Check for updates

Tandemly repeated NBPF HOR copies (Olduvai triplets): Possible impact on human brain evolution

Matko Glunčić[1] , Ines Vlahović[2], Marija Rosandić[3,4], Vladimir Paar[1,4]

Previously it was found that the neuroblastoma breakpoint family (*NBPF*) gene repeat units of ~1.6 kb have an important role in human brain evolution and function. The higher order organization of these repeat units has been discovered by both methods, the higher order repeat (HOR)-searching method and the HLS searching method. Using the HOR searching method with global repeat map algorithm, here we identified the tandemly organized NBPF HORs in the human and nonhuman primate NCBI reference genomes. We identified 50 tandemly organized canonical 3mer NBPF HOR copies (Olduvai triplets), but none in nonhuman primates chimpanzee, gorilla, orangutan, and Rhesus macaque. This discontinuous jump in tandemly organized HOR copy number is in sharp contrast to the known gradual increase in the number of Olduvai domains (NBPF monomers) from nonhuman primates to human, especially from ~138 in chimpanzee to ~300 in human genome. Using the same global repeat map algorithm method we have also determined the 3mer tandems of canonical 3mer HOR copies in 20 randomly chosen human genomes (10 male and 10 female). In all cases, we found the same 3mer HOR copy numbers as in the case of the reference human genome, with no mutation. On the other hand, some point mutations with respect to reference genome are found for some NBPF monomers which are not tandemly organized in canonical HORs.

Introduction

The ~1.6-kb repeat units in the neuroblastoma breakpoint family (NBPF) genes

NBPF genes in human chromosome 1 contain a pronounced repetitive structure of ~1.6-kb repeat units (Fortna et al, 2004; Vandepoele et al, 2005; Popesco et al, 2006), with sizable divergence (~20%) among neighboring repeat units. These repeat units encode the protein domain previously named DUF1220 domain (Vandepoele et al, 2005), which was changed to Olduvai domain (Sikela & van Roy,

2017). These repeat units within *NBPF* genes were previously also called NBPF repeats or NBPF monomers (Paar et al, 2011) (Fig 1), in accordance with repeat terminology used by Willard (1985) in studies of *α* satellite repeats. Each of Olduvai domains (NBPF repeat units) is characterized by a structure with two evenly spaced exons (Popesco et al, 2006; Paar et al, 2011). Comprehensive studies have shown that the Olduvai domain copy number was correlated with brain size, cortical neuron number, IQ scores, cognitive aptitude, evolution, and with brain pathologies (autism, schizophrenia, microcephaly, macrocephaly, and neuroblastoma) (Vandepoele et al, 2005, 2008; Popesco et al, 2006; Andries et al, 2012; Dumas et al, 2012; Davis et al, 2014; Keeney et al, 2014; Quick et al, 2016; Astling et al, 2017; Mitchell & Silver, 2018; Fiddes et al, 2019; Heft et al, 2020). O'Bleness et al (2012, 2014) showed that expanded Olduvai triplets are responsible for the extreme increase in Olduvai copy number in humans. The linking of Olduvai copy number and human brain evolution was first suggested in Popesco et al (2006).

HLS-searching method

Comparing sequences of repeated Olduvai domains, it was found that they are predominantly of three types, which were referred to as HLS-1, HLS-2, and HLS-3, and most often they appear in triplets (O'Bleness et al, 2012; O'Bleness et al, 2014) (Fig 1). Because of recent change of terminology, which was fully justified (Sikela & van Roy, 2017), these HLS triplets were referred to as Olduvai triplets. HLS searching method identifies Olduvai domains in a given genomic sequence in the first step and compares their divergence to determine Olduvai triplets in the second step.

Higher order repeat (HOR)-searching method

Centromeric regions of primate chromosomes are largely built from *α* satellite repeat units of of length ~171 bp, which have been extensively studied (Manuelidis, 1978; Willard, 1985; Jorgensen et al, 1987; Tyler-Smith & Brown, 1987; Waye &Willard, 1987; Choo et al, 1991; Jurka et al, 1996, 2005; Warburton & Willard, 1996; Alexandrov et al, 2001; Rosandic et al, 2003; Rudd & Willard, 2004; Paar et al, 2005; Warburton et al, 2008; Aldrup-Macdonald & Sullivan, 2014; Miga,

---

[1]Faculty of Science, University of Zagreb, Zagreb, Croatia   [2]Algebra University College, Zagreb, Croatia   [3]University Hospital Centre Zagreb (ret), Zagreb, Croatia   [4]Croatian Academy of Sciences and Arts, Zagreb, Croatia

Correspondence: matko@phy.hr

HLS terminology

Olduvai triplet   Olduvai triplet   Olduvai triplet

HOR terminology

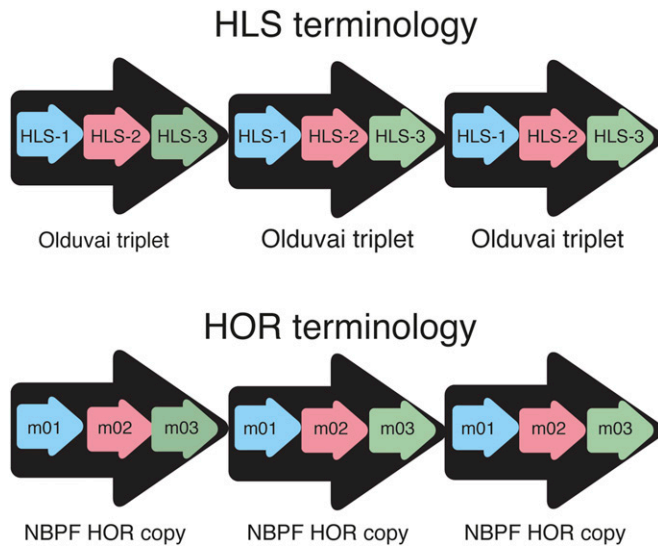NBPF HOR copy   NBPF HOR copy   NBPF HOR copy

Figure 1. HLS versus higher order repeat terminology.

2017; Sullivan et al, 2017; Lower et al, 2018; Gluncic et al, 2019; Uralsky et al, 2019). Neighboring α satellite monomers (primary repeat units) diverge sizably from each other (~20–40%), but some stretches of n monomers are often organized into secondary repeat $n$mer HOR copies, with small mutual divergence between HOR copies (less than 5%, in some cases even below 1%). Thus, divergence between HOR copies is much smaller than between monomers within each HOR copy (Warburton & Willard, 1996). Pronounced α satellite $n$mer HORs in human genome are chromosome specific, for example, 2mer and 11mer in chromosome 1; 15mer in chromosome 8; 5mer in chromosome 11, 12mer in chromosome X, and 34mer in chromosome Y (Alexandrov et al, 2001).

An efficient and robust computational tool to identify large HOR copies, as for example, α satellite HORs, is the global repeat map (GRM) algorithm (Paar et al, 2011; Gluncic & Paar, 2013; Gluncic et al, 2019; Vlahović et al, 2020). HOR method to determine α satellite repeat structure is based on identifying α satellite HOR copies in a given sequence in the first step, and in the second step deducing α satellite monomers from the HOR structure. Thus, HOR method enables identification of HOR copies without any prior knowledge of the primary repeats on which the HORs are based. Because the divergence among HOR copies is much smaller than the divergence among its underlying primary repeats, the scope of computations is simpler in the case of using HOR method than in the method of using monomer identification.

GRM algorithm identifies all pronounced HORs in a given genomic sequence. For example, applying in 2011 the GRM algorithm to the Builds 36.3 assembly of human chromosome 1 we obtained three HORs (Paar et al, 2011):

(1) The already known α satellite 11mer HOR (based on ~171-bp primary α satellite repeat unit);
(2) The novel 3mer HOR, based on previously known ~1.6-kb primary repeat units from NBPF genes in human chromosome 1;
(3) The novel quartic HOR, based on known 39-bp hornerin primary repeat unit in human chromosome 1.

Whereas the 11mer α satellite HORs are located in the centromere, the last two HORs are located within NBPF genes and Hornerin genes, respectively.

We also note that by applying GRM algorithm to genomic sequences of insects, the HOR pattern (based on T-cast-360 monomers of 331, 361, 362, and 369-bp primary repeat units), has been also discovered in beetle *Tribolium castaneum* (Vlahovic et al, 2017), which is evolutionary very distant from primates.

Using the robust GRM algorithm, we have first discovered in Build 36.3 assembly of human chromosome 1 the ~4,770-bp peak of 3mer HOR (Paar et al, 2011). The GRM diagram for this ~4,770-bp consensus sequence revealed two additional internal GRM peaks, at ~1.6 and ~3.2 kb. The GRM diagram for the 3.2-kb peak also shows that its consensus sequence consists of two ~1.6-kb monomer repeats. Altogether, the ~4,770-bp consensus sequence consists of a tandem of three ~1.6-kb monomer repeats of consensus lengths 1,623, 1,593, and 1,554 bp, respectively. The corresponding consensus monomers were denoted m01, m02, and m03, respectively. In this way, we discovered that the ~4,770-bp repeat copies are the 3mer HORs. Because these 3mer HORs are located within the *NBPF* genes in human chromosome 1, we named them NBPF 3mer HORs, and their constituent ~1.6-kb monomers were referred to as NBPF monomers in accordance with Willard's terminology for HORs (Warburton & Willard, 1996). Each consensus monomer consists of two exons and three introns. Each consensus monomer contains in succession intron-exon-intron-exon-intron. In the 1,554-bp consensus monomer their lengths amount to ~8%, ~4%, ~39%, ~12%, and ~37% of monomer length. In the 1,593-bp consensus monomer the length of second intron increases by ~8%, and in the 1,623-bp consensus monomer additionally the length of the first intron increases by ~6%.

In the Build 36.3 genomic assembly of human chromosome 1, we identified 47 HOR copies forming tandems, but only 34 of these 47 were canonical as shown in Fig 8 from Reference Paar et al (2011). The three constituent monomers of ~1.6 kb correspond to HLS DUF1220 domains of Fortna et al (2004), Vandepoele et al (2005), and Popesco et al (2006). As pointed out in Andries et al (2012), before the discovery of NBPF 3mer HOR by Paar et al (2011), it was not realized by HLS-searching method that the HLS domains form the triplet organization, that is, that the diverging DUF1220 HLS domains are of three types, forming a remarkable triplet organization—the HORs. On the other hand, the Build 2.1 (2010) genomic assembly of chimpanzee chromosome 1 did not show tandemly organized NBPF HOR copies, although 14 individual NBPF HOR copies (not organized into tandem) were identified and the total number of dispersed NBPF monomers was 48 (Paar et al, 2011). Also, tandemly organized NBPF HOR copies were not found in genomic assembly WUSTL Pongo_albelii-2.0.2 (2010) of orangutan and in Build 1.1 (2010) of Rhesus macaque (Paar et al, 2011). On this basis we hypothesized that the NBPF HOR tandem repeats (47 NBPF HOR copies in human versus 0 in nonhuman primate genome) reveal the NBPF HOR tandem repeats are human specific, possibly contributing to human brain evolution and human-chimpanzee divergence (Paar et al, 2011).

A possible challenge to such hypothesis has arisen from another complex HOR pattern, discovered by GRM algorithm within the hornerin (HRNR) gene in Build 36.3 genome assembly of human chromosome 1 (Paar et al, 2011): nine 39-bp primary repeat HRNR

units organized into ~0.35-kb secondary repeat units → two ~0.35-kb secondary HOR repeat units organized into ~0.7-kb tertiary HOR repeat units → and finally two 0.7-kb tertiary HOR repeat units organized into ~1.4-kb quartic HOR repeat units. In this, the 1,410-bp repeat unit represents the quartic HOR repeat unit, which is tandemly repeated, as the most complex multi-step HOR pattern discovered so far. In Reference Paar et al (2011), the quartic HOR was detected in the human genome only, and no quartic HOR counterpart was found in then-available sequenced chimpanzee genome.

The more recent investigation by Romero et al (2018) of HRNR from more complete NCBI human genomic assembly NC_000001.11 (2018) fully confirmed the HRNR quartic HOR formation described by Paar et al (2011). However, Romero et al (2018) found that the tandem quartic HOR formation of HRNR was conserved also in more recent reference genomes of chimpanzee and other nonhuman primates, except the crab-eating macaque. This showed that the tandemly organized HRNR quartic HOR copies are not exclusively human specific, as appeared for reference genome in 2011, but are also present in more recent NCBI reference genome sequences of most primates (Romero et al, 2018).

Prompted by that development, here we investigate whether the tandemly organized canonical NBPF 3mer HORs are present in current genome assemblies in both human and nonhuman primate genomes, or are exclusively human-specific, as was indicated by previous chimpanzee reference genome.

Results

Computed GRM diagrams for human genome assembly NC_000001.11 (2022) of human chromosome 1 and resulting NBPF HOR/monomer pattern

GRM diagram for NC_000001.11 (2022) human genome assembly of chromosome 1 shows a peak at ~4.8 kb, corresponding to the ~4.8 kb tandemly organized canonical NBPF 3mer HOR copies, based on ~1.6 kb-NBPF monomer (Fig 2A). Its constituents ~1.6-kb NBPF monomers contribute also to the GRM peaks at ~1.6 and ~3.2 kb.

Using GRM algorithm, we obtain that the arrays of tandemly organized canonical NBPF HOR copies are located in human genes *NBPF20* (19 canonical HOR copies), *NBPF19* (13 canonical HOR copies), *NBPF10* (11 canonical HOR copies), and *NBPF14* (7 canonical HOR copies), in total 50 tandemly organized canonical NBPF HOR copies. In addition, we identified 22 noncanonical NBPF HOR copies (Table S2). Canonical 3mer HOR copy is defined as follows: (i) each HOR copy is constituted of three monomers m1, m2, and m3, where each monomer diverges from its consensus by less than 5% and (iii) each HOR copy is in tandem with two or more HOR copies. Human consensus NBPF monomers, determined from canonical NBPF HORs, are displayed in Table S1.

The human NBPF HOR pattern with 50 canonical NBPF HOR copies encompasses 34 canonical HOR copies identified previously in NCBI (2010) genome assembly (Paar et al, 2011) and is similar to the HLS DUF1220 triplet repeat pattern identified by monomer searching method (O'Bleness et al, 2014; Heft et al, 2020).

Distribution of NBPF monomers among *NBPF* genes is shown in Fig 2, as obtained by HOR-searching method. Additional isolated NBPF monomer or HOR copies could be found by increasing divergence interval. For example, in the gene *NBPF12* a doublet of HOR copies, with one monomer missing in the second HOR copy, could change into a doublet of two canonical HOR copies by extending divergence limit for recognition of the third monomer. Analogously, an isolated noncanonical HOR copy could become canonical.

Divergence between canonical NBPF HOR copies is an order of magnitude smaller than divergence between monomers within each NBPF HOR copy. For example, in the NBPF20 array of 19 canonical NBPF HOR copies, the average divergence between NBPF HOR copies is ~1.6%, whereas the average divergence between monomers within each NBPF HOR copy is ~18%. This pattern is consistent with divergence pattern for the well-known α satellite HORs constituted from the 171-bp α satellite monomers (Warburton & Willard, 1996).

Computed GRM diagrams and NBPF monomer distributions for genome assemblies of nonhuman chromosomes 1 of chimpanzee (NC_036879.1, 2022), gorilla (NC_044602.1, 2022), orangutan (NC_036903.1, 2022), and rhesus macaque (NC_041754.1, 2022), and resulting NBPF HOR/monomer pattern

GRM diagrams for nonhuman genome assemblies of the whole chromosome 1 are displayed in Fig 2B–E (left panel), which shows that the peak at ~4.8 kb, corresponding to the tandemly repeated NBPF 3mer HORs, is absent. This is even more evident in GRM diagrams for genome segments which contain NBPF genes (Fig 2B–E, right panel).

As an alternative method to search for distribution of NBPF monomers in nonhuman primate genomes, we applied Edlib (Sosic & Sikic, 2017), an open-source C/C ++ library for exact pairwise sequence alignment using edit (Levenshtein) distance, with human NBPF consensus monomers m1, m2, m3, obtained by GRM algorithm. Distribution and divergence of NBPF monomers, 3mer HORs and 2mer variants in primate genomes (human, chimpanzee, gorilla, orangutan, and Rhesus macaque) are displayed in Table S2. Distribution of NBPF monomers in nonhuman primate genomes shows the absence of tandemly repeated canonical NBPF HORs (Fig 3), in full accordance with the absence of the 4.8 kb GRM peak in GRM diagrams (Fig 2B–E).

Both methods (1) computing GRM diagram for each nonhuman primate genome and looking for existence of the ~4.8-kb GRM peak, and (2) using the human NBPF consensus monomers, determined by GRM algorithm, to analyze genomes of nonhuman primates, prove the absence of tandemly organized canonical NBPF HORs in nonhuman primate genomes.

Comparison of copy number and structural organization of tandemly organized NBPF 3mer HOR copies of 20 randomly chosen individual human genomes from The International Genome Sample Resource and human reference genome GRCh38

Variations in Olduvai triplet copy number among various human genomes have been reported (Heft et al, 2020). To observe which is the diversity of this genomic portion in terms of copy number and
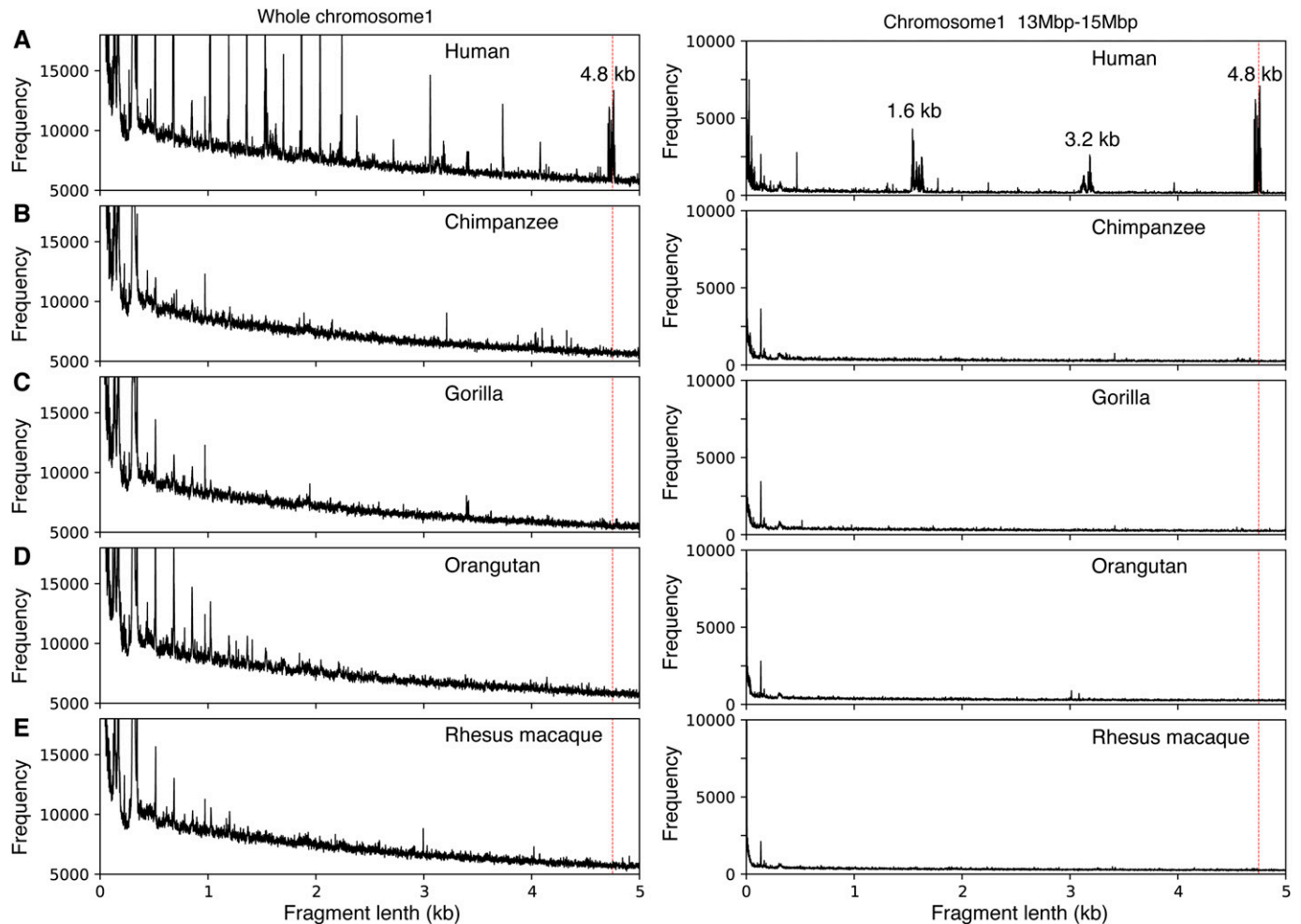
Figure 2. Global repeat map diagrams for National Center for Biotechnology Information (2022) primate genome assemblies.
(A) Human, NC_000001.11. **(B)** Chimpanzee, NC_036879.1. **(C)** Gorilla, NC_044602.1. **(D)** Orangutan, NC_036903.1. **(E)** Rhesus macaque, NC_041754.1. Peaks at ~4.8 kb correspond to the NBPF 3mer higher order repeat. These peaks are present in human and absent in nonhuman primate genomes. These diagrams demonstrate that the NBPF 3mer higher order repeat is exclusively characteristic of human genome, completely absent in National Center for Biotechnology Information (2022) assemblies of nonhuman primate genomes.

mutations in different genomes, we have applied our HOR-finding method to sequences of 20 randomly chosen individuals from The International Genome Sample Resource (The 1000 Genomes Project: http://www.internationalgenome.org/data-portal) (Table S3). In all 20 genomes, we obtained exactly the same copy numbers and the same consensus HOR sequences (with no mutations) equal to results for reference genome GRCh38, as illustrated by Table 1. (For full results on all 20 genomes, see Table S4, where monomers marked by asterisk do not belong to canonical 3mer HOR copies and some have a point mutation with respect to results for reference genome GRCh38.) In the table, the same value of divergence for the corresponding monomers in canonical HOR copies for all 20 individual genomes and reference genome means that monomers in canonical HOR copies in all 20 individual genomes have no mutation with respect to the reference genome. On the other hand, a sizable number of monomers from HOR copies outside of tandemly organized 3mer HOR copies have some point mutations with respect to the reference genome, although their HOR copy number

and structural organization (monomer types) are the same as for reference genome. For example, the every monomer No. 30 in 20 individual genomes from Table S3 has the same monomer length 1,529 bp and monomer type m01, but only eight of them have the same divergence value (14.54%) as the corresponding monomer No. 30 for reference genome (i.e., no point mutation with respect to the reference genome), whereas the remaining 12 individual genomes have slightly different divergences (14.60%, 14.67%) (i.e., point mutation with respect to the reference genome). Because of large length of NBPF monomer (~1.6 kb) such rare point mutations cannot influence the NBPF HOR structure and copy numbers, but are restricted only to slight difference of monomer sequences. In this sense, the tandemly repeated 3mer consensus HOR pattern may be considered as a kind of characteristic signature of human genome. A recent study indicates that human NBPF proteins that contain many tandemly repeated Olduvai triplets are posttranslationally cleaved by the furin protease once at each triplet (Pacheco et al, 2022 Preprint). This
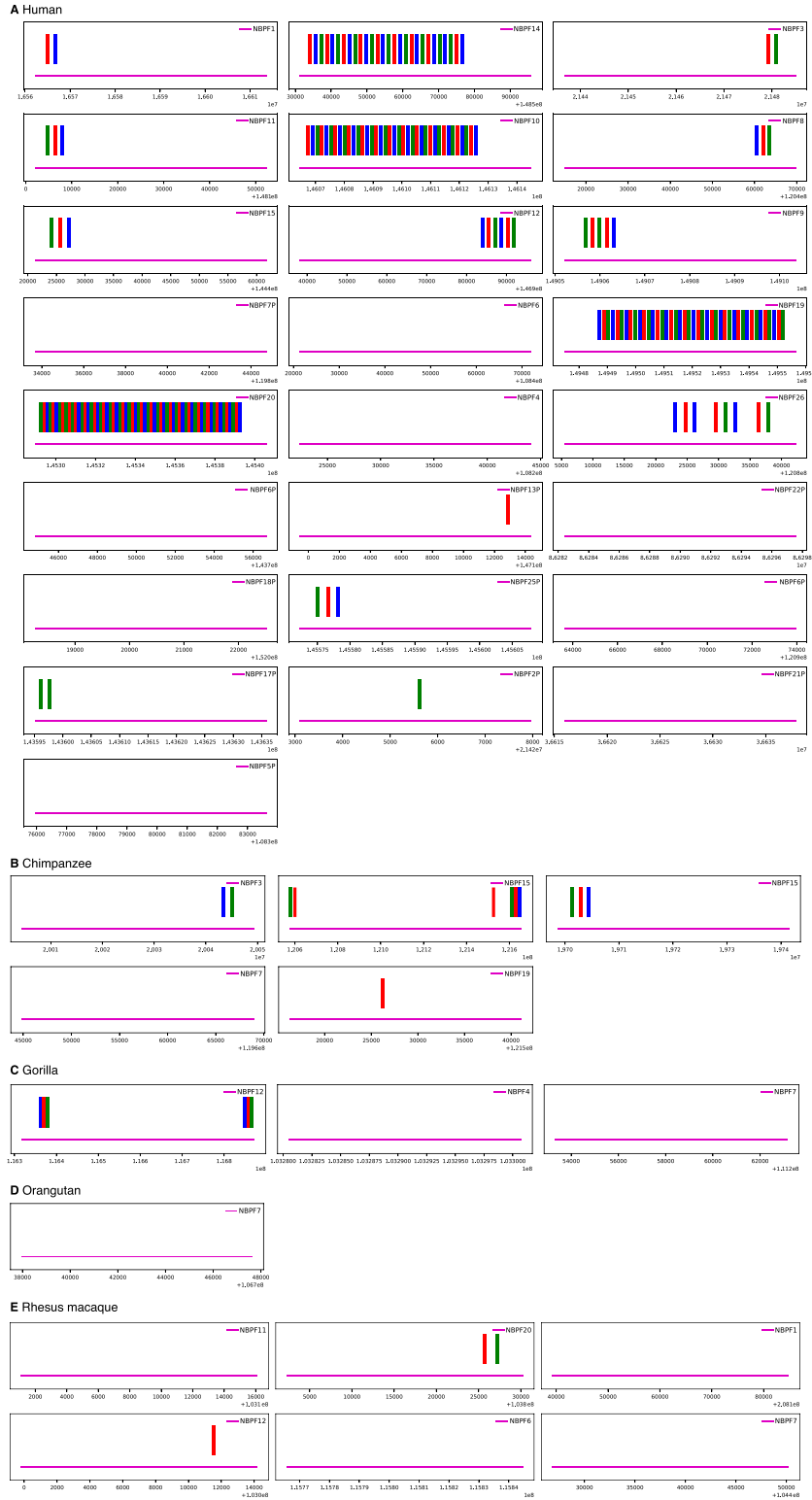
Figure 3. NBPF monomer distribution along the National Center for Biotechnology Information (2022) genome assemblies.
(A) Human, NC_000001.11. **(B)** Chimpanzee, NC_036879.1. **(C)** Gorilla, NC_044602.1. **(D)** Orangutan, NC_036903.1. **(E)** Rhesus macaque, NC_041754.1. Horizontally displayed is position along genome sequence within each NBPF gene. Each NBPF monomer is shown by a short vertical line: m1 (green), m2 (red), and m3 (blue). The canonical NBPF 3mer copies in tandem are present in four human NBPF genes: NBPF20, NBPF19, NBPF10, and NBPF14.

Table 1.   Comparison of the human reference genome GRCh38 copy number and structural organization for an illustrative segment of tandemly organized NBPF higher order repeat copies to the result for 20 randomly chosen individual human genomes from the International Genome Sample Resource (the 1000 Genomes Project: http://www.internationalgenome.org/data-portal).

Reference genome GRCh38				20 individual genomes from Table S3		
Mon. No.	L(bp)	Mon. type	Div. (%)	L(bp)	Mon. type	Div. (%)
67	1,544	m03	1,78	1,544	m03	1,78
68	1,628	m02	0,44	1,628	m02	0,44
69	1,584	m01	1,56	1,584	m01	1,56
70	1,542	m03	1,29	1,542	m03	1,29
71	1,625	m02	1,78	1,544	m02	1,78
72	1,587	m01	1,78	1,544	m01	1,78
73	1,542	m03	1,84	1,542	m03	1,84
74	1,629	m02	2,27	1,629	m02	2,27
75	1,599	m01	1,62	1,599	m01	1,62
76	1,542	m03	2,09	1,542	m03	2,09
77	1,631	m02	1,45	1,631	m02	1,45
78	1,591	m01	1,62	1,591	m01	1,62
79	1,542	m03	1,72	1,542	m03	1,72
80	1,622	m02	1,33	1,622	m02	1,33
81	1,584	m01	1,49	1,584	m01	1,49
82	1,541	m03	1,78	1,541	m03	1,78
83	1,628	m02	0,76	1,628	m02	0,76
84	1,584	m01	1,49	1,584	m01	1,49
85	1,542	m03	1,84	1,542	m03	1,84
86	1,629	m02	1,20	1,629	m02	1,20
87	1,589	m01	1,62	1,589	m01	1,62

Mon. No., ordinal number of NBPF monomer/Olduvai domain in global repeat map (GRM) analysis; L(bp), length of monomer; Mon. type, monomer type from GRM analysis; Div. (%), divergence of monomer with respect to the GRM consensus of the corresponding type. This illustrative segment presents the subsequence of monomers No. 67–87, which belong to tandemly organized canonical 3mer higher order repeat copies. Start position of the monomer no. 67 in the reference genome GRCh838 is 144424115.

indicates that Olduvai triplet proteins are the ultimate functional unit for Olduvai in humans.

Discussion

Results of both HLS-searching and HOR-searching methods are largely congruent, in spite of using computational procedures with different divergence range. In the HOR copy searching procedure divergence range is much lower than in the HLS domain searching. On the other hand, the HLS-searching method provides in the first step also domains conserved among primates (CON), which can be obtained in the HOR-searching method only in the second step, after determining 3mer HORs and their repeat units.

From analysis of divergence in HOR-searching method by using GRM algorithm, the identified NBPF HOR copies have a pattern comparable with classical Willard's α satellite HORs in the centromeric regions (Warburton & Willard, 1996). For example, the average divergence among canonical HOR copies in HOR array within the NBPF20 gene is ~1%, whereas among monomers within each HOR copy divergence is sizable, ~17–19%. Results obtained for NBPF monomer repeats and 3mer HOR copies by using GRM method are similar to the corresponding Olduvai domains and Olduvai triplets, respectively, obtained by using HLS searching method, with slight difference due to different range of divergences. Because of smaller divergence between repeats detected in the first step of analyzing a given genomic sequence, it is simpler to use the HOR-searching than the domain (monomer) searching method, but the final results for NBPF HORs/Olduvai triplets are nearly the same.

As pointed out in Reference Andries et al (2012), before the discovery of NBPF 3mer HORs in Reference Paar et al (2011) by using the GRM HOR-searching method, it was not noted that diverging HLS domains, now called Olduvai domains, are of three types, and forming a remarkable 3mer HOR pattern. The main advantage of the GRM HOR-searching algorithm is the characteristic small divergence between HOR copies. After initial discovery of NBPF 3mer HORs (Paar et al, 2011), the alternative way to identify HORs in a given sequence was the HLS domains searching in the first step, and then, by analyzing their mutual divergences, in the second step

to identify their triplets (i.e., Olduvai triplets/HOR copies) with small mutual divergence among triplets (O'Bleness et al, 2012; O'Bleness et al, 2014; Astling et al, 2017). In comparison of results obtained by these two methods, it should be taken into account that different terminology is used for the same pattern: DUF1220 HLS domains (now Olduvai domains) instead of NBPF monomer and HLS DUF1220 triplet (now Olduvai triplets) instead of NBPF 3mer HOR copy. It is noted that the HOR terminology was established and widely used in the pioneering work of Willard and co-workers for α satellite HORs (Warburton & Willard, 1996).

Using the HLS searching method, it was discovered that the ~1.6 kb repeat units in primate NBPF genes are human specific; their copy number gradually increases from nonhuman primate genomes to human genome: rhesus macaque 74 → common marmoset 75 → olive baboon 75 → gorilla 97 → orangutan 101 → chimpanzee 138 → human 302 (Fortna et al, 2004; Vandepoele et al, 2005; Popesco et al, 2006; Dumas & Sikela, 2009; Dumas et al, 2012; Zimmer & Montgomery, 2015). Here we point out to even more dramatical human exclusive discontinuous jump of the number of tandemly organized NBPF canonical 3mer HOR copies (Olduvai triplets) from 0 in nonhuman primates to 50 in the evolutionary step from chimpanzee to human genome. One might hypothesize whether this nonlinearity might involve some additional coherence in human specific brain evolution, beyond the merely summation of individual ~1.6-kb Olduvai domain contributions.

Additional two pronounced HORs in human chromosome 1 are the quartic-order 2mer (2mer [9mer]) HOR based on 39-bp primary hornerin repeat unit in hornerin gene and the α satellite 11mer HOR in the centromere region (Warburton & Willard, 1996; Paar et al, 2011). A study of more recent chimpanzee reference genome has revealed the presence of tandemly organized hornerin quartic order HORs in more complete sequence of chimpanzee chromosome 1, that is, that tandemly organized quartic order hornerin HOR copies are present also in chimpanzee, thus excluding the existence of exclusively human tandemly organized hornerin quartic copies (Romero et al, 2018).

On the other hand, the problem of chimpanzee α satellite 11mer HOR remains open because of still large unsequenced gaps (Table S5) in the current reference genome NC_036879.1 from The National Center for Biotechnology Information (NCBI). Thus, the only proven exclusively human tandemly organized HOR copies in primates' chromosome 1 are the NBPF 3mer HOR/Olduvai triplet copies.

Extensive studies of evolution of Olduvai triplets in *NBPF* genes have been performed using Olduvai method. Thanks to rather complete sequencing of *NBPF* genes, it was possible to identify Olduvai triplets along the evolutionary chain. It was shown that the number of Olduvai triplets gradually expands along the evolutionary chain to human genome. Because of equivalence of Olduvai triplets and NBPF 3mer HOR copies, the same results hold for the HOR-searching method.

The evolutionary problem is more difficult for α satellite 11mer HOR. GRM of human genome assembly NC_000001.11 shows a pronounced peak for α satellite 11mer HOR, which is absent in current genome assemblies of nonhuman primates chimpanzee (NC_036879.1), gorilla (NC_044602.1), orangutan (NC_036903.1) and Rhesus macaque (NC_041754.1) (Fig S1). This is obviously due to large gaps in available current sequencing of centromeric region of

nonhuman genomes. Therefore, it is not possible at present to study evolution of 11mer HOR in primates.

Materials and Methods

NBPF HORs are identified in NCBI assembly (2022) of human and nonhuman primate genomes (https://www.ncbi.nlm.nih.gov/genome/), using GRM algorithm. GRM algorithm is efficient and robust method convenient to identify and study in a given genomic sequence very large repeat units, like HORs (Paar et al, 2011; Gluncic & Paar, 2013; Gluncic et al, 2019; Vlahović et al, 2020). The computational noise in detecting repeats increases with increased length and/or complexity of HOR repeat unit, which can mask some peaks corresponding to HOR copies. Such background noise is significantly reduced by using GRM algorithm. The novelty of GRM approach is a direct mapping of symbolic DNA sequence into frequency domain using complete K-string ensemble instead of statistically adjusted individual K-strings optimized locally. This provides a straightforward identification of DNA repeats in frequency domain but avoids mapping of symbolic DNA sequence to numerical sequence, and uses K-string matching but avoids statistical methods and locally optimizing individual K-strings (Paar et al, 2011; Gluncic & Paar, 2013). In this way, characteristics of GRM algorithm are robustness with respect to deviations from ideal repeats (substitutions, insertions, deletions), straightforward and parameter-free identification of repeats, applicability to very large repeat units—both primary repeats and HORs, and straightforward determination of consensus lengths and consensus sequences for primary repeats and HORs. The GRM method is a straightforward method to provide a global repeat map in a GRM diagram, identifying all pronounced repeats in a given sequence, without any prior knowledge on repeats. In addition, once the consensus repeat unit is determined using GRM, in the next step it can be combined with search for dispersed HOR copies or individual constituting monomers.

Conclusions

Using robust GRM algorithm, here the tandem repeats of NBPF HOR copies are searched in NCBI (2022) genome assemblies for nonhuman primates, chimpanzee, gorilla, orangutan, and Rhesus macaque, in comparison with human genome. The number of tandemly organized canonical NBPF HOR copies in human genome is 50, whereas in nonhuman primates is 0. It is hypothesized that the impact of NBPF monomners in human brain evolution is enhanced as a coherent effect when monomers are organized in tandem of HORs. This exclusively human specific NBPF HOR symmetry could involve a link in human cognitive development.

Supplementary Information

Supplementary Information is available at https://doi.org/10.26508/lsa.202101306.

Acknowledgements

Author Contributions

M Glunčić: conceptualization, resources, data curation, software, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, project administration, and writing—original draft, review, and editing.
I Vlahović: data curation, software, and formal analysis.
M Rosandić: data curation, investigation, and visualization.
V Paar: conceptualization, investigation, visualization, methodology, and writing—original draft, review, and editing.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

References

Aldrup-Macdonald ME, Sullivan BA (2014) The past, present, and future of human centromere genomics. *Genes (Basel)* 5: 33–50. doi:10.3390/genes5010033

Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y (2001) Alpha-satellite DNA of primates: Old and new families. *Chromosoma* 110: 253–266. doi:10.1007/s004120100146

Andries V, Vandepoele K, van Roy F (2012) The NBPF gene family. In *Neuroblastoma – Present and Future*, Shimada H (ed), pp 185–214. Rijeka, Croatia: InTech. doi:10.5772/28470

Astling DP, Heft IE, Jones KL, Sikela JM (2017) High resolution measurement of DUF1220 domain copy number from whole genome sequence data. *BMC Genomics* 18: 614. doi:10.1186/s12864-017-3976-z

Choo KH, Vissel B, Nagy A, Earle E, Kalitsis P (1991) A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res* 19: 1179–1182. doi:10.1093/nar/19.6.1179

Davis JM, Searles VB, Anderson N, Keeney J, Dumas L, Sikela JM (2014) DUF1220 dosage is linearly associated with increasing severity of the three primary symptoms of autism. *PLoS Genet* 10: e1004241. doi:10.1371/journal.pgen.1004241

Dumas L, Sikela JM (2009) DUF1220 domains, cognitive disease, and human brain evolution. *Cold Spring Harb Symp Quant Biol* 74: 375–382. doi:10.1101/sqb.2009.74.025

Dumas LJ, O'Bleness MS, Davis JM, Dickens CM, Anderson N, Keeney JG, Jackson J, Sikela M, Raznahan A, Giedd J, et al (2012) DUF1220-domain copy number implicated in human brain-size pathology and evolution. *Am J Hum Genet* 91: 444–454. doi:10.1016/j.ajhg.2012.07.016

Fiddes IT, Pollen AA, Davis JM, Sikela JM (2019) Paired involvement of human-specific Olduvai domains and NOTCH2NL genes in human brain evolution. *Hum Genet* 138: 715–721. doi:10.1007/s00439-019-02018-4

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, et al (2004) Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* 2: E207. doi:10.1371/journal.pbio.0020207

Gluncic M, Paar V (2013) Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. *Nucleic Acids Res* 41: e17. doi:10.1093/nar/gks721

Gluncic M, Vlahovic I, Paar V (2019) Discovery of 33mer in chromosome 21-the largest alpha satellite higher order repeat unit among all human somatic chromosomes. *Sci Rep* 9: 12629. doi:10.1038/s41598-019-49022-2

Heft IE, Mostovoy Y, Levy-Sakin M, Ma W, Stevens AJ, Pastor S, McCaffrey J, Boffelli D, Martin DI, Xiao M, et al (2020) The driver of extreme human-specific Olduvai repeat expansion remains highly active in the human genome. *Genetics* 214: 179–191. doi:10.1534/genetics.119.302782

Jorgensen AL, Bostock CJ, Bak AL (1987) Homologous subfamilies of human alphoid repetitive DNA on different nucleolus organizing chromosomes. *Proc Natl Acad Sci U S A* 84: 1075–1079. doi:10.1073/pnas.84.4.1075

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110: 462–467. doi:10.1159/000084979

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR: A program for identification and elimination of repetitive elements from DNA sequences. *Comput Chem* 20: 119–121. doi:10.1016/s0097-8485(96)80013-1

Keeney JG, Dumas L, Sikela JM (2014) The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion. *Front Hum Neurosci* 8: 427. doi:10.3389/fnhum.2014.00427

Lower SS, McGurk MP, Clark AG, Barbash DA (2018) Satellite DNA evolution: Old ideas, new approaches. *Curr Opin Genet Dev* 49: 70–78. doi:10.1016/j.gde.2018.03.003

Manuelidis L (1978) Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma* 66: 23–32. doi:10.1007/bf00285813

Miga KH (2017) The promises and challenges of genomic studies of human centromeres. *Prog Mol Subcell Biol* 56: 285–304. doi:10.1007/978-3-319-58592-5_12

Mitchell C, Silver DL (2018) Enhancing our brains: Genomic mechanisms underlying cortical evolution. *Semin Cell Dev Biol* 76: 23–32. doi:10.1016/j.semcdb.2017.08.045

O'Bleness M, Searles VB, Dickens CM, Astling D, Albracht D, Mak ACY, Lai YYY, Lin C, Chu C, Graves T, et al (2014) Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics* 15: 387. doi:10.1186/1471-2164-15-387

O'Bleness MS, Dickens CM, Dumas LJ, Kehrer-Sawatzki H, Wyckoff GJ, Sikela JM (2012) Evolutionary history and genome organization of DUF1220 protein domains. *G3 (Bethesda)* 2: 977–986. doi:10.1534/g3.112.003061

Paar V, Gluncic M, Rosandic M, Basar I, Vlahovic I (2011) Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. *Mol Biol Evol* 28: 1877–1892. doi:10.1093/molbev/msr009

Paar V, Pavin N, Rosandic M, Gluncic M, Basar I, Pezer R, Zinic SD (2005) ColorHOR–novel graphical algorithm for fast scan of alpha satellite higher-order repeats and HOR annotation for GenBank sequence of human genome. *Bioinformatics* 21: 846–852. doi:10.1093/bioinformatics/bti072

Pacheco A, Issaian A, Davis JM, Anderson N, Nemkov T, Vogeli B, Hansen K, Sikela JM (2022) Proteolytic activation of human-specific Olduvai domains by the furin protease. *BioRxiv*. doi:10.1101/2021.07.06.450945 (Preprint posted July 07, 2021)

Popesco MC, Maclaren EJ, Hopkins J, Dumas L, Cox M, Meltesen L, McGavran L, Wyckoff GJ, Sikela JM (2006) Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* 313: 1304–1307. doi:10.1126/science.1127980

Quick VBS, Davis JM, Olincy A, Sikela JM (2016) DUF1220 copy number is associated with schizophrenia risk and severity: Implications for

understanding autism and schizophrenia as related diseases. *Transl Psychiatry* 6: e735. doi:10.1038/tp.2016.11

Romero V, Nakaoka H, Hosomichi K, Inoue I (2018) High order formation and evolution of hornerin in primates. *Genome Biol Evol* 10: 3167–3175. doi:10.1093/gbe/evy208

Rosandic M, Paar V, Basar I (2003) Key-string segmentation algorithm and higher-order repeat 16mer (54 copies) in human alpha satellite DNA in chromosome 7. *J Theor Biol* 221: 29–37. doi:10.1006/jtbi.2003.3165

Rudd MK, Willard HF (2004) Analysis of the centromeric regions of the human genome assembly. *Trends Genet* 20: 529–533. doi:10.1016/j.tig.2004.08.008

Sikela JM, van Roy F (2017) Changing the name of the NBPF/DUF1220 domain to the Olduvai domain. *F1000Res* 6: 2185. doi:10.12688/f1000research.13586.2

Sosic M, Sikic M (2017) Edlib: A C/C ++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* 33: 1394–1395. doi:10.1093/bioinformatics/btw753

Sullivan LL, Chew K, Sullivan BA (2017) α satellite DNA variation and function of the human centromere. *Nucleus* 8: 331–339. doi:10.1080/19491034.2017.1308989

Tyler-Smith C, Brown WR (1987) Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J Mol Biol* 195: 457–470. doi:10.1016/0022-2836(87)90175-6

Uralsky LI, Shepelev VA, Alexandrov AA, Yurov YB, Rogaev EI, Alexandrov IA (2019) Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome assembly. *Data Brief* 24: 103708. doi:10.1016/j.dib.2019.103708

Vandepoele K, Andries V, Van Roy N, Staes K, Vandesompele J, Laureys G, De Smet E, Berx G, Speleman F, van Roy F (2008) A constitutional translocation t(1;17)(p36.2;q11.2) in a neuroblastoma patient disrupts the human NBPF1 and ACCN1 genes. *PLoS One* 3: e2207. doi:10.1371/journal.pone.0002207

Vandepoele K, Van Roy N, Staes K, Speleman F, van Roy F (2005) A novel gene family NBPF: Intricate structure generated by gene duplications

during primate evolution. *Mol Biol Evol* 22: 2265–2274. doi:10.1093/molbev/msi222

Vlahović I, Glunčić M, Dekanić K, Mršić L, Jerković H, Martinjak I, Paar V (2020) Global repeat map algorithm (GRM) reveals differences in alpha satellite number of tandem and higher order repeats (HORs) in human, Neanderthal and chimpanzee genomes – novel tandem repeat database. 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia. 237–242. doi:10.23919/MIPRO48935.2020.9245278

Vlahovic I, Gluncic M, Rosandic M, Ugarkovic Đ, Paar V (2017) Regular higher order repeat structures in beetle Tribolium castaneum genome. *Genome Biol Evol* 9: 2668–2680. doi:10.1093/gbe/evw174

Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G (2008) Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* 9: 533. doi:10.1186/1471-2164-9-533

Warburton PE, Willard HF (1996) Evolution of centromeric alpha satellite DNA: Molecular organisation within and between human primate chromosomes. In *Human Genome Evolution*, Oxford: pp 121–145. BIOS Scientific Publisher.

Waye JS, Willard HF (1987) Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: A survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res* 15: 7549–7569. doi:10.1093/nar/15.18.7549

Willard HF (1985) Chromosome-specific organization of human alpha satellite DNA. *Am J Hum Genet* 37: 524–532.

Zimmer F, Montgomery SH (2015) Phylogenetic analysis supports a link between DUF1220 domain number and primate brain expansion. *Genome Biol Evol* 7: 2083–2088. doi:10.1093/gbe/evv122