Evolution of a Human-Specific Tandem Repeat Associated with ALS

Meredith M. Course,[1] Kathryn Gudsnuk,[1] Samuel N. Smukowski,[1] Kosuke Winston,[1] Nitin Desai,[1] Jay P. Ross,[2,3] Arvis Sulovari,[4] Cynthia V. Bourassa,[2,5] Dan Spiegelman,[2,5] Julien Couthouis,[6] Chang-En Yu,[7] Debby W. Tsuang,[7] Suman Jayadev,[1,8] Mark A. Kay,[6,9] Aaron D. Gitler,[6] Nicolas Dupre,[10] Evan E. Eichler,[4,11] Patrick A. Dion,[2,5] Guy A. Rouleau,[2,3,5] and Paul N. Valdmanis[1,4,*]

Tandem repeats are proposed to contribute to human-specific traits, and more than 40 tandem repeat expansions are known to cause neurological disease. Here, we characterize a human-specific 69 bp variable number tandem repeat (VNTR) in the last intron of *WDR7*, which exhibits striking variability in both copy number and nucleotide composition, as revealed by long-read sequencing. In addition, greater repeat copy number is significantly enriched in three independent cohorts of individuals with sporadic amyotrophic lateral sclerosis (ALS). Each unit of the repeat forms a stem-loop structure with the potential to produce microRNAs, and the repeat RNA can aggregate when expressed in cells. We leveraged its remarkable sequence variability to align the repeat in 288 samples and uncover its mechanism of expansion. We found that the repeat expands in the 3′-5′ direction, in groups of repeat units divisible by two. The expansion patterns we observed were consistent with duplication events, and a replication error called template switching. We also observed that the VNTR is expanded in both Denisovan and Neanderthal genomes but is fixed at one copy or fewer in non-human primates. Evaluating the repeat in 1000 Genomes Project samples reveals that some repeat segments are solely present or absent in certain geographic populations. The large size of the repeat unit in this VNTR, along with our multiplexed sequencing strategy, provides an unprecedented opportunity to study mechanisms of repeat expansion, and a framework for evaluating the roles of VNTRs in human evolution and disease.

Introduction

More than 40 tandem repeat expansions in the human genome are known to cause neurological disease.[1–3] This number continues to increase with the growing adoption of long-read sequencing technology, which can sequence longer repeats like variable number tandem repeats (VNTRs; repeats with a repeat unit of seven or more nucleotides). Until now, most of the tandem repeats associated with disease have been short tandem repeats (STRs; repeats with a repeat unit of six or fewer nucleotides), and the mechanism by which disease-associated repeats expand have been difficult to study, since their repeat tracts are generally uninterrupted, and thus their exact locations of expansion are ambiguous. Long-read sequencing technology reveals that many VNTRs are far more polymorphic than the reference human genome suggests. Their length and variability provide us with an unprecedented opportunity to observe their mechanism of expansion.

So far, two VNTRs have been extensively studied in neurological disease: one in ATP binding cassette subfamily A member 7 (*ABCA7* [MIM: 605414]) associated with Alzheimer disease (MIM: 104300)[4] and one in calcium voltage-gated channel subunit alpha1 C (*CACNA1C* [MIM: 114205]) associated with schizophrenia (MIM: 181500) and bipolar disorder (MIM: 125480).[5] Both of these VNTRs were studied because they were found in close proximity to a genome-wide association study signal for the associated disease. The high incidence of neurological disease in humans is partly attributed to rapid changes in genes involved in brain function,[6–8] and tandem repeats are proposed to contribute to these human-specific traits.[9] To better understand the role that tandem repeats could play in human-specific brain health and disease, we took a genome-wide approach, looking for VNTRs that expanded only in humans and exhibited far greater variability in a neurological disease population, as compared to the reference genome.

One neurodegenerative disease in which repeat expansions contribute to a substantial number of cases is amyotrophic lateral sclerosis (ALS [MIM: 105400]). ALS is a rapidly progressive and uniformly fatal motor neuron disease. Currently, the most common variant found in both familial and sporadic cases is an intronic hexanucleotide tandem repeat expansion in C9orf72-SMCR8 complex subunit (*C9orf72* [MIM: 614260]).[10,11] Another repeat expansion in ataxin 2 (*ATXN2* [MIM: 601517]) modifies disease in ALS, when repeat copy number is between 27

[1]Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA 98195, USA; [2]Montreal Neurological Institute and Hospital, McGill University, Montreal, QC H3A 2B4, Canada; [3]Department of Human Genetics, McGill University, Montreal, QC H3A 0C7, Canada; [4]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [5]Department of Neurology and Neurosurgery, McGill University, Montreal, QC H3A 2B4, Canada; [6]Department of Genetics, Stanford University, Stanford, CA 94305, USA; [7]Geriatric Research, Education, and Clinical Center, VA Puget Sound Health Care System, Seattle, WA 98108, USA; [8]Department of Neurology, University of Washington School of Medicine, Seattle, WA 98195, USA; [9]Department of Pediatrics, Stanford University, Stanford, CA 94305, USA; [10]Neuroscience Axis, CHU de Québec-Université Laval & Department of Medicine, Université Laval, Quebec City, QC G1J 1Z4, Canada; [11]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
*Correspondence: paulnv@uw.edu
https://doi.org/10.1016/j.ajhg.2020.07.004.

to 33 (but not > 34, which causes spinocerebellar ataxia 2 [SCA2 (MIM: 183090)]).[12,13] Other key genes with pathogenic coding changes include superoxide dismutase 1 (*SOD1* [MIM: 147450]),[14] TAR DNA binding protein (*TARDBP*, which encodes TDP-43 [MIM: 605078]),[15,16] and FUS RNA binding protein (*FUS* [MIM: 137070]).[17,18] Still, though, these and other genetic variants account for only ~50%–60% of ALS cases with a strong family history of ALS (fALS) and 10%–20% of those with sporadic ALS (sALS).[19] The missing heritability for this complex disease could in part be explained by other unidentified tandem repeat expansions, which could act as causes or modifiers of disease.[20] Indeed, two or more mutations that individually have a low probability of causing disease, but act synergistically to cause disease, have already been identified in a small number (1%–4%) of fALS and sALS cases.[21]

In addition to uncovering disease-causing variants, characterizing the full-length sequence of VNTRs can increase our understanding of the mechanisms underlying tandem repeat expansion. The mechanism by which tandem repeats expand has so far been difficult to pinpoint, as the uniformity of most STRs (e.g., CAG trinucleotide repeats) obscures the exact location where expansion occurs. In these cases, it is hard to know by how many repeat copies a repeat expands, and whether the expansions occur from the 3′ end, 5′ end, or even in the middle of the repeat sequence. In contrast, by looking at the full-length sequence of VNTRs, which exhibit significantly more internal nucleotide variability, we can track the exact breakpoints of an expansion.

Here, we first identified possible disease-associated VNTRs by inferring the lengths of human-specific tandem repeats in a whole genome database of sALS cases. Of these, we found a 69 bp repeat in WD repeat domain 7 (*WDR7* [MIM: 613473])—a gene involved in synaptic transmission[22,23]—that varied markedly compared to the human reference genome. This repeat also lacked any adjacent regions that could have facilitated its expansion (like retroelements). Subsequent long-read sequencing confirmed the variability in length that we observed, and revealed a unique pattern of internal variability that provided us with an unparalleled opportunity to visualize and analyze its pattern of expansion. We found that higher copy number of the VNTR was enriched in cases of sALS, and that the VNTR can form repeat-derived microRNAs and RNA aggregates *in vitro*. In addition to its association with disease, the variability of this repeat also provided a unique opportunity to observe the timing and location of its expansion, as well as its state in ancient genomes and diverse modern-day populations.

Material and Methods

DNA Samples

Informed consent for the *ALS3*-affected family was obtained from McGill University.[24] Blood samples were collected and converted to lymphoblast cell lines. DNA was extracted from cell lines using a DNeasy kit (QIAGEN 69504). Additional 96-well plates of DNA samples were obtained from the Coriell Institute. In total, DNA from 376 individuals with ALS, 531 individuals with Parkinson disease, and 639 control subjects was obtained. DNA from 64 individuals with Alzheimer disease and 32 age- and sex-matched control subjects was obtained from the University of Washington (UW) Alzheimer Disease Research Center (ADRC).

PCR

To PCR amplify the *WDR7* repeat, standard *Taq* DNA Polymerase (New England BioLabs M0273) was used, following the manufacturer's protocol for sample preparation with about 25 ng of DNA. The annealing temperature of the thermocycler was reduced from 60°C to 55°C over a 10 cycle period, followed by 25 cycles at 55°C, and the extension time was set at 3 min per cycle. This method successfully covered both alleles of each sample about 70% of the time. For those samples in which one or both alleles did not amplify, the PCR was repeated using LongAmp *Taq* DNA Polymerase (New England BioLabs M0323). Samples were similarly prepared, following the manufacturer's protocol. The annealing temperature was set at 55°C and the extension time was set at 4 min per cycle. The entire 25 μL sample was then loaded with loading dye onto a 1% agarose gel. The repeat copy number in each allele was calculated from the band size on the gel. Primers used for the repeat lay 82 bp before and 84 bp after the repeat location: 5′-GCCGAATTTGAAGAGGTCATAG-3′ (forward) and 5′-TAGAA AAGGCCATTACAACTGG-3′ (reverse).

PacBio SMRT Sequencing

We first PCR amplified the *WDR7* repeat using the same primers listed above for 10 cycles, using 100 ng of DNA. Then, primers with 12 nt barcodes were used to amplify for 25 more cycles. We used combinations of 12 forward and 12 reverse primers to yield 144 distinct amplicons. Samples were pooled and purified and their concentration was measured. Free primers were then removed with Ampure beads. The sequencing library was prepared and SMRT bell adaptors were ligated by UW PacBio Sequencing Services. Samples were run on a PacBio Sequel System using v.2.1 chemistry in the first round, and a Sequel II System using v.3.0 chemistry in the second round.

Sample De-multiplexing and Analysis

Samples were sorted by their combination of forward and reverse barcodes, yielding ~132 distinct files (for example, 92% samples were successfully amplified in the first round). Each sample, grouped by its barcodes, was then merged to identify multiple instances of the same identical sequence using the FASTX-toolkit collapsed command, which we took to be an example of the true sequence. In events where we identified only singleton reads, we performed ClustalW alignment of these reads. We then took the consensus sequence from each as the representative sequence from this individual. By aligning and quantifying repeat units across all samples, we obtained a sequence logo using WebLogo.[25]

Sequence Alignment Plot Generation

To align the two consensus alleles from each sample after de-multiplexing, we first trimmed all sequences before and after the repeat. Each repeat unit was assigned a single letter code. We did not use the letters A, C, G, or T, to avoid any ambiguity associated

with using DNA nucleotides. If a sequence did not correspond to the top 18 resulting letters/repeat units, it was assigned an "X." We then used ClustalW protein alignment tools to align each string of letters from each allele as if they were amino acids. When visualizing the resulting sequence, we noticed that most of the sequence conservation occurred at the 3′ end, which was much less variable than the 5′ region. Therefore, we right-justified each of the reads, while maintaining some short gaps where contractions were most parsimonious with the alignment. We then manually swapped sequences to most closely resemble their neighbors. Each letter was then converted to its own unique color to improve visualization. Finally, samples were de-coded to reveal their clinical status.

Repeat Length Estimation

We adopted a previously published method[5] to estimate read depth using whole-genome sequencing data. Reads were counted that mapped to the repeat, compared to a 100 kb window of genomic sequence around the *WDR7* repeat. The fraction of enrichment or depletion of reads was used to calculate the estimated length compared to the reference human genome (hg38, which shows six copies of the repeat). Whole-genome sequencing datasets used included those from the National Institute on Aging Genetics of Alzheimer Disease Data Storage Site (NIAGADS) and Answer ALS. Quebec samples come from Canadians of mostly French ancestry living in Quebec. All samples analyzed were aligned to hg38. Samples were of European descent (self-reported).

Calculation of Inter-repeat Distance

To calculate the distance between individual repeat units, we counted the distance between one repeat unit and all subsequent repeat units in that sample. For instance, if one repeat unit appeared at positions 2, 6, and 12 in a repeat, the numbers 4 (6 − 2), 10 (12 − 2), and 6 (12 − 6) would be scored. This process was repeated iteratively across all repeat units from each sample that underwent long-read sequencing.

Ancient DNA Calculations

Raw data for ancient DNA calculations were obtained from published genomes of Altai Neanderthal[26] and Denisovan[27] samples. Regions corresponding to the *WDR7* repeat were converted to SAM files and individual reads were queried for the presence of a complete 69 bp repeat unit contained within the 100 bp sequence read.

1000 Genomes Project Sample Analysis

Reads that mapped to the *WDR7* repeat (hg19 co-ordinates chr18:54,691,726–54,692,180) were extracted from BAM files for each individual from the 1000 Genomes Project.[28] The presence of each repeat unit was counted across each sample. To identify novel repeat units not identified in our initial PacBio screen, we searched for sequences that had a match to either the 12 nt sequences at the beginning or the end of the repeat and were 67–70 nt long.

Construct Generation

The same PCR primers used to amplify the *WDR7* repeat were used for cloning, with the addition of overhangs to clone the product into the pIRES-NEO vector using the In-Fusion HD Cloning Plus kit (Takara 638910). Repeats of 1, 4, 6, 10, 15, 27, and 36 copies in length were amplified from case and control DNA samples from the Coriell Institute.

Cell Culture and Transfection

HEK293 cells were obtained from the American Type Culture Collection (ATCC). They were grown in 10% DMEM and transfected using Lipofectamine 3000 Transfection Reagent (Invitrogen L3000015) following the manufacturer's instructions. RNA was extracted 48 h later using QIAzol Lysis Reagent (QIAGEN 79306). MEF cells were also obtained from the ATCC and were grown in 15% DMEM. They were transfected in the same manner as the HEK293 cells. Lymphoblasts were cultured in 10% IMDM.

RNA Localization

Custom Stellaris fluorescent *in situ* hybridization (FISH) probes labeled with a Quasar 570 dye were ordered from LGC Biosearch Technologies. HEK293 and MEF cells were treated with the probe following the manufacturer's instructions, then imaged on an Olympus FluoView FV1000 confocal microscope. Images were analyzed in CellProfiler,[29] using the Speckle Counting pipeline.

Small RNA Sequencing

Small RNA sequencing was performed as previously described.[30,31] Briefly, 3 μg of RNA from HEK293 cells were ligated to 3′ Universal miRNA Cloning Linker (New England Biosciences S1315) using T4 RNA Ligase 1 (New England Biosciences M0204) without ATP, then run on a 15% urea-polyacrylamide gel. 17–28 nt fragments were excised and ligated to 5′ barcodes, again using T4 RNA ligase, then multiplexed and sequenced on an Illumina 50 bp miSeq machine at the UW Center for Precision Medicine. Adaptors and barcodes were trimmed, retaining small RNAs $>$ 18 nt in length, and then sequences were aligned to human microRNAs on miRBase (release 15)[32] using Bowtie v.0.12.7, allowing for 2 mismatches.[33]

MicroRNA Target Prediction

To evaluate predicted targets for the microRNAs produced from the *WDR7* repeat, we used the seed sequence of the most abundant microRNA sequence detected from small RNA sequencing (UCACAUA) as input for the TargetScan v7.2 program.[34] Considering the human-specific nature of the repeat, we did not use species conservation as a consideration when reporting target mRNAs.

RNA Folding

RNA from non-human primates, the reference human genome (GRCh38), and one affected individual from the *ALS3*-affected family was entered in the RNA-fold program—part of the ViennaRNA Package 2.0.[35] Centroid plots demonstrating base-pair probabilities are presented here.

Phylogenetic Tree Analysis

A total of 28 samples from 5 different 1000 Genomes Project[28] populations (ASW, CEU, CHB, JPT, and YRI) were included in our long-read sequencing pipeline. We used the VCFtoTree program[36] to extract sequence data before and after the *WDR7* repeat and generate alignments of the samples. We included 10 kb of sequence immediately before the repeat (hg19 chr18: 54677408–54687408). The sequence immediately after the repeat had few single nucleotide polymorphisms (SNPs), precluding resolution of the different populations on a tree; as a result, we used a sequence starting 10 kb after the repeat until 20 kb after the repeat (hg19 chr18: 54,694,892–54,704,892). We used FastTree 2[37] to generate approximately-maximum-likelihood phylogenetic trees of the resulting alignments. We then overlaid the two alleles

obtained from long-read sequencing for each individual and kept the allele that most closely matched its neighbor. While the VCFtoTree alignment is phased for each individual, SNPs were not present in the section of the repeat that was amplified, so we could not match each repeat allele to the phased tree segments.

RNA Editing Analysis

We extracted RNA-seq reads that map to the *WDR7* repeat from brain samples from the Mount Sinai Brain Bank that have been deposited in the Accelerating Medicines Partnership – Alzheimer disease (AMP-AD) database.[38] Reads were aggregated across all samples. In instances where an A-to-G change was identified in the RNA sequencing reads but not from our combined DNA sequence analysis, we calculated the percent frequency of each A-to-G mismatch out of total reads.

Statistical Analysis

Statistical analyses were performed using Prism 8.0.1 (GraphPad Software). The D'Agostino-Pearson omnibus K2 test was used to test whether the data were normally distributed. A two-tailed Mann-Whitney test was used to compare groups of two with nonparametric distribution, for which median and interquartile range (IQR) as well as mean and standard deviation (SD) are given. A one-way Kruskal-Wallis test was used to compare groups greater than two with nonparametric distribution, followed by Dunn's multiple comparisons test if the Kruskal-Wallis gave $p < 0.05$. R^2 values on scatterplots were produced using simple linear regression.

Results

To identify VNTRs that could act as disease modifiers for ALS, we evaluated VNTRs that are human specific, located in introns, and highly variable according to analysis of 15 individuals whose genomes were sequenced and phased in their entirety using long-read sequencing.[39] We focused on intronic regions (as opposed to intergenic regions) to identify transcribed regions most likely to influence disease, similar to the hexanucleotide repeat in *C9orf72*.[10,11] We also excluded repeats that were part of repetitive elements, especially SINE-VNTR-Alu repeats, which have a variable length, but map to many locations in the human genome. A total of 20 repeats matched this criteria, with a VNTR repeat unit length > 25 nt that mapped uniquely to one location in the human genome (Table S1). Their genomic co-ordinates were then used to obtain estimated read length relative to the reference human genome in a dataset of 97 genomes from Answer ALS (Figure 1A). One 69 bp VNTR in the final intron of *WDR7* was particularly striking in its large size and variable length (Figure 1A). *WDR7* encodes the β-subunit of rabconnectin-3, which is involved in the Ca^{2+}-dependent exocytosis of neurotransmitter and is highly expressed in the brain.[40] Furthermore, *WDR7* was compelling given that its genomic location is within the bounds of the *ALS3* (MIM: 606640) locus on chr18q21.[24] While the family used to map the locus was subsequently found to have a variant in *FUS*,[17] 14 of 15 affected individuals nonetheless still map to the chr18q21 region, indicating that a modifying gene may be present.

The *WDR7* Repeat Sequence Is Primate Specific and Expands Only in Humans

Through phylogenetic analysis, we determined that the *WDR7* region of interest is present only in primates (Figure 1B). By analyzing the reference genomes of non-human primates and phased genomes of several non-human great apes, we found that squirrel monkeys, marmosets, green monkeys, baboons, rhesus macaques, and gibbons share a 47 bp region, which corresponds to about two-thirds of the 69 bp repeat unit (Figures 1B and 1C). Conservation of this 47 bp sequence drops to 74% in mouse lemurs. Orangutans, chimpanzees, gorillas, and bonobos all have exactly one copy of the full 69 bp sequence in addition to the 47 bp, and the sequence itself is identical between species (Figure S1). The human reference genome, by contrast, shows 6 copies of the 69 bp sequence, in addition to the 47 bp. Intriguingly, the 69 bp sequence found in great apes is predicted to form a unique secondary structure, with complementary regions that form a hairpin (Figure 1C). Expansion of the repeat in humans appears *ab initio*, meaning without any adjacent or internal repeat sequences that could facilitate its expansion, making it unique among VNTRs.[39] Overall, this analysis shows that expansion of the repeat is specific to the human lineage.

The *WDR7* Repeat Is Highly Variable in Length and Longer in Case Subjects with sALS

Given its unique properties, we further characterized the distribution of the VNTR length by PCR amplification followed by SMRT sequencing in 376 individuals with sALS, 531 individuals with sporadic Parkinson disease (sPD [MIM: 168600]), and 639 control samples obtained from the Coriell Institute. The distribution of the longer amplified allele per individual ranged from 1 to 86 copies (Figure 2A). The mean and median copy number of the longest allele per individual was also significantly higher in individuals with ALS than control subjects ($p = 0.0003$; Mann-Whitney test), suggesting that the repeat may modify ALS susceptibility. Notably, the largest repeat length we observed (86 copies) was detected in an individual who developed sALS at age 72.

To validate this finding, we estimated *WDR7* repeat copy number in whole-genome sequencing data from NIAGADS ($n = 917$ case subjects with Alzheimer disease and 675 control subjects), Quebec ($n = 159$ case subjects with sALS and 311 control subjects), and Answer ALS databases ($n = 307$ case subjects and 53 control subjects). SALS samples from the Quebec cohort and the Answer ALS cohort exhibited significantly higher repeat copy number than the Quebec control subjects, as well as the NIAGADS control subjects (Figure 2B; $p < 0.0001$; Kruskal-Wallis test followed by Dunn's multiple comparisons). For comparison, the repeat copy number in sporadic Alzheimer disease (AD) samples available through NIAGADS was no different than the control subjects (Figure 2B). While in all cohorts we observed a relationship between longer repeat length and ALS risk, we did not observe a relationship between longer repeat

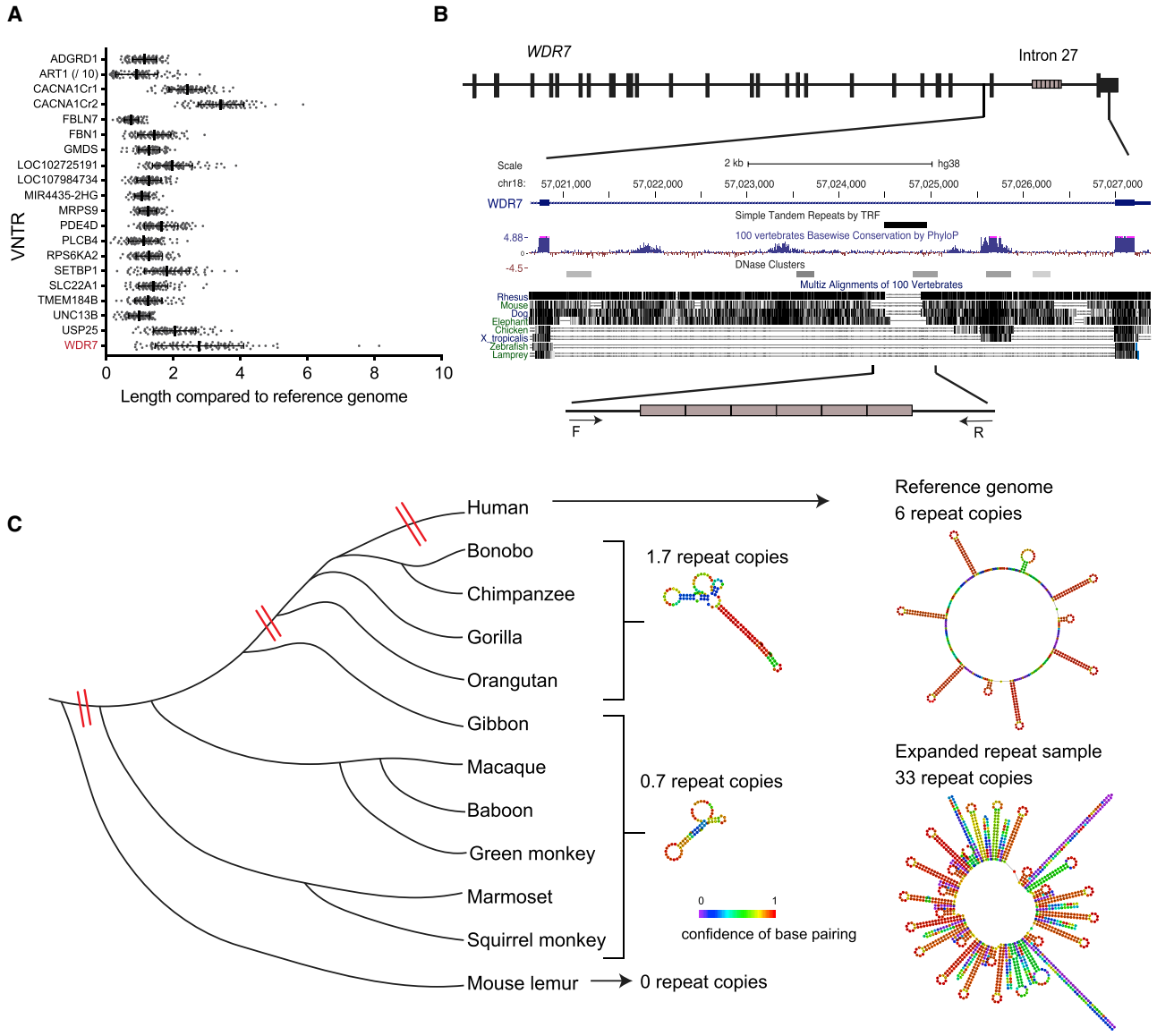

Figure 1. A VNTR in *WDR7* Is Highly Variable in Humans

(A) Estimated read length of several human-specific and intronic VNTRs that are not derived from repetitive elements. Read length is given relative to the reference genome, to evaluate how variable the read length of each VNTR can be. Read length was obtained from the Answer ALS database (n = 97 samples). Black lines show mean and standard deviation. The VNTR in *WDR7* exhibits greater variability than the other VNTRs.

(B) Position of the VNTR in *WDR7* intron 27, adjacent to a DNase I hypersensitivity site, and regions of multi-species conservation. The repeat itself is not conserved across species. The bottom schematic shows the region that was included in subsequent PCR amplification.

(C) Phylogenetic tree showing events of repeat region evolution in humans and non-human primates (red hashmarks). Adjacent are the predicted RNA structures for each iteration.

length and age of disease onset, similar to what has been reported for intermediate lengths in *ATXN2* trinucleotide repeat expansions in ALS (Figure S2).[41]

The *WDR7* Repeat Is Highly Variable in Internal Repeat Sequence

We next characterized differences in the internal repeat sequence, since sequence slippage or internal nucleotide variation within the repeat itself has previously been shown to lead to disease.[42–44] To determine the exact nucleotide composition of the sequences repeated in our samples, we used PacBio single molecule, real-time (SMRT) sequencing technology. We PCR amplified the *WDR7* repeat region and multiplexed 144 samples with a repeat length of 20 or greater into one lane of SMRT sequencing. Samples were of European ancestry (self-reported) and included 40 control subjects, 35 sALS samples, 22 sPD samples, and 43 affected and unaffected members of the family with ALS that maps to the *ALS3* locus. We captured the sequences of the 288 alleles for >90% of samples.

Intriguingly, we determined that exactly 6 of the 69 bp in the repeat unit were variable (Figure 2C). None of the

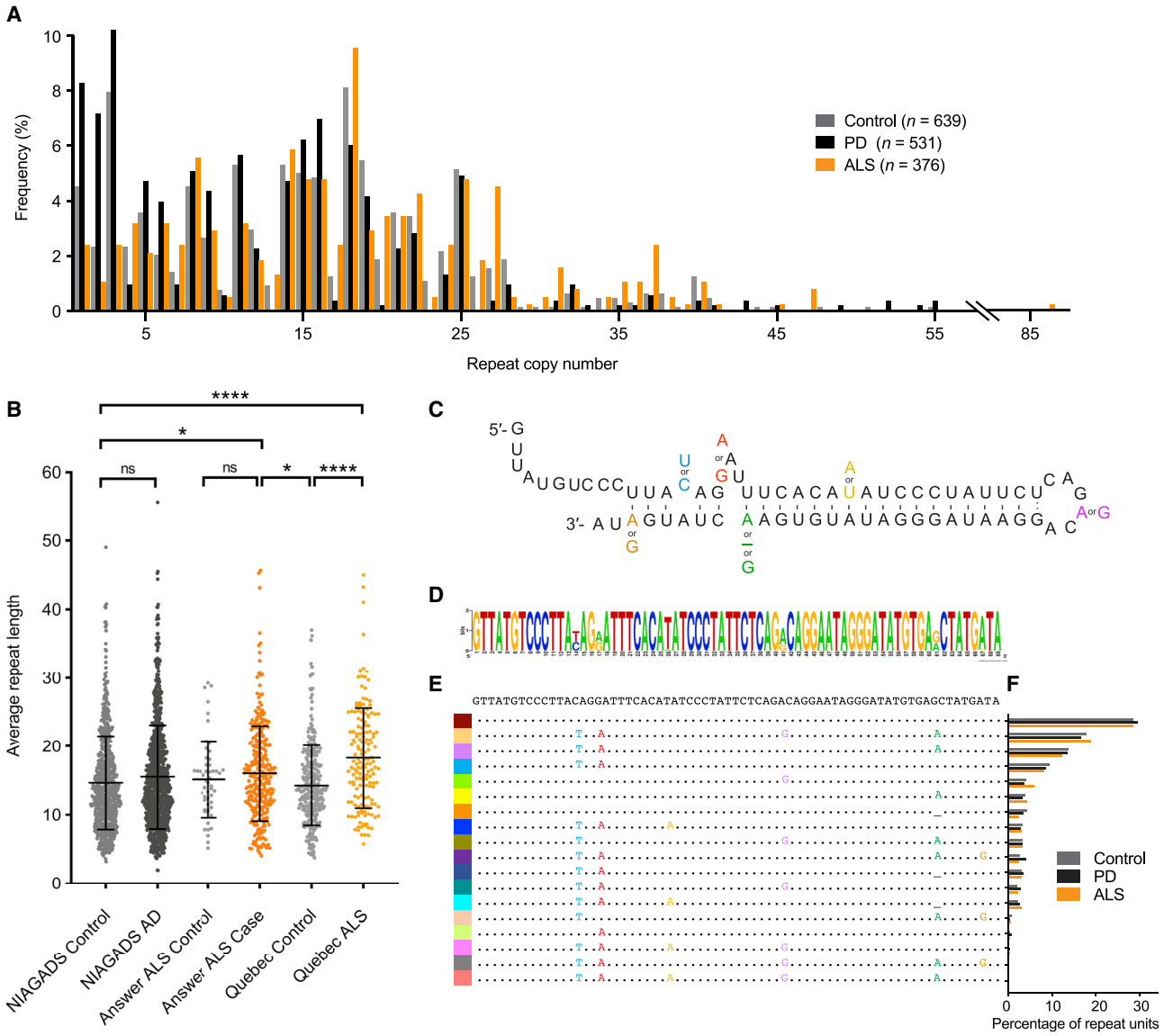

Figure 2. The *WDR7* VNTR Is Longer in Cases of sALS and Variable in Internal Sequence

(A) *WDR7* VNTR copy number distribution in the longest amplified alleles from individuals with sALS and sPD and control subjects obtained from the Coriell Institute. Repeat copy number is significantly higher in cases of ALS (median [IQR] = 17.5 [9–24], mean ± SD = 17.7 ± 10.4) versus control subjects (median [IQR] = 15 [8–21], mean ± SD = 15.2 ± 9.66; p = 0.0003; Mann-Whitney test).

(B) Comparison of *WDR7* VNTR copy number estimated from whole-genome sequencing. Read lengths were obtained from NIAGADS (n = 917 case subjects with AD and 675 control subjects), Quebec (n = 159 case subjects with ALS and 311 control subjects), and Answer ALS databases (n = 307 case subjects and 53 control subjects). Black lines show mean and standard deviation. p values were determined by a Kruskal-Wallis test (which gave $p < 0.0001$), followed by Dunn's multiple comparisons. *$p < 0.05$, ***$p < 0.001$, ****$p < 0.0001$, "ns" means "not significant."

(C) Hairpin secondary structure of one repeat copy, with the 6 of 69 variable base pair positions highlighted.

(D) Repeat logo of the *WDR7* VNTR, indicating relative conservation or variability at each nucleotide position.

(E) The 18 "parent" repeat unit sequences that account for >99% of repeat units identified. Variable nucleotides are highlighted. Each parent sequence was assigned a color, shown on the left, used for alignment and visualization in Figure 3.

(F) The frequency of each repeat unit across all individuals (n = 144 individuals, totaling 6,041 distinct repeat units).

observed changes were compensatory changes (i.e., when one nucleotide of a self-complementary region is modified to restore base-pairing with a nucleotide on the other side of the duplex). By aligning repeat units across all samples, we obtained a sequence logo (Figure 2D). We also estimated concordance between SMRT long-read sequence length and length estimated by PCR (Figure S3; $R^2 = 0.479$). Our high depth of coverage obviated any issues associated with single nucleotide variants, especially short insertions or deletions from long-read sequencing. More than 99% of repeat units resolved into 18 parent sequences, with no intervening nucleotides between repeats and no insertions, deletions, or nucleotide changes at any of the 63 of 69 fixed positions.

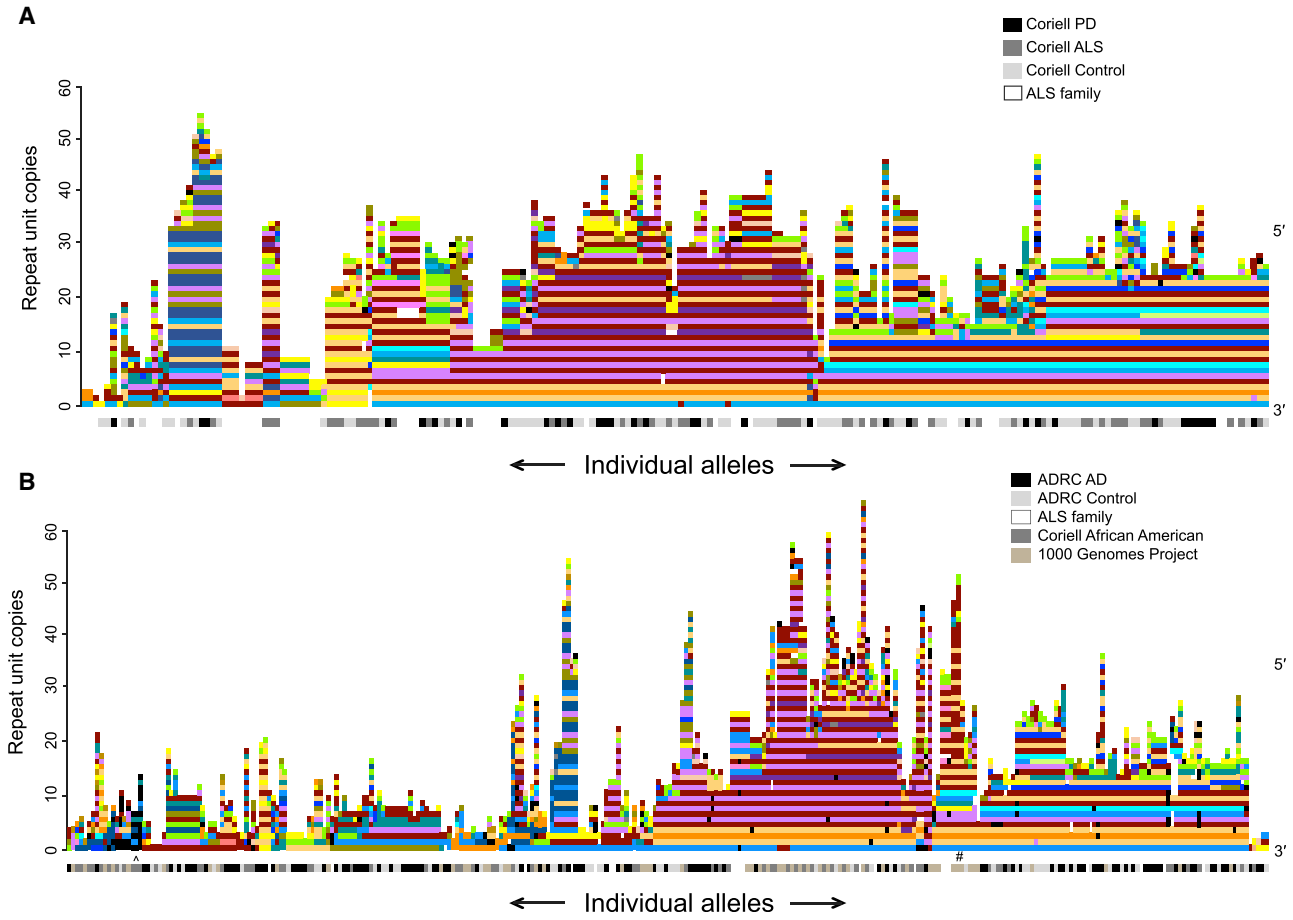

Figure 3. Alignment of Repeat Units across Individuals Is Clustered and Reproducible
(A) Repeat units were color coded (colors assigned in Figure 2E; black indicates a rare repeat unit without an assigned color), and then each individual's alleles were plotted as a series of colors, from the 5′ to 3′ end (top to bottom) of the VNTR. Here, alleles are aligned based on similarity at the 3′ end of the sequence. Repeat copy number is shown on the y axis. Samples used were the same as those presented in Figure 2A. For this round of sequencing, samples selected had at least one of the two alleles 20 repeats or greater.
(B) The same visualization and alignment strategy was applied to a second cohort of individuals. This cohort was expanded to include samples from individuals with AD, samples from the 1000 Genomes Project, and samples from African American individuals, obtained through Coriell. For this round of sequencing, all samples were sequenced irrespective of repeat length. An allele unique to samples from individuals of African descent is denoted by a "^," and an allele found only in samples from individuals of Han Chinese descent is denoted by a "#."

These 18 parent sequences were comprised of several combinations of the six variable nucleotides present in each repeat unit. The most abundant repeat unit accounted for 28.6% of total repeat units. The next two accounted for 17.2% and 13.6%, and the rest under 10% each (Figure 2E). Any other repeat sequences accounted for only 0.3% or less of reads and were also predominantly composed of different combinations of the six variable nucleotides. The distribution of each repeat unit was markedly similar between sALS-affected subjects, sPD-affected subjects, and control subjects (Figure 2F).

Alignment of *WDR7* Repeat Alleles Reveals Remarkable and Reproducible Patterning

To visualize and align the repeat units across individuals, we first color-coded each repeat unit sequence (assigned colors shown in Figure 2E). Plotting each full repeat allele for the 144 SMRT-sequenced samples as a stack of colors and then aligning them using ClustalW revealed a remarkable bias of alignment to the 3′ end of the repeat (Figure 3A). This divergence is in the opposite direction of transcription. In contrast, we did not observe the same pattern or structure when the samples were ranked by length and forced to align to the 5′ end (Figure S4). Plotting the data in this color-coded way also revealed several broad families of repeat patterns, indicated by "clustering" of similar patterns, suggesting that some variation originated independently (Figure 3A).

Given the novelty of the pattern we identified, we next determined whether a second cohort of individuals—including those of non-European ancestry—would exhibit a similar clustering pattern. In addition to 64 samples from individuals with AD and 32 matched control subjects, we sequenced 28 samples from the 1000 Genomes Project and 22 African American samples from the Coriell Institute. For this round of sequencing, we did not size select. Even so,

we observed the same striking clustering pattern (Figure 3B). We did note some super-population-specific regions, such as an expansion in samples from individuals of Han Chinese descent—which started about halfway into a repeat shared with other population samples—and repeat unit sequences unique to samples from individuals of African descent.

Repeat Expansion Occurs in One Direction and in Multiples of Two

The mechanism by which tandem repeats expand is hotly debated, and the location of the expansion is difficult to discern in STRs because each repeat unit is identical (e.g., CAG). We leveraged the heterogeneity of the observed repeat units to gain insight into how the *WDR7* repeat expanded. We first found several instances of repeat unit blocks that were duplicated within the same repeat allele (Figures 3 and 4A), which could represent homologous recombination events. Duplications appear to be relatively rare events; however, when they occur, they become the new baseline length upon which additional repeats are added. Interestingly, the duplications were composed of complete repeat units with no intervening insertions or deletions.

Surprisingly, there were few instances of the same repeat units situated immediately adjacent to one another. Instead, the same repeat units more often appear every two repeat units. In many cases, combinations of the same two repeat units appeared consecutively for a long stretch, exemplified by the dark red and purple combination in the middle of the plot (Figure 3). To quantify this observation, we calculated the distance between the first instance of a repeat unit and the next, for each repeat unit in each allele (Figures 4A and 4B). When these calculations were summed, we found a remarkable periodicity of the distribution of repeat units, such that they appear almost exclusively in multiples of two. This pattern was so fixed that we were more likely to observe a repeat unit 20 repeats away from itself (1.4 kb) than immediately adjacent or three repeat units away from itself (Figure 4B).

In addition, repeat units that appeared infrequently were biased to the 5′ end of the repeat. If the rare repeat unit was duplicated, its neighboring repeat unit was also duplicated (Figures 3 and 4A). This observation is consistent with our initial finding that the alleles aligned at the 3′ end of the repeat but not at the 5′ end. Taken together, these data indicate that the variability—and thus "leading edge"—of the expansion appears at the 5′ end of the repeat, and that the expansion occurs in combinations of two repeat units.

The *WDR7* Repeat Does Not Exhibit Intergenerational Instability in a Family with ALS

Among the samples we SMRT sequenced were 43 affected and unaffected members of a family with ALS. Across the three generations sequenced, we observed no evidence of variation in repeat length or sequence. The sequence composition from each parent to child showed complete segregation, including a 33-copy repeat that segregated with ALS disease state (Figure 4C). We therefore concluded that expansion of this repeat occurs on a longer timescale than generation-to-generation change.

The *WDR7* Repeat Is Expanded in Ancient Genomes

Understanding how human-specific repeats expand can give insight into human migration patterns. We inferred *WDR7* repeat length in short-read whole-genome sequencing datasets by counting the number of reads that align to the repeat relative to adjacent DNA segments. We were also able to quantify the number of unique repeat units that were present in a sequenced individual when the full repeat unit was captured within a sequence read (which was often 100 bp long).

We first estimated the length and internal composition of the repeat in two ancestral genomes: Neanderthal and Denisovan. The Denisovan genome showed a pile-up of reads, suggesting a repeat length longer than the reference human genome, while the Neanderthal genome did not (Figure 5A). On closer inspection, the Neanderthal genome had four repeats, suggesting it had either one and three or two and two repeat copies on either allele (Figure 5B). Meanwhile, the Denisovan genome had approximately 50 repeats in total, across the two alleles (Figure 5C). Based on the repeat unit composition, we were able to match the Denisovan repeats to some of the modern-day alleles we obtained from long-read sequencing (Figure 5C). Collectively, these results suggest that expansion of the *WDR7* repeat pre-dates modern humans.

Geographical Patterns of the *WDR7* Repeat Length and Composition

We next examined the distribution of the top 18 repeat units across the 1000 Genomes Project populations and observed largely equal abundance and range (Figure 6A). While searching for novel repeat units across these samples, we observed a concordance between average repeat length inferred by calculating short-read relative density across the repeat and actual length of the two alleles revealed by long-read sequencing (Figure S5; $R^2 = 0.61$). Moreover, the composition of repeat units extracted from short-read sequencing matched those present in long-read sequencing. Some novel repeat units were identified in African super-populations that had a C>T transition at base pair 56 of the repeat unit, always in combination with a more frequent 1 bp deletion at position 61 and other variable nucleotides at the other five common variable nucleotide positions. To further interrogate population structure of the repeat units, we identified the cumulative abundance of each repeat unit (Figure 6B). This analysis revealed that while the major repeat units were consistent among populations, some less frequent repeat units were enriched or depleted in certain super-populations (Figures 6B and 6C).

When we evaluated the location of these rare variants in the full repeat expansion obtained by SMRT sequencing, we found a strong bias to the 5′ end, consistent with our previous observation that expansion of the repeat was directional. The rare variants only start to appear 14 units

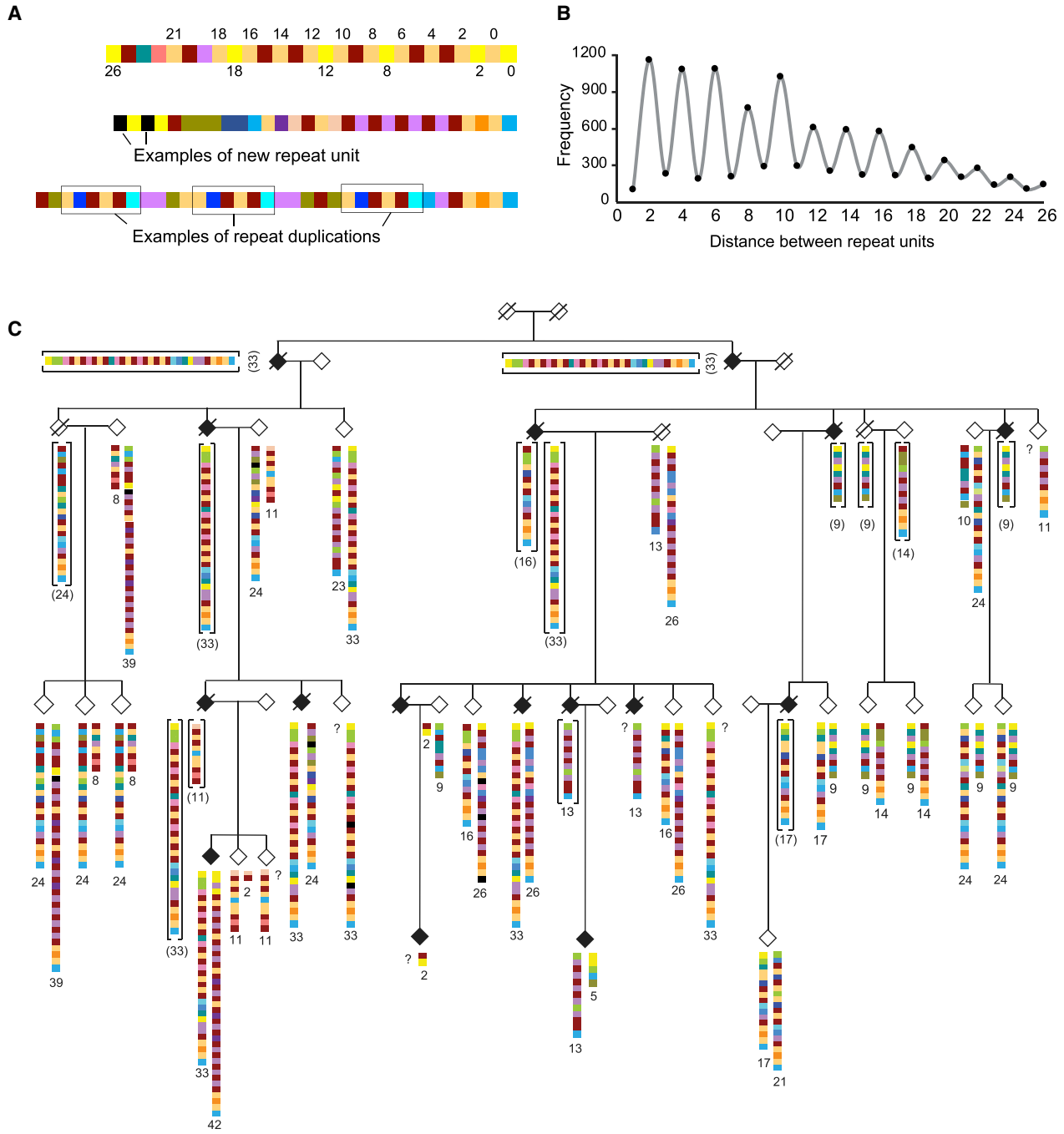

Figure 4. *WDR7* VNTR Repeat Unit Heterogeneity Reveals Patterns of Expansion
(A) Representative examples of observed repeat sequence patterns. Top: example of how we calculated the distance between the first instance of a repeat unit and the next, for each repeat unit in each allele. Middle: example showing that if a rare repeat unit (black) was duplicated, its neighboring repeat unit (yellow) was also duplicated. Bottom: example of larger repeat duplications.
(B) Summation of the distance between each repeat unit and itself. Repeat units largely occur with a two-unit periodicity (combined results across all sequenced individuals from Figure 3A).
(C) Segregation of the *WDR7* VNTR in a pedigree of a family with ALS. Sequences shown in brackets are inferred. A question mark means that the allele did not amplify successfully.

away from the 3′ end of the repeat, and the repeat units that are enriched in super-populations appear closer to this 5′ leading edge (Figures 6D and S6). Therefore, by looking at repeat composition alone, we can approximate the population origin of a DNA sample. The reverse is also true: by looking at the surrounding genetic structure using phased SNP data, we can estimate what the repeat will look like, using a phylogenetic analysis (Figure S7).

The *WDR7* Repeat Can Produce MicroRNAs *In Vitro*

Each repeat unit has up to 21 nt of self-complementary sequence, which, according to RNAfold prediction

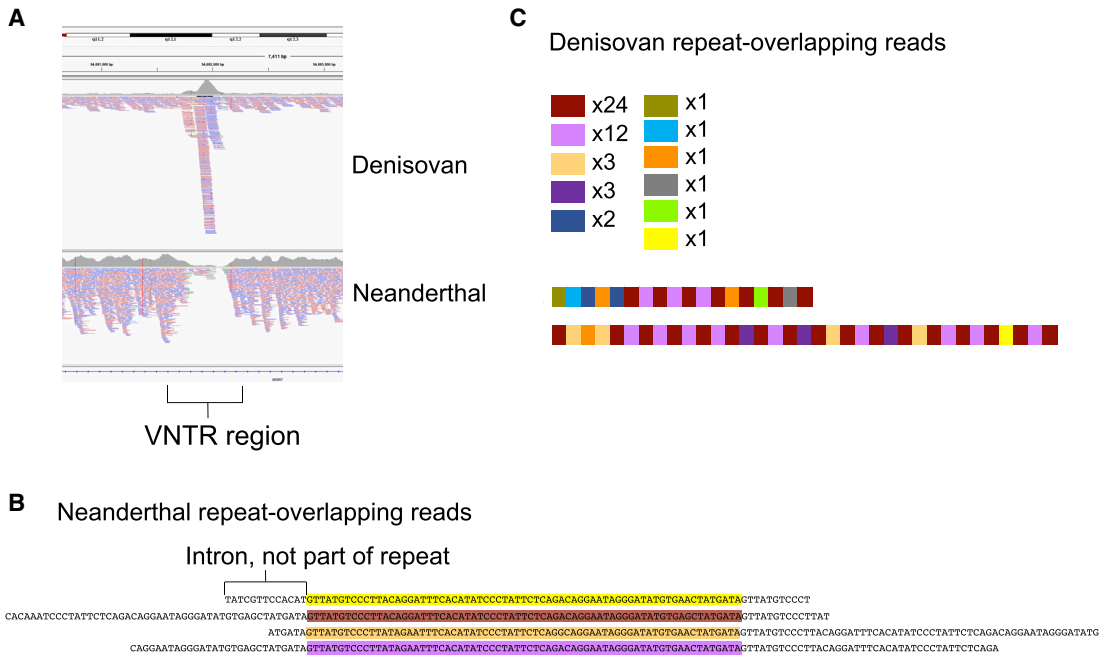

Figure 5. *WDR7* Repeat Length and Sequence Estimation in Ancient Genomes
(A) Alignment of whole genome sequencing reads for Altai Neanderthal and Denisovan genomes. Read distribution in the *WDR7* VNTR region indicates a build-up of reads in the Denisovan genome, and a paucity of reads in the Neanderthal genome.
(B) Individual repeat units identified in the Neanderthal genome. Units were identified in reads with a full 69 bp repeat sequence present in the ~100 bp sequence read (average length of a read).
(C) Individual repeat units identified in the Denisovan genome, the relative abundance of each unit, and correlating modern-day human repeat alleles that match those abundances.

software, likely folds into a stem-loop structure reminiscent of a microRNA precursor (Figure 2C). We therefore hypothesized that this repeat expansion could form microRNAs. To test this hypothesis, we cloned several expansions from individuals with sALS and control subjects, transfected human embryonic kidney (HEK293) cells, and performed small RNA sequencing. We observed a strong correlation ($R^2 = 0.977$) between the number of repeat copies expressed and the abundance of microRNAs produced from the repeat structure, indicating that the repeat could indeed form microRNAs (Figure 7A). We confirmed these results in a second independent transfection and sequencing reaction (Figure S8, $R^2 = 0.918$). Possible targets of the most abundantly detected microRNA, as predicted by TargetScan, are provided in Table S2. While we did not identify the sequence in neuronal datasets, suggesting that the microRNA may not accumulate to appreciable levels *in vivo*, it is possible that in certain disease-relevant regions, this microRNA species could be expressed and influence the transcriptional milieu.

The *WDR7* Repeat Forms RNA Aggregates

To find out whether the repeat was transcribed and exported from the nucleus, we performed RNA FISH using probes against the *WDR7* repeat. We first confirmed specificity of the probes by testing them in mouse embryonic fibroblasts (MEFs), which do not contain the repeat. These cells showed an absence of non-specific binding of the probe (Figure 7B). The MEFs were then transfected with one or 36 copies of the repeat. These cells exhibited a clear "speckling" pattern indicative of RNA aggregates in the cytoplasm of the cells, with significantly more aggregates in the cells transfected with 36 copies of the repeat, as compared to the untransfected cells ($p < 0.0001$; Kruskal-Wallis test followed by Dunn's multiple comparisons) and those transfected with one copy of the repeat ($p = 0.0052$). We repeated the same experiment in HEK293 cells and observed a similar pattern, with significantly more aggregates in the cells transfected with 36 copies of the repeat, as compared to the untransfected cells ($p = 0.0007$; Figure 7B).

The *WDR7* Repeat Is Subject to RNA Editing

Of note, the *WDR7* repeat has stop codons in all six reading frames (Figure S9), diminishing the likelihood of repeat-associated non-ATG translation as a pathological mechanism, which has been proposed for individuals with *C9orf72* repeat expansions.[45] We also noted no instances of alternative splicing or intron detention[46] in qPCR and RNA-seq analysis of lymphoblasts from individuals with ALS (data not shown). This finding is especially relevant considering that *WDR7* is intolerant to loss-of-function variants according to the Genome Aggregation Database (gnomAD; pLI = 1).[47]

However, regions of extensive complementarity are often substrates for editing by adenosine deaminases acting on RNA enzymes.[48] Our profiling of the repeat in thousands

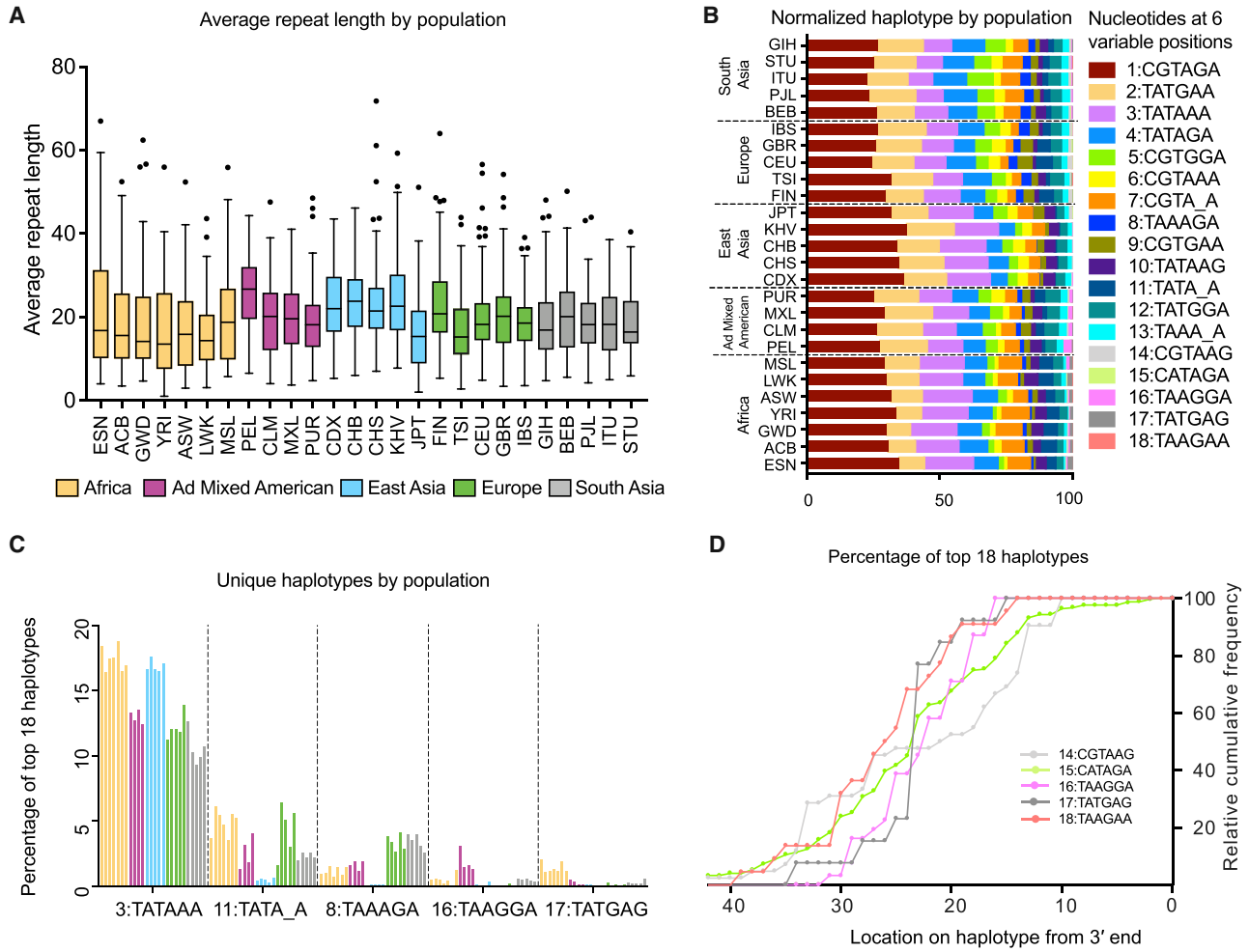

Figure 6. *WDR7* Repeat Length and Sequence in the 1000 Genomes Project Populations
(A) Distribution of *WDR7* VNTR length in 1000 Genomes Project samples, grouped by super-population.
(B) Distribution of the 18 most frequent repeat units within each population. In the legend, the numbers rank the frequency of the sequence overall, along with the nucleotides present at each of the six variable positions in the repeat. An underscore represents a deletion.
(C) Enrichment or depletion patterns of specific repeat units are unique to certain super-populations.
(D) Location of rarest repeat units on the full allele, normalized to 100% for each repeat unit. All 18 repeat units are shown in Figure S6.

of DNA samples permitted us to identify nucleotide changes specific to RNA transcripts, suggestive of RNA editing events. To evaluate the potential of the repeat to be edited, we took advantage of a large set of RNA-seq data from the AMP-AD.[38] This sequencing uses ribosomal RNA depletion rather than poly-A purification as an RNA cloning strategy, which increases the levels of intronic reads. Compiling reads from >900 sequenced individuals revealed the same parent repeat sequences we had previously observed; however, we also identified four positions with A-to-G mismatches, which are the consequence of A-to-I editing (Figure S10).

Discussion

The human genome contains at least 21,442 polymorphic STRs and VNTRs, of which 1,584 were recently classified as human-specific tandem repeat expansions.[39] So far, more than 40 STRs have been linked to neurodegenerative diseases,[1,2] and the recent advent of long-read sequencing technology is already uncovering larger VNTRs that are also associated with disease. For instance, a 30 bp repeat in *CACNA1C* predisposes individuals to bipolar disorder and schizophrenia,[5] while a 25 bp VNTR in the intron of *ABCA7* has been identified as a risk factor for individuals with AD.[4] While greater copy number of the repeat in *WDR7* is linked to ALS, several additional factors set this repeat apart from the others described so far.

This VNTR is comprised of a large 69 bp repeat unit and is notable for its remarkable variability in both length and internal sequence (Figures 1A and 3). Each repeat unit is complementary and predicted to form a stable hairpin (Figures 1C and 2C), and the internal sequence variation for each unit is strictly patterned, enabling an orientation-specific interrogation of read composition and length (Figures 2C–2E). The repeat expands specifically in the human lineage,

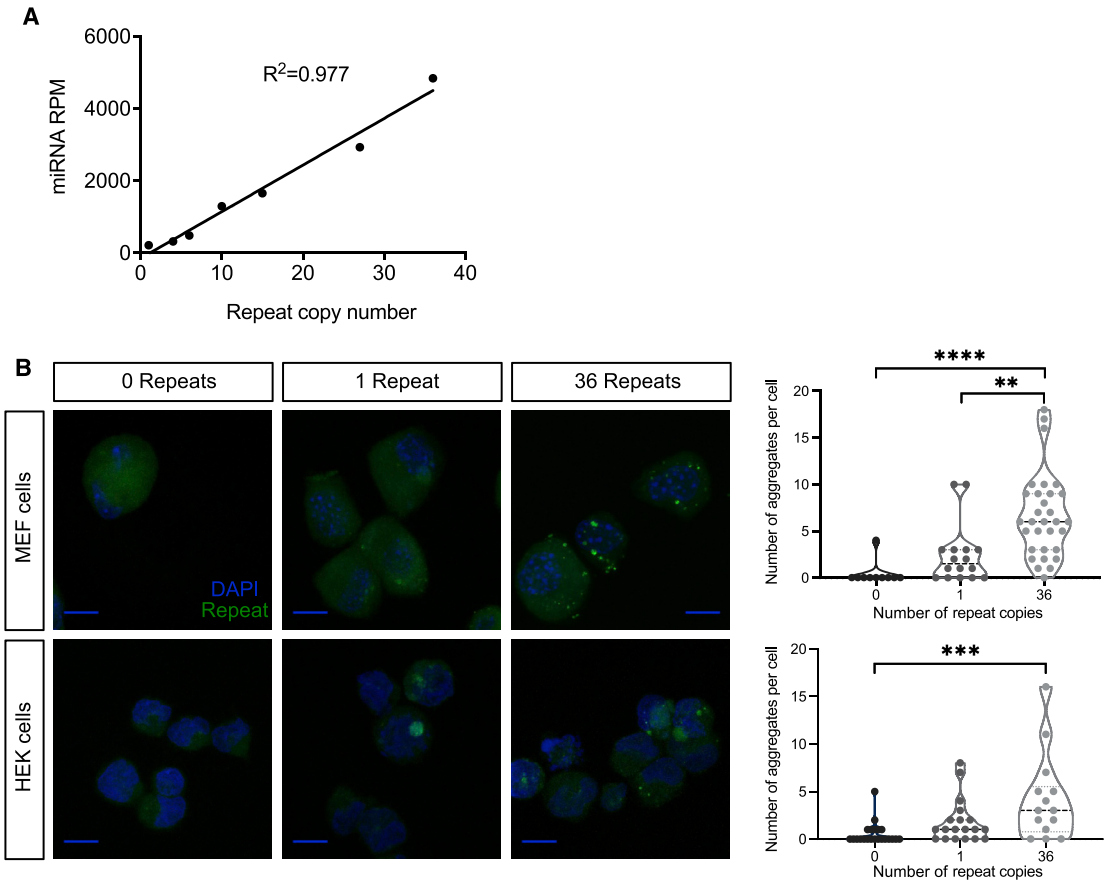

Figure 7. Functional Consequences of *WDR7* Repeats

(A) Length of *WDR7* VNTR expressed in HEK293 cells plotted against the normalized microRNAs produced from the *WDR7* hairpin, as determined by small RNA sequencing. Linear regression gives $R^2 = 0.977$. RPM is reads per million.

(B) RNA FISH probes targeting the *WDR7* repeat in MEF cells or HEK293 cells transfected with constructs containing 0 (untransfected), 1, or 36 copies of the repeat. Scale bar is 10 μm. Quantification of speckles per cell is given at right. p values were determined by a Kruskal-Wallis test (which gave $p < 0.0001$ for MEF cells and $p = 0.0009$ for HEK293 cells), followed by Dunn's multiple comparisons. **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

not just *Homo sapiens*. Moreover, it is one of a very rare set of *ab initio* repeats that are not precipitated by a repetitive element.[39] Yet, once the repeat expanded from one copy in primates to two or more in humans, widespread expansion occurred (Figure 1C). The two other large VNTRs characterized so far by long-read sequencing appear less biased in orientation than the *WDR7* VNTR. The *CACNA1C* repeat has variability primarily in two regions in the middle of the repeat.[5] The *ABCA7* repeat is dynamically expanded in non-human primates and has much more internal variability.[4]

Even for individuals who share the same repeat length, the internal arrangement of repeat units is often quite different (Figures 3 and 4C). Therefore, genotyping this repeat alone could theoretically allow for identifying individual relatedness, especially as it does not show evidence of generation-to-generation instability in a large family with ALS (Figure 4C). The enrichment and depletion of certain repeat units in various super-populations could also allow for approximating the population origin of a DNA sample, simply by observing this one VNTR. Indeed, the clustering of similar repeat patterns is reminiscent of mitochondrial or Y chromosome haplogroups (that is, groups that share a common ancestor; Figure 3). Greater sequence coverage of more diverse human populations can further inform us about the repeat units that are specific to individual populations.

The strict patterning of the nucleotide variation in the *WDR7* repeat—only six locations change out of 69 in >99% of repeat units—suggests that the sequence is conserved as a consequence of function, or mechanism of expansion, or both. One consideration is that the repeat is proximal to a DNase hypersensitivity site (Figure 1B), which may impart some selective pressure to preserve sequence specificity. We were also surprised to identify so many different repeat units present in the Denisovan genome, which overlap in sequence composition with alleles of a modern-day individual. Our findings in both Neanderthals and Denisovans suggest that the origins of the expansion pre-date modern humans (Figure 5). The ancient origin of the repeat is consistent with the generation-to-generation stability observed in the *ALS3* pedigree. Repeat length variability is a common feature of STRs,

most notably CAG repeats in Huntington disease; however, here we observe a fixed length of the repeat across multiple generations in a pedigree, suggesting that expansions in larger VNTRs are rare, stepwise events (Figure 4C).

Several types of errors can occur during DNA replication. One possible type of error that could explain the pattern of expansion that we observe for *WDR7* is template switching. Template switching occurs when the nascent leading strand switches from its template to the nascent lagging strand, because the 3′ end of the nascent leading strand is complementary to sequences on the nascent lagging strand. This type of replication error is rare, but has been proposed to contribute to tandem repeat expansions.[49,50] It occurs mainly when a repetitive tract is sufficiently long, especially longer than one Okazaki fragment, and leads to an expansion roughly the size of one Okazaki fragment: ~140 bp in primates.[2] Template switching is therefore consistent with our observation that the repeat expands in the 3′ to 5′ direction, and in units of two, which together equal 138 bp. The hairpin we observe for the RNA is also predicted to occur in single-stranded DNA, which could help explain why the repeat is being amplified.

WD repeat domains are found in >250 proteins with diverse functions, but which generally facilitate assembly of multimeric protein complexes. WDR7 is also known as rabconnectin-3β and associates with guanine exchange factors and GTPase-activating protein to facilitate RAB3-mediated recruitment and release of synaptic vesicles.[22] RAB3 is involved in synaptic vesicle exocytosis and recycling, which facilitates the tightly regulated process of calcium flux and neurotransmitter release.[23] This repeat expansion may impair the role that *WDR7* plays in synaptic function, thus suggesting its relevance to neurodegenerative disease.

The essential role of *WDR7* in synaptic health is reflected by the fact that it is intolerant to loss-of-function variants (pLI = 1). Indeed, an individual with a large 18q deletion that included part of *WDR7* experienced significant developmental delay and dysmorphic features.[41] In addition, there is a human accelerated region in the 20th intron of *WDR7* (this repeat is in the 27th), as well as nearby in the intergenic region before *WDR7*.[51] Similarly, *WDR7* exhibits human-specific downregulation in the brain, as a target of miR-941.[52] The microRNA that is produced from the *WDR7* repeat itself, coming from a human-specific expansion, could also have targets that are human specific—though this fact eliminates the ability to use conservation as a parameter for identifying possible microRNA targets, thus producing a large dataset of predicted targets (Table S2).

WDR7 repeat expansions may act to modify disease in ALS, in a manner similar to *ATXN2*, where intermediate CAG repeat expansions numbering 27 to 33 (but not >34, which causes SCA2)[12,13] are enriched in individuals with ALS. This finding supports the idea that tandem repeats may account for some of the missing heritability in disease.[20] Each *WDR7* repeat has predicted binding sites for ALS-related RNA-binding proteins TDP-43 (UGUG) and heterogeneous nuclear ribonucleoprotein A1 (HNRNPA1 [MIM: 164017])[53] (UAGGGA).[34,54] Recent reports also show that the ALS-related RNA-binding protein FUS preferentially associates with exposed loops of stem-loop hairpin RNAs.[55,56] The exposed loop present in the *WDR7* tandem repeat may act as a binding site for multiple molecules of FUS, and therefore as an RNA species that facilitates FUS intracellular aggregation. Additionally, cross-linking and immunoprecipitation (CLIP) data also suggest that TDP-43 interacts with a binding site adjacent to a neural-specific exon 17 of *WDR7*,[57] which is about 250 kb from the location of the repeat expansion in intron 27. It is possible that long-range chromatin interactions may take place between the repeat and alternative exon 17. Furthermore, RNA editing events may strengthen certain interactions. For instance, one of the observed A-to-G editing events alters a sequence from UAUGUG to UGUGUG, which is an even stronger TDP-43 interactor. Finally, TDP-43, FUS, and HNRNPA1 are all involved in microRNA processing,[58–61] which could influence microRNAs produced from the *WDR7* repeat.

VNTRs are some of the most polymorphic and mutable parts of the genome. The VNTR described here is particularly large among VNTRs, and remarkably variable in both size and internal sequence. Furthermore, it is human specific and among a select group of VNTRs that arose *ab initio* in the human lineage. Its complementary sequence forms a stable hairpin, which in turn has the potential to form microRNAs, and though intronic, it appears to be exported to the cytoplasm, where it aggregates. It likely expands through a combination of template switching and duplication events, which first occurred in ancient human genomes. Higher repeat copy number is significantly associated with cases of ALS, suggesting that it contributes to ALS susceptibility. Together, our detailed interrogation of this VNTR demonstrates the value of high-depth, long-read sequencing of human-specific repetitive regions that expand in the genome.

Data and Code Availability

Data for Neanderthal and Denisovan genomes can be accessed from ENA archives ERP002097 and ERP001519, respectively. RNA sequence reads for calculating RNA editing are available from synapse.org with accession number syn7416949. Non-human primate data are available from Zenodo with accession number 10.5281/zenodo.3401477. The remaining data are not publicly available due to institutional ethics restrictions.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.07.004.

Acknowledgments

This work is supported by the Robert F. Schoeni Award for Research from Ann Arbor Active Against ALS (to P.N.V.), the National

Institute of General Medical Sciences (5T32GM007454-38 to M.M.C.), a Frederick Banting & Charles Best Doctoral Scholarship (FRN 159279 to J.P.R.), the US National Institutes of Health (NIH) R01DK078424 (to M.A.K.), and NIH HG010169 (to E.E.E.). E.E.E. is an investigator of the Howard Hughes Medical Institute. We acknowledge the support of the ALS Association, the ALS Society of Canada, Brain Canada, and the Canadian Institutes of Health Research. This publication was supported in part by data provided by the Answer ALS Consortium – administered by the Robert Packard Center for ALS at Johns Hopkins University. In addition, data for this study were prepared, archived, and distributed by the National Institute on Aging Alzheimer Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689), funded by the National Institute on Aging. We would like to thank Dr. Jason Underwood for his valuable insight into SMRT sequencing strategies, as well as all individuals who donated biospecimens for their willingness to contribute to scientific research.

Declaration of Interests

A.D.G. has served as a consultant for Aquinnah Pharmaceuticals, Prevail Therapeutics, and Third Rock Ventures and is a scientific founder of Maze Therapeutics. E.E.E. is on the scientific advisory board of DNAnexus. All other authors declare no competing interests.

Received: April 23, 2020
Accepted: July 8, 2020
Published: August 3, 2020

Web Resources

The 1000 Genomes Project, https://www.internationalgenome.org/
AMP-AD Knowledge Portal, https://www.synapse.org/
Answer ALS, https://www.answerals.org/
Cell Profiler, https://cellprofiler.org/
European Nucleotide Archive, http://www.ebi.ac.uk/ena
FastTree, http://www.microbesonline.org/fasttree/
gnomAD, https://gnomad.broadinstitute.org/
NIAGADS, https://www.niagads.org/
OMIM, https://omim.org/
RNAfold, http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi
Synapse, https://www.synapse.org/
TargetScan, http://www.targetscan.org/
VCFtoTree, https://github.com/duoduoo/VCFtoTree_3.0.0
Web Logo, https://weblogo.berkeley.edu/
Zenodo, https://zenodo.org/

References

1. Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet. *19*, 286–298.
2. Pearson, C.E., Nichol Edamura, K., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. *6*, 729–742.
3. Todd, P.K., and Paulson, H.L. (2010). RNA-mediated neurodegeneration in repeat expansion disorders. Ann. Neurol. *67*, 291–300.
4. De Roeck, A., Duchateau, L., Van Dongen, J., Cacace, R., Bjerke, M., Van den Bossche, T., Cras, P., Vandenberghe, R., De Deyn, P.P., Engelborghs, S., et al.; BELNEU Consortium (2018). An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. Acta Neuropathol. *135*, 827–837.
5. Song, J.H.T., Lowe, C.B., and Kingsley, D.M. (2018). Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. Am. J. Hum. Genet. *103*, 421–430.
6. Oksenberg, N., Stevison, L., Wall, J.D., and Ahituv, N. (2013). Function and regulation of AUTS2, a gene implicated in autism and human evolution. PLoS Genet. *9*, e1003221.
7. Srinivasan, S., Bettella, F., Frei, O., Hill, W.D., Wang, Y., Witoelar, A., Schork, A.J., Thompson, W.K., Davies, G., Desikan, R.S., et al. (2018). Enrichment of genetic markers of recent human evolution in educational and cognitive traits. Sci. Rep. *8*, 12585.
8. Srinivasan, S., Bettella, F., Mattingsdal, M., Wang, Y., Witoelar, A., Schork, A.J., Thompson, W.K., Zuber, V., Winsvold, B.S., Zwart, J.A., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, The International Headache Genetics Consortium (2016). Genetic Markers of Human Evolution Are Enriched in Schizophrenia. Biol. Psychiatry *80*, 284–292.
9. Nithianantharajah, J., and Hannan, A.J. (2007). Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. BioEssays *29*, 525–535.
10. DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J., Nicholson, A.M., Finch, N.A., Flynn, H., Adamson, J., et al. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron *72*, 245–256.
11. Renton, A.E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J.R., Schymick, J.C., Laaksovirta, H., van Swieten, J.C., Myllykangas, L., et al.; ITALSGEN Consortium (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron *72*, 257–268.
12. Pulst, S.M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X.N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A., et al. (1996). Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat. Genet. *14*, 269–276.
13. Elden, A.C., Kim, H.J., Hart, M.P., Chen-Plotkin, A.S., Johnson, B.S., Fang, X., Armakola, M., Geser, F., Greene, R., Lu, M.M., et al. (2010). Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. Nature *466*, 1069–1075.
14. Rosen, D.R., Siddique, T., Patterson, D., Figlewicz, D.A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J.P., Deng, H.X., et al. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature *362*, 59–62.
15. Kabashi, E., Valdmanis, P.N., Dion, P., Spiegelman, D., McConkey, B.J., Vande Velde, C., Bouchard, J.P., Lacomblez, L., Pochigaeva, K., Salachas, F., et al. (2008). TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. Nat. Genet. *40*, 572–574.
16. Sreedharan, J., Blair, I.P., Tripathi, V.B., Hu, X., Vance, C., Rogelj, B., Ackerley, S., Durnall, J.C., Williams, K.L., Buratti, E.,

et al. (2008). TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. Science *319*, 1668–1672.
17. Kwiatkowski, T.J., Jr., Bosco, D.A., Leclerc, A.L., Tamrazian, E., Vanderburg, C.R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E.J., Munsat, T., et al. (2009). Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. Science *323*, 1205–1208.
18. Vance, C., Rogelj, B., Hortobágyi, T., De Vos, K.J., Nishimura, A.L., Sreedharan, J., Hu, X., Smith, B., Ruddy, D., Wright, P., et al. (2009). Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. Science *323*, 1208–1211.
19. Chia, R., Chiò, A., and Traynor, B.J. (2018). Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. Lancet Neurol. *17*, 94–102.
20. Hannan, A.J. (2010). Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. Trends Genet. *26*, 59–65.
21. Kenna, K.P., McLaughlin, R.L., Byrne, S., Elamin, M., Heverin, M., Kenny, E.M., Cormican, P., Morris, D.W., Donaghy, C.G., Bradley, D.G., and Hardiman, O. (2013). Delineating the genetic heterogeneity of ALS using targeted high-throughput sequencing. J. Med. Genet. *50*, 776–783.
22. Kawabe, H., Sakisaka, T., Yasumi, M., Shingai, T., Izumi, G., Nagano, F., Deguchi-Tawarada, M., Takeuchi, M., Nakanishi, H., and Takai, Y. (2003). A novel rabconnectin-3-binding protein that directly binds a GDP/GTP exchange protein for Rab3A small G protein implicated in Ca(2+)-dependent exocytosis of neurotransmitter. Genes Cells *8*, 537–546.
23. Schlüter, O.M., Schmitz, F., Jahn, R., Rosenmund, C., and Südhof, T.C. (2004). A complete genetic analysis of neuronal Rab3 function. J. Neurosci. *24*, 6629–6637.
24. Hand, C.K., Khoris, J., Salachas, F., Gros-Louis, F., Lopes, A.A., Mayeux-Portas, V., Brewer, C.G., Brown, R.H., Jr., Meininger, V., Camu, W., and Rouleau, G.A. (2002). A novel locus for familial amyotrophic lateral sclerosis, on chromosome 18q. Am. J. Hum. Genet. *70*, 251–256.
25. Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.
26. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.
27. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.
28. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
29. McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B.A., Karhohs, K.W., Doan, M., Ding, L., Rafelski, S.M., Thirstrup, D., et al. (2018). CellProfiler 3.0: Next-generation image processing for biology. PLoS Biol. *16*, e2005970.
30. Valdmanis, P.N., Kim, H.K., Chu, K., Zhang, F., Xu, J., Munding, E.M., Shen, J., and Kay, M.A. (2018). miR-122 removal in the liver activates imprinted microRNAs and enables more effective microRNA-mediated gene repression. Nat. Commun. *9*, 5321.
31. Course, M.M., Gudsnuk, K., and Valdmanis, P.N. (2019). A Complete Pipeline for Isolating and Sequencing MicroRNAs, and Analyzing Them Using Open Source Tools. J. Vis. Exp. (150).
32. Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.
33. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.
34. Agarwal, V., Bell, G.W., Nam, J.W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife *4*, 4.
35. Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.
36. Xu, D., Jaber, Y., Pavlidis, P., and Gokcumen, O. (2017). VCFtoTree: a user-friendly tool to construct locus-specific alignments and phylogenies from thousands of anthropologically relevant genome sequences. BMC Bioinformatics *18*, 426.
37. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.
38. Wang, M., Beckmann, N.D., Roussos, P., Wang, E., Zhou, X., Wang, Q., Ming, C., Neff, R., Ma, W., Fullard, J.F., et al. (2018). The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci. Data *5*, 180185.
39. Sulovari, A., Li, R., Audano, P.A., Porubsky, D., Vollger, M.R., Logsdon, G.A., Warren, W.C., Pollen, A.A., Chaisson, M.J.P., Eichler, E.E.; and Human Genome Structural Variation Consortium (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc. Natl. Acad. Sci. USA *116*, 23243–23253.
40. Nagano, F., Kawabe, H., Nakanishi, H., Shinohara, M., Deguchi-Tawarada, M., Takeuchi, M., Sasaki, T., and Takai, Y. (2002). Rabconnectin-3, a novel protein that binds both GDP/GTP exchange protein and GTPase-activating protein for Rab3 small G protein family. J. Biol. Chem. *277*, 9629–9632.
41. Sproviero, W., Shatunov, A., Stahl, D., Shoai, M., van Rheenen, W., Jones, A.R., Al-Sarraj, S., Andersen, P.M., Bonini, N.M., Conforti, F.L., et al. (2017). ATXN2 trinucleotide repeat length correlates with risk of ALS. Neurobiol. Aging *51*, 178.e1–178.e9.
42. Schüle, B., McFarland, K.N., Lee, K., Tsai, Y.C., Nguyen, K.D., Sun, C., Liu, M., Byrne, C., Gopi, R., Huang, N., et al. (2017). Parkinson's disease associated with pure *ATXN10* repeat expansion. NPJ Parkinsons Dis. *3*, 27.
43. Cortese, A., Simone, R., Sullivan, R., Vandrovcova, J., Tariq, H., Yau, W.Y., Humphrey, J., Jaunmuktane, Z., Sivakumar, P., Polke, J., et al. (2019). Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat. Genet. *51*, 649–658.
44. Ishiura, H., Doi, K., Mitsui, J., Yoshimura, J., Matsukawa, M.K., Fujiyama, A., Toyoshima, Y., Kakita, A., Takahashi, H., Suzuki, Y., et al. (2018). Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nat. Genet. *50*, 581–590.
45. Zu, T., Liu, Y., Bañez-Coronel, M., Reid, T., Pletnikova, O., Lewis, J., Miller, T.M., Harms, M.B., Falchook, A.E.,

Subramony, S.H., et al. (2013). RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. Proc. Natl. Acad. Sci. USA *110*, E4968–E4977.
46. Boutz, P.L., Bhutkar, A., and Sharp, P.A. (2015). Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev. *29*, 63–80.
47. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
48. Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R., Ramaswami, G., Ariyoshi, K., et al.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; and Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz (2017). Dynamic landscape and regulation of RNA editing in mammals. Nature *550*, 249–254.
49. Cherng, N., Shishkin, A.A., Schlager, L.I., Tuck, R.H., Sloan, L., Matera, R., Sarkar, P.S., Ashizawa, T., Freudenreich, C.H., and Mirkin, S.M. (2011). Expansions, contractions, and fragility of the spinocerebellar ataxia type 10 pentanucleotide repeat in yeast. Proc. Natl. Acad. Sci. USA *108*, 2843–2848.
50. Shishkin, A.A., Voineagu, I., Matera, R., Cherng, N., Chernet, B.T., Krasilnikova, M.M., Narayanan, V., Lobachev, K.S., and Mirkin, S.M. (2009). Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. Mol. Cell *35*, 82–92.
51. Doan, R.N., Bae, B.I., Cubelos, B., Chang, C., Hossain, A.A., Al-Saad, S., Mukaddes, N.M., Oner, O., Al-Saffar, M., Balkhy, S., et al.; Homozygosity Mapping Consortium for Autism (2016). Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell *167*, 341–354.e12.
52. Hu, H.Y., He, L., Fominykh, K., Yan, Z., Guo, S., Zhang, X., Taylor, M.S., Tang, L., Li, J., Liu, J., et al. (2012). Evolution of the human-specific microRNA miR-941. Nat. Commun. *3*, 1145.
53. Kim, H.J., Kim, N.C., Wang, Y.D., Scarborough, E.A., Moore, J., Diaz, Z., MacLea, K.S., Freibaum, B., Li, S., Molliex, A., et al. (2013). Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. Nature *495*, 467–473.
54. Ray, D., Kazan, H., Chan, E.T., Peña Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat. Biotechnol. *27*, 667–670.
55. Tan, L., Yu, J.T., and Tan, L. (2015). Causes and Consequences of MicroRNA Dysregulation in Neurodegenerative Diseases. Mol. Neurobiol. *51*, 1249–1262.
56. Loughlin, F.E., Lukavsky, P.J., Kazeeva, T., Reber, S., Hock, E.M., Colombo, M., Von Schroetter, C., Pauli, P., Clery, A., Muhlemann, O., et al. (2019). The Solution Structure of FUS Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape Specificity. Mol. Cell *73*, 490–504.
57. Rogelj, B., Easton, L.E., Bogu, G.K., Stanton, L.W., Rot, G., Curk, T., Zupan, B., Sugimoto, Y., Modic, M., Haberman, N., et al. (2012). Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. Sci. Rep. *2*, 603.
58. Morlando, M., Dini Modigliani, S., Torrelli, G., Rosa, A., Di Carlo, V., Caffarelli, E., and Bozzoni, I. (2012). FUS stimulates microRNA biogenesis by facilitating co-transcriptional Drosha recruitment. EMBO J. *31*, 4502–4510.
59. Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235–240.
60. Guil, S., and Cáceres, J.F. (2007). The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nat. Struct. Mol. Biol. *14*, 591–596.
61. Michlewski, G., and Cáceres, J.F. (2010). Antagonistic role of hnRNP A1 and KSRP in the regulation of let-7a biogenesis. Nat. Struct. Mol. Biol. *17*, 1011–1018.