Human Genetics (2019) 138:661–672
https://doi.org/10.1007/s00439-019-02017-5

ORIGINAL INVESTIGATION

De novo emergence and potential function of human-specific tandem repeats in brain-related loci

Kwondo Kim[1,2] · Sohyun Bang[1,2] · DongAhn Yoo[1] · Heebal Kim[1,2,3,4] · Shunsuke Suzuki[4,5]

Received: 2 November 2018 / Accepted: 16 April 2019 / Published online: 8 May 2019

Abstract

Tandem repeats (TRs) are widespread in the genomes of all living organisms. In eukaryotes, they are found in both coding and noncoding regions and have potential roles in the regulation of cellular processes such as transcription, translation and in the modification of protein structure. Recent studies have highlighted TRs as a key regulator of gene expression and a potential contributor to human evolution. Thus, TRs are emerging as an important source of variation that can result in differential gene expression at intra- and inter-species levels. In this study, we performed a genome-wide survey to identify TRs that have emerged in the human lineage. We further examined these loci to explore their potential functional significance for human evolution. We identified 152 human-specific TR (HSTR) loci containing a repeat unit of more than ten bases, with most of them showing a repeat count of two. Gene set enrichment analysis showed that HSTR-associated genes were associated with biological functions in brain development and synapse function. In addition, we compared gene expression of human HSTR loci with orthologues from non-human primates (NHP) in seven different tissues. Strikingly, the expression level of HSTR-associated genes in brain tissues was significantly higher in human than in NHP. These results suggest the possibility that de novo emergence of TRs could have resulted in altered gene expression in humans within a short-time frame and contributed to the rapid evolution of human brain function.

List of abbreviations

TR	Tandem repeat
bp	Base pair
HSTR	Human-specific tandem repeat
NHP	Non-human primates
HTT	Huntingtin
STR	Short tandem repeat
GWAS	Genome-wide association study
HTR	Human tandem repeat
IS	Intervening sequence
TRF	Tandem repeat finder
RBH	Reciprocal best hit
VNTR	Variable number tandem repeat
CDS	Coding sequence
GO	Gene ontology
OMIM	Online Mendelian inheritance in man
logFC	log2-fold change

Kwondo Kim and Sohyun Bang contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00439-019-02017-5) contains supplementary material, which is available to authorized users.

✉ Heebal Kim
heebal@snu.ac.kr

✉ Shunsuke Suzuki
ssuzuki@shinshu-u.ac.jp

Kwondo Kim
bigkd@snu.ac.kr

Sohyun Bang
sohyunbk@gmail.com

DongAhn Yoo
dongahn.yoo@gmail.com

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea

[2] C&K Genomics, C-1008, H Businesspark, 26, Beobwon-ro 9-gil, Songpa-gu, Seoul, Republic of Korea

[3] Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea

[4] Department of Interdisciplinary Genome Sciences and Cell Metabolism, Institute for Biomedical Sciences, ICCER, Shinshu University, 8304 Minami-Minowa, Kami-Ina, Nagano 399-4598, Japan

[5] Department of Agricultural and Life Sciences, Faculty of Agriculture, Shinshu University, 8304 Minami-Minowa, Kami-Ina, Nagano 399-4598, Japan

DEHG	Differentially expressed HSTR-associated gene
HSP	High-scoring segment pair
SRA	Sequence read archive
RPKM	Reads per kilobase per million
GEO	Gene expression omnibus

Introduction

Tandem repeats (TR) are DNA segments, where a pattern of one or more nucleotides is repeated in the directly adjacent sequence (Lander et al. 2001). TRs are also known as satellite DNA, since they were initially detected as satellite bands in density-gradient centrifugal DNA separation. Although there is no clear definition, TRs with a unit ranging from 1 to 9 nucleotides in length are generally referred to as microsatellites, while those with a longer repeat unit ($\geq$ 10 nucleotides) are known as minisatellites (Gemayel et al. 2010). Typical minisatellites range in size up to 100 bp (Vergnaud and Denoeud 2000); however, some minisatellites up to 500 bp have been identified in the human genome (Lander et al. 2001).

TRs are ubiquitous in eukaryotic genomes. In the human genome, TRs comprise 3% of its length (Lander et al. 2001). TRs are found in both coding and noncoding regions and are often associated with alterations to cellular processes including transcription, translation as well as altering protein structure, depending on their locations (Gemayel et al. 2010). Huntington's disease is a classic example, where affected individuals have a larger number of CAG sequence repeats in the first exon of their *HTT* (huntingtin) gene (Walker 2007). The expansion of TRs in the protein-coding region of *HTT* gene gives rise to abnormal protein, which gradually damages cells in the brain (Gymrek et al. 2017; Usdin 2008b). TRs located in the noncoding region are also able to modify the binding of transcription factors. TRs in the promoter region of the *PIG3* gene directly interact with the P53 transcription factor to mediate its induction (Contente et al. 2002). Expression of the dopamine transporter gene (*DAT1*) is similarly regulated by the number of TRs located in the 3′ UTR (Mill et al. 2002).

Recent studies have highlighted the relationship between TR variation and gene expression (Gymrek et al. 2016; Sonay et al. 2015). Gymrek et al. surveyed genome-wide short tandem repeats (STRs) and identified 2060 STRs within promoter regions, which were significantly associated with altered expression of the neighboring genes. By analyzing genome-wide association studies (GWAS), they predicted that TRs were associated with various clinical conditions (Gymrek et al. 2016). Another study by Sonay et al. showed that polymorphic TRs in promoters alter gene expression across human and primate in several different tissues. This study also suggested that TRs are associated with biological processes such as stimulus detection, sensory perception, and skin development, which are related to the evolution of human cognitive traits and adaptation to new environments (Sonay et al. 2015).

Much research has focused on genomic alterations contributing to the unique evolutionary phenotypes in human (O'bleness et al. 2012a). These studies suggest that alterations to the DNA sequence are likely to have contributed to the evolution of human-specific phenotypes, prompting us to explore the potential relevance of human-specific TRs to the evolution of human-specific traits (Dumas et al. 2012; Enard et al. 2009; Suzuki et al. 2018).

Mutation rates of TRs are 10 to 100,000 times higher than that in other parts of the genome. Variation in these regions is generated by strand slippage and homologous recombination errors rather than point mutations (Legendre et al. 2007). We hypothesized that under neutral condition, TRs are unlikely to be fixed in length and sequence composition. Therefore, de novo TRs that are fixed could indicate a positive selection signature and loci that are related to advantageous traits.

In this study, we extracted the loci, where putative de novo TRs emerged and are fixed, specifically in the human lineage by comparing the genome sequences of humans with three non-human primates (NHP) (Chimpanzee, Orangutan, and Gorilla). Then, we analyzed their potential association with gene expression and their functional role in the evolution of human-specific traits.

Results

Tandem repeats present only in the human lineage

To retrieve as many candidates as possible, we started with all possible TRs loci, detected by tandem repeat finder (TRF) on the human reference genome (GRCh38), which resulted in approximately one million loci (1,014,188) (Fig. 1a). We next filtered for TR loci that have flanking sequences with orthologous blast hits in each NHP genome. This resulted in the elimination of 255,465/238,644/284,520 in the chimpanzee, gorilla, and orangutan, respectively (loci were eliminated, since they had *e* value above 10; Table 1, 1. Blast hit). The remaining loci (600,132/606,841/598,864) were then filtered for orthologous queries that covered the entire target sequence (Table 1, 2. Coverage). A further, 136,167/144,338/116,802 loci were removed using reciprocal best hits (RBH) analysis between flanking sequences of human and NHP TR loci (Table 1, 3. RBH). Only a few loci (386/375/1220) had non-matching flanking sequences in both 3′ and 5′ directions and were removed (Table 1, 4. Pair). Consequently, 22,038, 23,999, and 13,472 loci in

the chimpanzee, gorilla, and orangutan, respectively, with shared orthologues between human and NHPs were retained.

After removing the loci that either have repeat count less than 2 (e.g., 1.5) or ambiguous base compositions (e.g., "N"), 17,840/19,287/10,326 orthologous loci remained. Among them, 23.8%/26.2%/30.5% were not detected as HSTRs after filtering for length and identity (Table 2), 69.8%/66.2%/56.4% due to length criterion (Table 2, length), and 2.7%/3.9%/8.2% due to identity criterion (Table 2, identity). Finally, 925/1131/866 loci remained in each human vs NHPs comparisons (Table 2, HSTR), and 152 loci that were shared across all three comparisons were considered as HSTR loci (Fig. 1c).

We then compared the HSTRs with the total TRs detected by TRF, in terms of unit length, number of repeats, total length, and percentage of match (Fig. 2). The mean unit length of HSTRs (30.3 bp) was longer than that of total TRs (28.4 bp), and showed a length distribution from 11 bp to 219 bp. This shows that HSTRs did not include TRs that have a unit length shorter than 10 bp. In the process of determining HSTR, 98.2% of microsatellites and 64.1% of all TRs (in average of three NHPs) were removed due to length

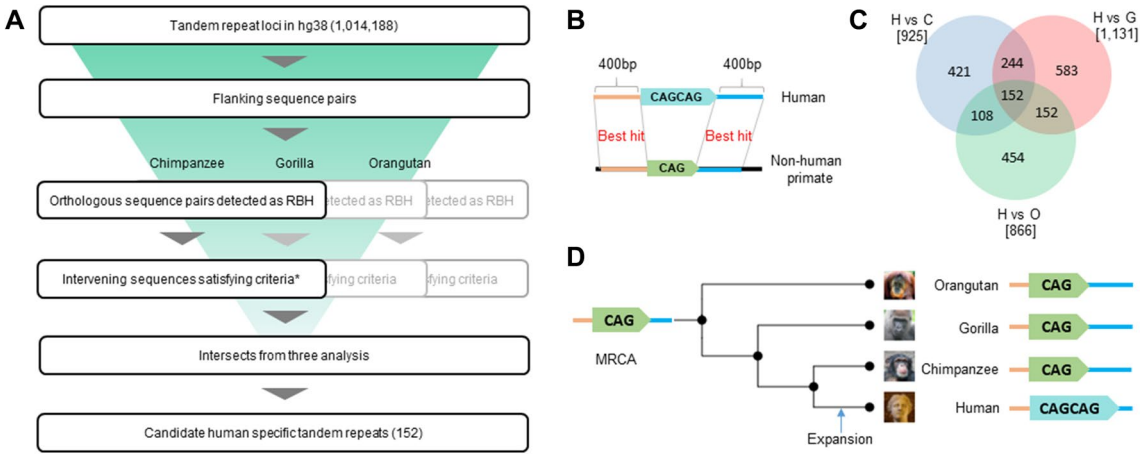

Fig. 1 Workflow for detecting human-specific tandem repeat. **a** Diagram representing whole analysis steps to detect HSTR. TRs were filtered by the criterion of each step (see "Materials and methods"). **b** Schematic of method for detecting HSTRs. The concept of method was modified from a previous study (Sen et al. 2006). **c** Number of TRs detected in each comparison (human vs chimpanzee, human vs gorilla, and human vs orangutan). **d** Type of TRs (HSTR) we expect to identify in this study

Table 1 Number of loci satisfying the criteria for orthologous flanking sequences

	1. Blast hit		2. Coverage		3. RBH		4. Pair	
	Not satisfying	Satisfying	Not satisfying	Satisfying	Not satisfying	Satisfying	Not satisfying	Satisfying
Chimpanzee	255,465	758,723	600,132	158,591	136,167	22,424	386	22,038
Gorilla	238,644	775,544	606,841	168,712	144,338	24,374	375	23,999
Orangutan	284,520	729,668	598,864	130,804	116,802	14,002	1220	13,472

The figures indicate the number of loci satisfying or not satisfying criteria in each filtering step (see "Materials and methods")

RBH reciprocal best hits

Table 2 Number of TR loci that did not pass criteria for determining HSTR

	Length and identity	Length	Identity	HSTR	Total
Chimpanzee	4252 (23.8%)	13,112 (69.8%)	476 (2.7%)	925 (4.9%)	18,765 (100%)
Gorilla	5063 (26.2%)	13,464 (66.2%)	760 (3.9%)	1131 (5.5%)	20,418 (100%)
Orangutan	3157 (30.5%)	6320 (56.4%)	849 (8.2%)	866 (7.7%)	11,192 (100%)

The figures in each column indicate the number of loci that did not pass each criteria. Length criteria are: (1) Human TR (HTR) length ≥ intervening sequence (IS) length in NHP *2 and (2) HTR length – IS length in NHP ≥ unit length of HTR. Identity selects loci with percent identity > 95% between sequence units of the HTR and IS in NHP. Table indicates the number of HSTRs that remain after filtering TRs through the previous three criteria ("Length & Identity", "Length" and "Identity")

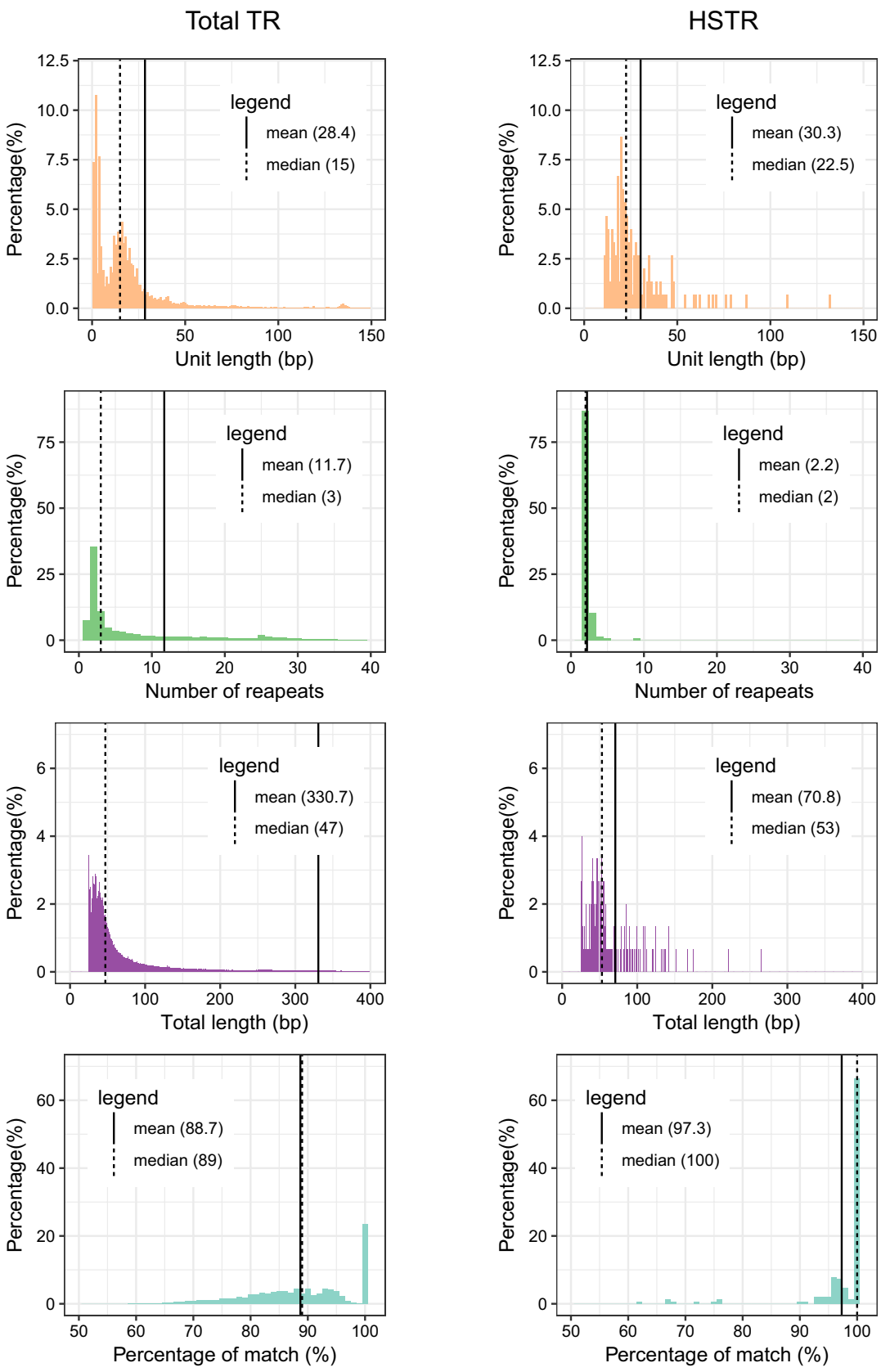

Fig. 2 Characteristics of total TRs and HSTRs. Characteristics (unit length, number of repeats, total length, and percentage of match) of total TR and HSTR

criterion (Tables 2, 3). This resulted in the complete removal of microsatellites from the final list of HSTRs shared by all three NHPs. Most HSTRs were repeated two times (mean: 2.2, median: 2), whereas total TRs were repeated more than two times (mean: 11.7, median: 3). The total TRs contained much longer sequences (mean: 330.7 bp, median: 47 bp) than HSTRs (mean: 70.8 bp, median: 53 bp), as expected, since the each total TRs contain more repeats than each HSTRs. The percentage match between repeat unit sequences was higher in HSTRs (average 97.3%) than in total TRs (average 88.7%), which was consistent in a comparison between HSTRs and control sets (see "Materials and Methods", Figure S1).

Validation and fixation of HSTRs

If the HSTRs that we have detected de novo emerged in the human lineage, they should be absent in other species as well as in the three NHPs. To assess this, we examined the presence/absence of sequences at the HSTR loci in an additional 99 species. In a multiple alignment of 100 species from the UCSC database, the HSTRs showed extremely high frequency of 'deletion' (a nucleotide that is absent in one of the species' sequence in the alignment) compared to flanking sequences (Fig. 3a, Wilcoxon rank sum test, p value < 2.2e−16). As expected, all NHPs had a deletion at the exact orthologous position (Fig. 3b, c).

It is highly likely that TRs associated with important genomic functions are fixed in the human population. We examined variations at the HSTR loci in human population data from the whole genome sequences of the 1000 genomes project (Consortium 2015). The sequence data from 24 individuals with high sequence coverage (153X on average) were used to reduce false negatives due to the lack of sequencing depth. Approximately 64% of the HSTRs' sequences were identified in all 24 individuals (95 out of total 148 HSTRs, 4 loci that were longer than the read length (> 250 bp) were excluded from the 152 HSTRs, Fig. 4a). Total 86.3% of HSTRs was identified in more than ten individuals (sum of 64.1% found in all samples, an additional 4.7% was found in 23 samples, an additional 2% in 22 samples, an additional 2.7% in 21 individuals, and 12.1% in 11–20 individuals). Only 9% of HSTRs (14 loci) were not detected in all individuals. The 24 individuals in the 1000

Table 3 Number of microsatellites (unit length ≤ 10 bp) that did not pass criteria for determining HSTR

	Length and identity	Length	Identity	HSTR	Total
Chimpanzee	32 (0.56%)	5610 (98.8%)	6 (0.1%)	26 (0.45%)	5674 (100%)
Gorilla	30 (0.51%)	5784 (98.7%)	12 (0.2%)	29 (0.49%)	5855 (100%)
Orangutan	51 (1.5%)	3242 (97.2%)	8 (0.24%)	31 (0.93%)	3332 (100%)

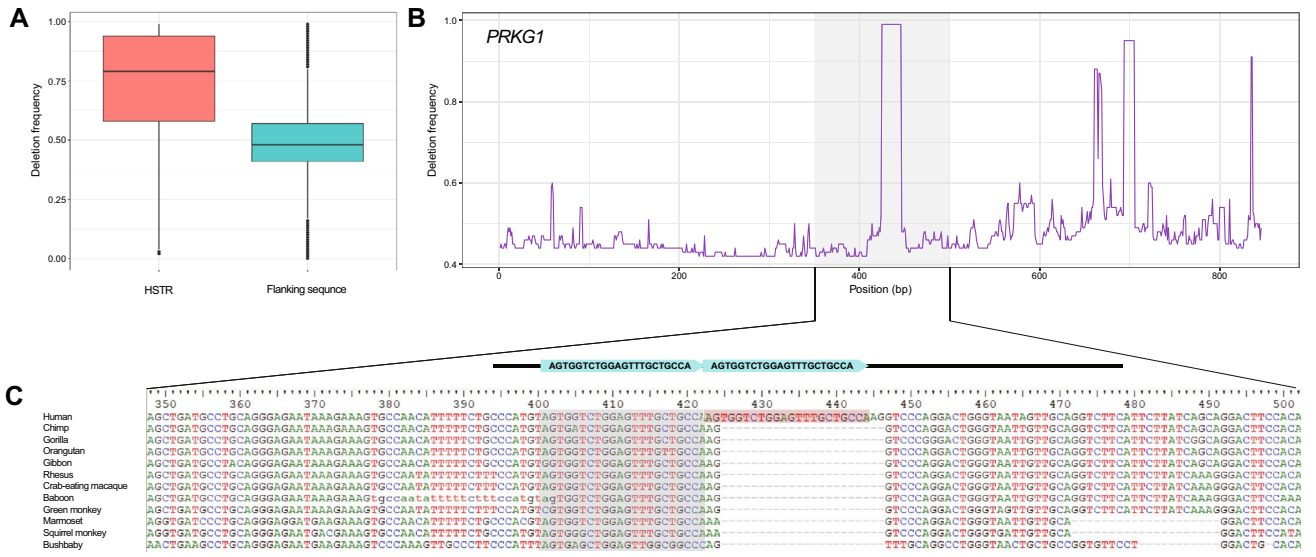

Fig. 3 Deletion frequency of identified HSTRs and flanking sequences. **a** Per base deletion frequencies across 100 vertebrate's alignments between sequence of HSTRs and Flanking sequences. The 152 HSTRs have extremely high frequency of deletions compared to flanking sequences (Wilcoxon rank sum test, p value < 2.2e−16). **b** An Example of deletion frequency trend for the HSTR in PRKG1 gene along with flanking sequences. Almost all species have deletion at the HSTR region except human. **c** Multiple sequence alignment of shaded area in (**b**). Of 100 vertebrates, only the alignment of 12 primates were shown

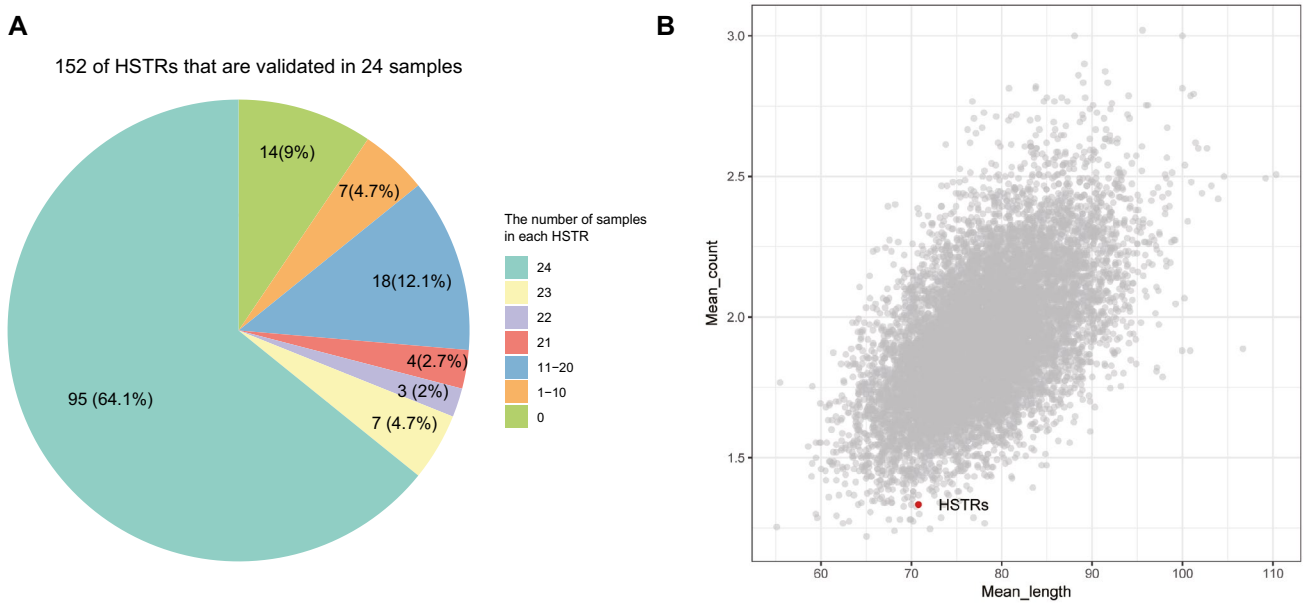

Fig. 4 Validation and fixation of HSTRs. **a** Distribution of samples that have intact HSTR sequences. **b** Mean count of common variants within HSTRs. Control sets generated by resampling 10,000 times were plotted as grey dots, and the HSTR set was plotted as a red dot. *X*- and *Y*-axes indicate mean length of TR and mean count of common variants for each set, respectively

genomes project come from diverse genetic backgrounds and represent 24 different ethnic groups. Nevertheless, 95 HSTR sequences were identified across all individuals. It is also worth noting that there may be more than 95 fixed HSTRs out of total 148 candidates, but these have been removed because of our strict criteria applied for identification of intact HSTR sequences in the 1000 genomes project data.

We next investigated the incidence of common variants retrieved from dbSNP database within HSTR regions. In general, TRs along with other types of repeat sequences have higher mutation rates than other genomic regions (Vergnaud and Denoeud 2000). We generated control sets consisting of only TR sequences with various lengths using bootstrapping and compared them with the HSTRs. The control sets showed a clear trend that longer TRs contain more variants. In contrast, HSTRs showed much lower variation than expected for their mean length compared with other TRs (Fig. 4b). In addition, separate comparisons for genic and intergenic regions showed the same tendency (Figure S2). This shows that the low incidence of variation in HSTRs is not simply attributed to their genomic location.

Enrichment of HSTR in brain-related functions

To assess potential functionality of HSTRs, we first investigated their position within human genome (Fig. 5a). Similar to a previous study (Legendre et al. 2007), we found the majority of our HSTRs mapped within genic regions (72.4%). Out of the HSTRs within genic regions, 59.9 (91), 9.9 (15), and 2.6 (4) % were located in intron, exon, or intron–exon boundaries, respectively. The proportion of genic HSTRs located in introns was greater than that for total TRs (42%). Of 15 HSTRs in exon, seven were located within the coding sequences.

Using gene annotation information, we performed gene set enrichment tests to identify potential functions of HSTRs. Out of the 152 HSTRs, 110 were assigned to at least one gene annotation, and 118 genes contained at least one HSTR. Out of the remaining 118 genes, 81 genes were assigned to Gene Ontology (GO) terms (Biological Process) and were used in gene set enrichment analysis (Fig. 5b). The most significant GO term was "receptor localization to synapse (GO:0097120)". In addition, among non-statistically significant results were several synapse-related terms; "innervation (GO:0060384)", "positive regulation of excitatory postsynaptic potential (GO:2000463)", and "synaptic vesicle exocytosis (GO:0016079)". Intriguingly, three genes (*DYNC2H1*, *PRKG1*, and *SSTR1*) belonging to the statistically significant term, "forebrain development (GO:0030900)", were found. Moreover, as the HSTRs might have long distance effects on gene regulation, we also expanded the gene list to include those within 100 kb of any gene TSS. If multiple TSS are located within 100 kb of any HSTR, the closest gene was selected. The expanded gene list consisting of 117 genes was still enriched in synapse-related or brain development, biological processes (Figure S3). 22 genes including 26 HSTRs were assigned to one of the genetic disorders in the OMIM (Online Mendelian Inheritance in Man) database, of

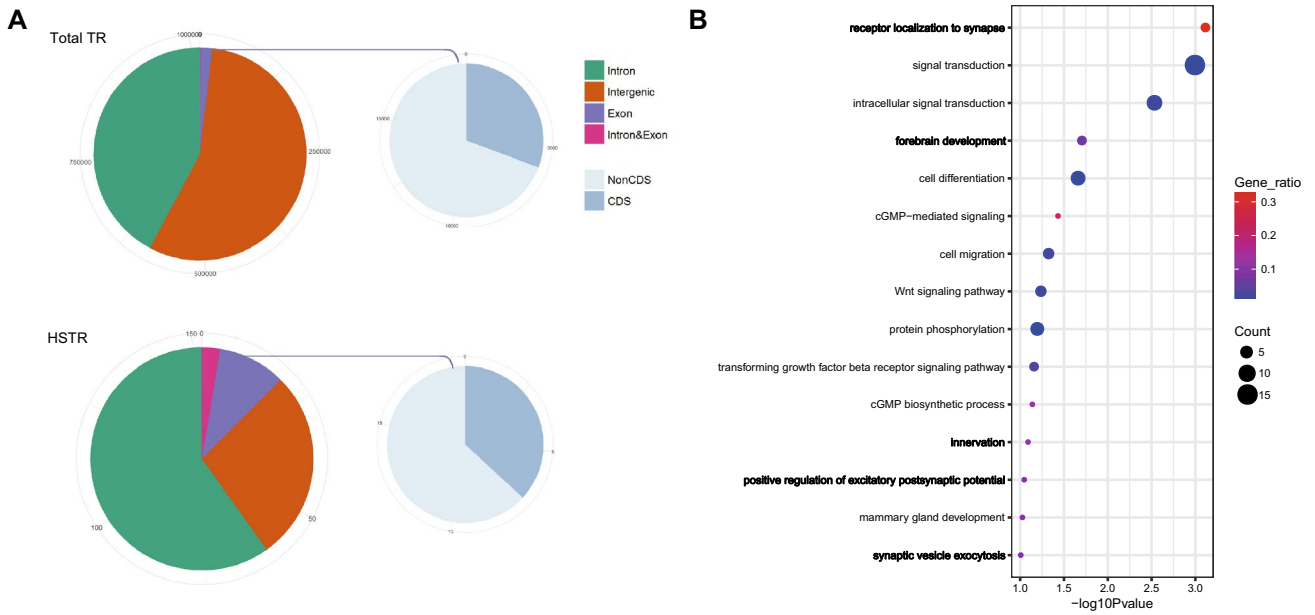

Fig. 5 Genomic location and functional enrichment of HSTRs. **a** Genomic location of total TR and HSTR. The ratio for the location of TRs (intron, intergenic, exon, and boundaries of intron and exon) was illustrated with pie chart. Among TRs located in exon, the ratio of TR in CDS (coding sequence) and untranslated region (NonCDS) was also shown. **b** Result of gene set enrichment analysis for the gene set which contains HSTRs

which six genes were related to neurodegenerative disorders (*VPS53*), mental retardation (*TNIK*, *DPP6,* and *KCNH1*), and brain-related diseases (*TRPC3* and *TDGF1*).

Brain-specific expression of HSTR-associated genes

Given that TRs can affect the expression of nearby genes (Bennett et al. 1995; Hamada et al. 1984; Streelman and Kocher 2002; Warpeha et al. 1999), we hypothesized that HSTR-associated genes would show a greater degree of differential expression levels between human and NHP compared to non-HSTR-associated genes. To examine this hypothesis, we used "log2-fold change" (logFC) as a measure of differential expression between human and NHP, and compared logFC of the HSTR-associated gene set with that of the total orthologous gene set (see "Materials and methods"). For investigating the expression of orthologues between humans and NHP, two sets of publicly available RNA-seq data were used. The first data set consists of six tissues from four species (human, chimpanzee, gorilla, and orangutan). Out of 15,508 orthologous gene sets defined in the four species, those genes which are located within 100 kb from HSTR were selected; 82 orthologues were selected and the mean of logFC was calculated. In the entire tissues, most of mean logFCs were close to zero with small deviations. The mean values of logFC distributions in each tissues were −0.033 (frontal cortex), −0.128 (cerebellum), −0.166 (liver), −0.116 (heart), −0.082 (kidney), and −0.072 (testis). Strikingly, enhancement of HSTR gene expression was only observed in two brain tissues. The frontal cortex had the highest mean logFC of HSTR-associated genes (Mean logFC: 0.204, Empirical *p* value: 0.059), and cerebellum had the second highest mean logFC (Mean logFC: 0.028, Empirical *p* value: 0.15) (Fig. 6). Increased gene expression was not observed in any other tissues with other tissues showing a decrease or similar expression (liver: −0.378, heart: −0.070, kidney: −0.089, and testis: −0.156).

Out of 16,377 orthologous gene sets defined in the second data set, those genes which are located within 100 kb from HSTR were selected the same as the first data set; 64 orthologues were selected and the mean of logFC was calculated. Similar to the first data set, the overall expression of the orthologues was higher in brain organoids from humans than chimpanzees (Mean logFC: −2.827, empirical *p* value: 0.12, Fig. 6). There were 68, differentially expressed HSTR-associated genes (DEHG) between human and NHP (29 in the frontal cortex, 18 in the cerebellum, and 21 in brain organoids, p value < 0.05). The mean logFCs of the DEHGs were 1.052, 1.062, and 1.474, respectively. Note that all the logFCs for the DEHGs were in the positive direction or zero, that is, the expressions of DEHGs are consistently higher in human than those of NHP in three brain tissues.

The results from the brain transcriptome and brain organoid data consistently show the increased expression of HSTR-associated genes specifically in brain tissues. Comparisons with non-HSTR-associated loci also support the

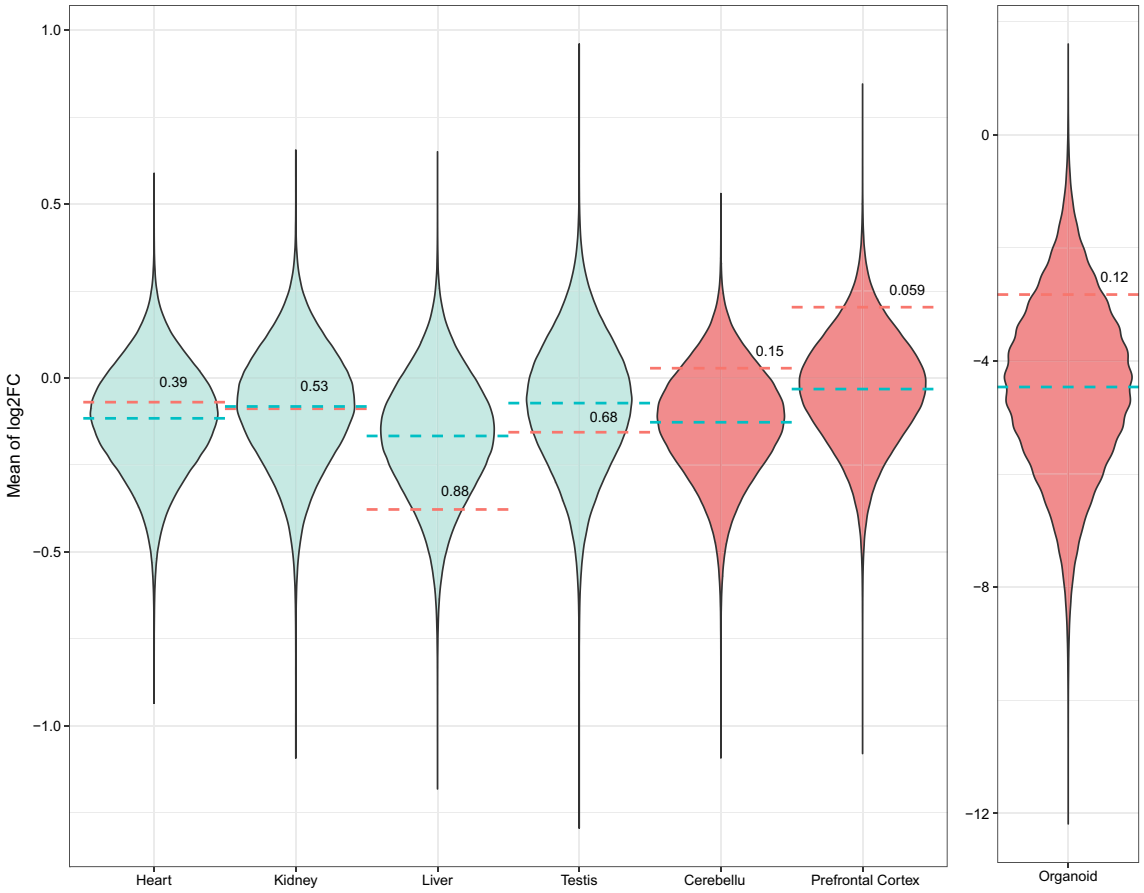

Fig. 6 Expression of HSTR genes across seven tissues. Distributions of mean logFC (log2-fold change) for random samples in seven tissues. Dashed line colored by red indicates mean logFC for total orthologous gene sets. Dashed line colored by blue indicates mean logFC for HSTR gene set

conclusion that altered expression is likely associated with HSTRs.

Potential roles of HSTRs in regulation of gene expression

The possibility of HSTRs causing altered gene expression led us to investigate their potential regulatory role. We first investigated signals of histone modification within the HSTR regions. 49 ChIP-seq data sets produced from human brain tissues were retrieved from the ENCODE database (Consortium 2007) and used for checking the potential promoter (H3K4me3 and H3K9ac) and enhancer (H3K4me1 and H3K27ac) activity of HSTR regions (Table S5). For each histone mark, we found 4 (H3K4me3), 15 (H3K9ac), 22 (H3K4me1), and 22 (H3K27ac) peaks, which covered ~ 28% of all HSTRs (Table S1). We then investigated HSTR sequences to identify potential binding motifs for transcription factors. Although the types of binding motifs were varied, ~ 35% of all HSTRs contained at least one binding motif (Table S1).

Discussion

The brain is one of the most active evolutionary sites in humans. Humans have the largest brain among extant primates with specialized neuronal connections (Hill and Walsh 2005; Sousa et al. 2017), which lead to superior abilities related to cognition and behavior such as long-term planning and the capacity to create art (Sousa et al. 2017). As such, the brain is regarded as the core component of the human identity. However, how the human brain evolved from that of closely related primates is not fully understood. Several studies suggested gene expression changes as one of the driving forces for the distinctive attributes of human brain function (Cáceres et al. 2003; Carroll 2005; Gu and Gu 2003; Khaitovich et al. 2008). However, the reason for difference in gene expression across different species is still an open question.

Recent studies have examined TR as a potential driver of altered gene expression. Their unstable nature facilitates rapid evolutionary events and can cause variations in gene expression and function within or between species (Gemayel

et al. 2010). Divergence in gene expression is observed in the presence or absence of TR using six tissues across four primates (Sonay et al. 2015). In this study, we examined the distribution and potential functional consequences of TRs that emerged de novo in the human lineage (HSTRs). The HSTRs we found showed a low number of repeats compared to total TRs in the human genome. As we did not limit the number of repeats in any of our steps for detecting HSTR, two unit repetitions in our HSTR are not likely to be caused by a systematic bias. Rather, it may be a unique characteristic of TRs that recently emerged. A previous study (Ahmed and Liang 2012) suggested that older TRs have more repeats than younger ones. In addition, the higher percentage match between repeats in the HSTRs provides further support for their recent emergence (Kumar and Subramanian 2002). In addition, the rareness of HSTR across 100 species suggests that the HSTRs not only recently emerged, but also de novo emerged after split from chimpanzee.

HSTRs were enriched within or nearby genes related to brain function. This is consistent with the previous studies (Legendre et al. 2007) examining unfiltered TRs. In addition, we observed an enrichment in synapse-related functions and showed that the presence of HSTRs was associated with increased gene expression, exclusively in human brain tissues.

Approximately 28% of HSTRs overlapped with histone marks associated with promoter or enhancer functions in brain tissues and ~ 35% of HSTR sequences contained at least one transcription factor binding motif (Table S1, Bonferroni corrected p value < 0.05). These results were consistent with a role for TRs in cis-regulatory elements such as promoters or enhancers (Gemayel et al. 2010; Usdin 2008a). Considering the genomic location of HSTRs (most of HSTRs were located in intron region), there might be other underlying mechanisms. Several previous studies suggested intron size as one factor related to splicing (Bell et al. 1998; Pai et al. 2017; Wieringa et al. 1984). Because most of our HSTRs have a relatively long unit length, the influence of TR expansion on the splicing of messenger RNAs is worth examining in evolutionary context. Moreover, among HSTRs located in exonic regions, the majority were contained in either the first or last exon. This may be due to their role in regulating gene expression as the first and last exons are related to the recognition of regulatory elements and polyadenylation processes, respectively (Bieberstein et al. 2012; Matoulkova et al. 2012).

Although more functional investigations are required, we found a large number of HSTRs which might play a role in human brain evolution (Figure S4). However, the HSTRs described in this study do not account for entire set in the human genome. We used flanking sequence similarity to detect orthologous regions in NHP; thus, we could only detect HSTRs in highly conserved regions. This caveat can explain the enrichment of HSTRs in genic regions. In addition, we only examined TRs with a fixed number of repeats; there might be HSTRs, where the number of repeats is not fixed in length, which show a defined range of repeat units that are unique to humans. For example, the DUF1220 protein domain, which is related to brain size (Dumas et al. 2012), has a human-specific copy number range (human have ~ 300 copies, while chimp and gorilla have 95–140 copies) (O'bleness et al. 2012b; Zimmer and Montgomery 2015). Hence, the list of HSTR we analyzed should be considered as a subset of the entire HSTRs. Finally, the imbalance of annotation information between human and NHP along with incomplete reference genomes limits our expression analysis, especially in defining the orthologous gene sets. More refined reference genomes and annotations might reduce the missing orthologues.

To the best of our present knowledge, this is the first study to suggest a relationship between de novo emerged HSTRs and gene expression. Moreover, our findings provide novel insights into the role of TR emergence in human brain evolution. Our computational approach can provide a useful tool for discovering additional candidate TRs responsible for phenotypic variation and help us further understand their contribution to the regulatory landscape.

Materials and methods

Detection of human-specific tandem repeat

A list of whole tandem repeats for the human genome (GRCh38) generated by TRF (Benson 1999) was downloaded from UCSC genome browser (https://genome.ucsc.edu/). Among the total 1,014,212 TR loci, 1,014,188 non-redundant TR loci were used for downstream analyses.

To identify TRs that have emerged only in the human lineage, we employed a method from a study examining Alu elements (Sen et al. 2006) (Fig. 1b). For each tandem repeat loci identified in the human genome, 400 bp of upstream and downstream were extracted from unmasked genome sequences. RBH (Reciprocal best hits) were detected to find orthologous sequences in masked NHP genomes using Blast with default settings (ver. 2.2.30) (Moreno-Hagelsieb and Latimer 2008). We used bitscore to find the best hits, and filtered out those that did not cover whole query sequence in both the upstream and downstream regions.

The intervening sequence (IS) between the two flanking sequences on NHP genomes was examined to determine HSTRs, and the loci with IS that include "*N*" base or do not have any sequence were removed. For the remaining loci, we only considered the loci, where NHP have repeat count of one to reduce false positives derived from variation within each NHP population (Fig. 1d). The criteria for this process

are as follows: (1) Human TR (HTR) length ≥ IS length*2; (2) HTR length – IS length ≥ unit length of HTR; and (3) at least one match with percent identity > 95% between unit sequence of HTR and IS. We repeated the above procedure for three comparisons (human vs chimpanzee, human vs gorilla, and human vs orangutan) and intersected three TR lists. The TRs in the list were considered as human-specific TRs that are expanded only in human genome.

Basic statistics and genomic location of HSTR

For the sequences of total TRs and HSTRs, five characteristics (total length, unit length, number of repeats, and percentage of match between repeats) were independently investigated. The percentage of match between repeat units of HSTRs was compared to total TRs using a random sample. The random sampling was performed with the subset of the total TRs that have same number of repeats to HSTRs. After generating 152 random samples (same as HSTRs), we calculated the mean percentage of match, and repeated this process 1000 times. The location of TR was classified into exon, intron, boundaries of exon and intron, CDS, and intergenic using Ensembl annotations (Hubbard et al. 2002).

In silico validation of HSTRs in 1000 genome data set

To investigate HSTR variation at the population level, we examined them in 24 ethnically diverse, high coverage genomes from the 1000 genomes project (Consortium 2015) (Table S2). BLAST (Altschul et al. 1990) databases were built from the whole genome sequence reads for each individual. We used HSTR sequences with 20 nucleotides of flanking sequence as queries. Since the read length of the data was only 250 bp, queries longer than 250 bp were excluded from the BLAST search. A hit was counted if the high-scoring segment pair (HSP) contained at least 98% of the query sequence.

Investigation of common variants in HSTR loci

A list of common variants was downloaded from dbSNP (build 151) (Sherry et al. 2001). In this list, common variants are defined as follows; variants with a minor allele of frequency ≥ 1% in at least one population of the 1000 genomes project and for which two or more founders contribute to that minor allele frequency. The common variant was counted if it was located in any HSTR regions. The list of common variants does not include unplaced scaffolds; thus, two HSTRs located in unplaced scaffolds were excluded. For comparison with non-HSTRs, 10,000 control sets were generated by random sampling TR sets that have the same number of TRs and minimum/maximum TR length the same as that of HSTRs.

Expression quantification of orthologous genes in human, chimpanzee, gorilla, orangutan, and brain organoids

RNA-seq data of human, chimpanzee, gorilla, and orangutan were obtained from the sequence read archive (SRA) database with accession number SRP007412 (Brawand et al. 2011). The data set consists of 48 tissue samples across frontal cortex, cerebellum, liver, heart, kidney, and testis (Table S3). The raw reads were quality checked and trimmed for low-quality regions and adaptor sequences using Trimmomatic v0.36 (Bolger et al. 2014). The clean reads were aligned to the reference genomes (GRCh38, panTro4, gorGor4, and ponAbe2) using HISAT v2-2.1.0 [2]. We then quantified gene expression by counting the aligned reads using FeatureCounts with annotation file (.GTF) from Ensembl (Hubbard et al. 2002). To compare gene expression between species, 15,508 orthologous groups for human and other primates were determined by intersecting pairwise 1-to-1 orthologous genes between human and three NHP using Ensembl bioMart. For each orthologous group, RPKM (Reads per kilobase per million) and quantile normalization were performed in each tissues. We also used the gene expression data set of cerebral organoids (52 human and 344 chimpanzee samples), which were obtained from GEO (gene expression omnibus) with accession number GSE86207 (Mora-Bermúdez et al. 2016) (Table S4). Because this data set has been already normalized by RPKM method and includes orthologous group information, RPKM normalization and orthologous group identification were not performed for this data set.

For each two RPKM normalized data, we independently performed re-normalization in each tissues using quantile normalization (Barbash and Sakmar 2017; Brawand et al. 2011). In each tissue, we calculated the logFC of HSTR genes by dividing the mean expression level of human by that of NHP ($\log_2$(Human/Other primates)). For comparing the logFC of HSTR genes with that of all genes, we resampled the same number of genes as HSTR-associated genes, and calculated logFC. By repeating this process one million times, we constructed a null distribution and calculated empirical *P* values by determining the position of HSTR logFC in that distribution.

Gene set enrichment analysis

For gene sets containing or nearby an HSTR, enrichment tests were performed using DAVID functional annotation (Dennis et al. 2003). *P* value < 0.05 was used as the cutoff value of enrichment tests.

Multiple alignments data of 99 vertebrate genomes with human genome

Multiple alignments of 100 vertebrates (GRCh38) were downloaded from UCSC genome browser (Casper et al. 2017).

ChIP-seq data for histone modification overlap with HSTRs

To identify histone modifications that are likely to be responsible for regulating gene expression in brain tissues, we used 49 bed files, from the ENCODE (Consortium 2007) database. Four histone marks were considered: H3K4me3 ($n = 15$) and H3K9ac ($n = 7$) for promoter activity, and H3K4me1 ($n = 15$) and H3K27ac ($n = 12$) for enhancer activity (Table S5). For each histone marks, consensus peaks across samples were generated by DiffBind (Stark and Brown 2011) with the following options; minOverlap = 2, summits = 250.

Scan for binding motif of transcription factor within HSTR sequences

To scan for transcription factor-binding motifs in our HSTRs, we used position-weighted matrices from the JASPAR 2016 database (Mathelier et al. 2015), which consist of 386 transcription factor motif profiles implemented in TFBSTools (Tan and Lenhard 2016). Multiple testing problems were corrected with the Bonferroni method (Dunn 1961) and adjusted p value < 0.05 was used as the cutoff value.

Acknowledgements Not applicable.

Author contributions HK and SS conceived and designed this study. KK and SB analyzed and interpreted all of the data, and drafted the manuscript. DY performed genomic data analysis regarding the tandem repeat identification. All authors read and approved the final manuscript.

Funding Not applicable

Compliance with ethical standards

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Data availability The data sets supporting the conclusions of this article are included within the article and its supplementary information.

Competing interests The authors declare that they have no competing interests.

References

Ahmed M, Liang P (2012) Transposable elements are a significant contributor to tandem repeats in the human genome Comparative and functional genomics 2012. Comp Funct Genom. https://doi.org/10.1155/2012/947089

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Barbash S, Sakmar TP (2017) Brain gene expression signature on primate genomic sequence evolution. Sci Rep 7:17329

Bell MV, Cowper AE, Lefranc M-P, Bell JI, Screaton GR (1998) Influence of intron length on alternative splicing of CD44. Mol Cell Biol 18:5930–5941

Bennett S et al (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat Genet 9:284

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573

Bieberstein NI, Carrillo Oesterreich F, Straube K, Neugebauer KM (2012) First exon length controls active chromatin signatures and transcription. Cell Rep 2:62–68. https://doi.org/10.1016/j.celrep.2012.05.019

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30:2114–2120

Brawand D et al (2011) The evolution of gene expression levels in mammalian organs. Nature 478:343

Cáceres M et al (2003) Elevated gene expression levels distinguish human from non-human primate brains. Proc Natl Acad Sci 100:13030–13035

Carroll SB (2005) Evolution at two levels: on genes and form. PLoS Biol 3:e245

Casper J et al (2017) The UCSC genome browser database: 2018 update. Nucleic Acids Res 46:D762–D769

Consortium EP (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799

Consortium EP (2015) A global reference for human genetic variation. Nature 526:68

Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M (2002) A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat Genet 30:315

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4:R60

Dumas LJ et al (2012) DUF1220-domain copy number implicated in human brain-size pathology and evolution. Am J Hum Genet 91:444–454

Dunn OJ (1961) Multiple comparisons among means. J Am Stat Assoc 56:52–64

Enard W et al (2009) A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell 137:961–971

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44:445–477

Gu J, Gu X (2003) Induced gene expression in human brain after the split from chimpanzee. Trends Genet 19:63–65

Gymrek M et al (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet 48:22

Gymrek M, Willems T, Reich D, Erlich Y (2017) Interpreting short tandem repeat variations in humans using mutational constraint. Nat Genet 49:1495–1501. https://doi.org/10.1038/ng.3952

Hamada H, Seidman M, Howard B, Gorman CM (1984) Enhanced gene expression by the poly (dT-dG). poly (dC-dA) sequence. Mol Cell Biol 4:2622–2630

Hill RS, Walsh CA (2005) Molecular insights into human brain evolution. Nature 437:64
Hubbard T et al (2002) The Ensembl genome database project. Nucleic Acids Res 30:38–41
Khaitovich P et al (2008) Metabolic changes in schizophrenia and human brain evolution. Genome Biol 9:R124
Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. Proc Natl Acad Sci 99:803–808
Lander ES et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
Legendre M, Pochet N, Pak T, Verstrepen KJ (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res 17:1787–1796
Mathelier A et al (2015) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 44:D110–D115
Matoulkova E, Michalova E, Vojtesek B, Hrstka R (2012) The role of the 3′ untranslated region in post-transcriptional regulation of protein expression in mammalian cells. RNA Biol 9:563–576. https://doi.org/10.4161/rna.20231
Mill J, Asherson P, Browes C, D'Souza U, Craig I (2002) Expression of the dopamine transporter gene is regulated by the 3′ UTR VNTR: evidence from brain and lymphocytes using quantitative RT-PCR. Am J Med Genet Part A 114:975–979
Mora-Bermúdez F et al (2016) Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. Elife 5:e18683
Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24:319–324. https://doi.org/10.1093/bioinformatics/btm585
O'bleness MS, Dickens CM, Dumas LJ, Kehrer-Sawatzki H, Wyckoff GJ, Sikela JM (2012a) Evolutionary history and genome organization of DUF1220 protein domains. G3 Genes Genomes Genet 2:977–986
O'bleness M, Searles VB, Varki A, Gagneux P, Sikela JM (2012b) Evolution of genetic and genomic features unique to the human lineage. Nat Rev Genet 13:853
Pai AA, Henriques T, Paggi J, Burkholder A, Adelman K, Burge CB (2017) Intron length and recursive sites are major determinants of splicing efficiency in flies. bioRxiv. https://doi.org/10.1101/107995
Sen SK et al (2006) Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet 79:41–53
Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311
Sonay TB et al (2015) Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. Genome Res 25:1591–1599
Sousa AM, Meyer KA, Santpere G, Gulden FO, Sestan N (2017) Evolution of the human nervous system function, structure, and development. Cell 170:226–247
Stark R, Brown G (2011) DiffBind: differential binding analysis of ChIP-Seq peak data R package version 100:4.3
Streelman JT, Kocher TD (2002) Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. Physiol Genom 9:1–4
Suzuki S, Miyabe E, Inagaki S (2018) Novel brain-expressed noncoding RNA, HSTR1, identified at a human-specific variable number tandem repeat locus with a human accelerated region. Biochem Biophys Res Commun 503:1478–1483
Tan G, Lenhard B (2016) TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics 32:1555–1556
Usdin K (2008a) The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res 18:1011–1019
Usdin K (2008b) The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res 18:1011–1019. https://doi.org/10.1101/gr.070409.107
Vergnaud G, Denoeud F (2000) Minisatellites: mutability and genome architecture. Genome Res 10:899–907
Walker FO (2007) Huntington's disease. Lancet 369:218–228
Warpeha K et al (1999) Genotyping and functional analysis of a polymorphic (CCTTT) n repeat of NOS2A in diabetic retinopathy. FASEB J 13:1825–1832
Wieringa B, Hofer E, Weissmann C (1984) A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. Cell 37:915–925
Zimmer F, Montgomery SH (2015) Phylogenetic analysis supports a link between DUF1220 domain number and primate brain expansion. Genome Biol Evol 7:2083–2088