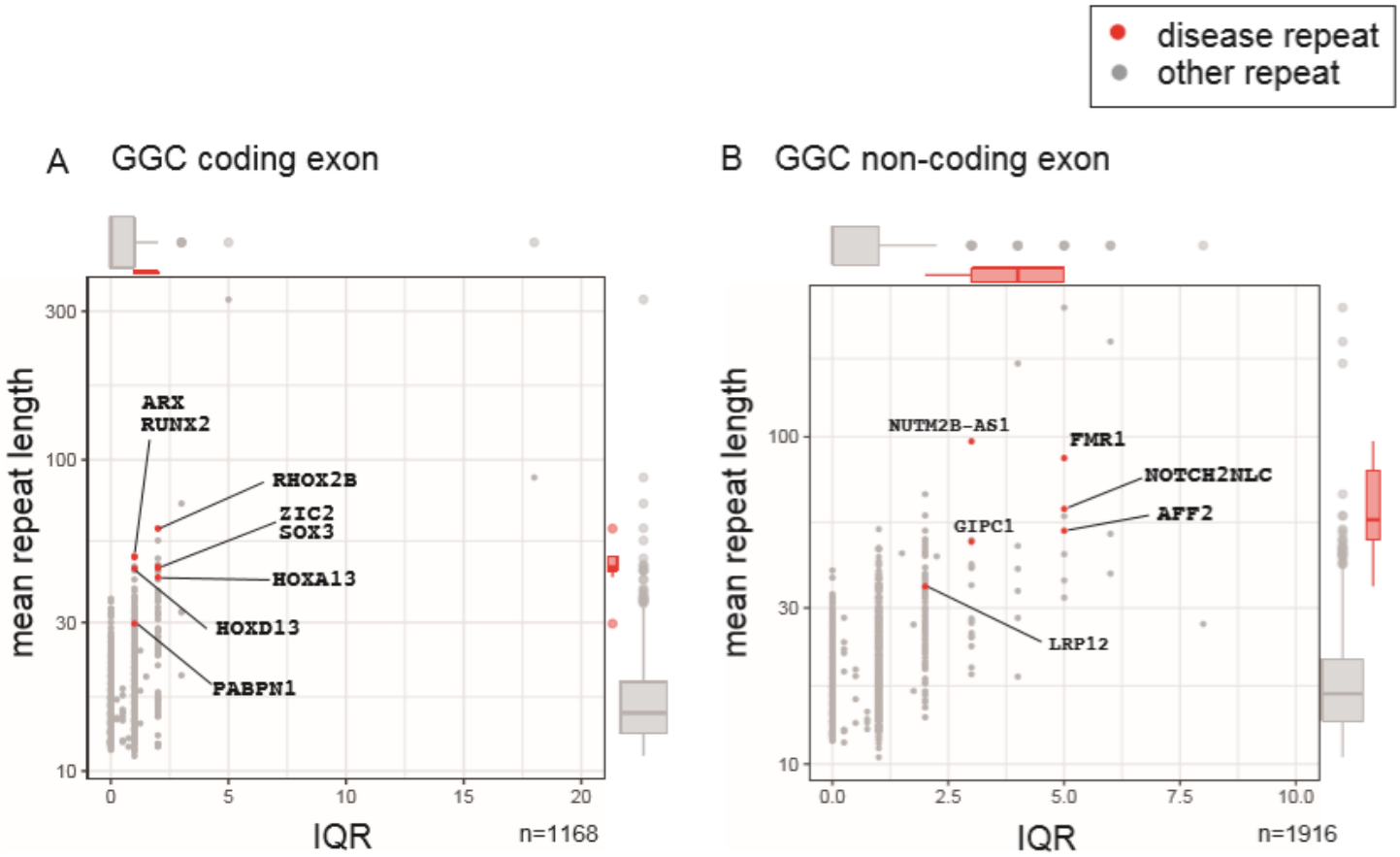

Figure 2

Variation (IQR) and length of repeats with disease-associated sequences. coding GGC repeats (A), and non-coding exonic GGC repeats (B). x-axis: IQR, y-axis: mean repeat length (bp). n provides the numbers of repeat loci. In merged boxplots on the right and upper, ranges are the 25th and 75th percentiles, dots are outliers and lines in boxes are median.

Figure 3

Variation (IQR) and length of repeats with disease-associated intronic AAAAT sequences. x-axis: IQR, y-axis: mean repeat length (bp). n provides the numbers of repeat loci. In merged boxplots on the right and upper, ranges are the 25th and 75th percentiles, dots are outliers and lines in boxes are median.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedData.txt
- SupplementalTables.xlsx
- SupplementaryFigures.pdf
- Table1.png