scientific reports

OPEN

Tandem NBPF 3mer HORs (Olduvai triplets) in Neanderthal and two novel HOR tandem arrays in human chromosome 1 T2T-CHM13 assembly

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

It is known that the ~ 1.6 kb Neuroblastoma BreakPoint Family (NBPF) repeats are human specific and contributing to cognitive capabilities, with increasing frequency in higher order repeat 3mer HORs (Olduvai triplets). From chimpanzee to modern human there is a discontinuous jump from 0 to ~ 50 tandemly organized 3mer HORs. Here we investigate the structure of NBPF 3mer HORs in the Neanderthal genome assembly of Pääbo et al., comparing it to the results obtained for human hg38. p14 chromosome 1. Our findings reveal corresponding NBPF 3mer HOR arrays in Neanderthals with slightly different monomer structures and numbers of HOR copies compared to humans. Additionally, we compute the NBPF 3mer HOR pattern for the complete telomere-to-telomere human genome assembly (T2T-CHM13) by Miga et al., identifying two novel tandem arrays of NBPF 3mer HOR repeats with 5 and 9 NBPF 3mer HOR copies. We hypothesize that these arrays correspond to novel NBPF genes (here referred to as NBPFA1 and NBPFA2). Further improving the quality of the Neanderthal genome using T2T-CHM13 as a reference would be of great interest in determining the presence of such distant novel NBPF genes in the Neanderthal genome and enhancing our understanding of human evolution.

High-quality Neanderthal genome sequence. Impressive progress by Pääbo et al. in high-quality sequencing of Neanderthal's genome has opened new avenues in studying relation of modern humans and our closest extinct relatives Neanderthals, in quest of searching "what makes us human"[1–6]. The high-quality genomes *Denisova 5 AltaiNea.hg19*[7,8], Chagyrskaya 8[6], and *Vindija 33.19*[5] were determined. Under the assumption that Neanderthals had the same mutation rate (1.45× mutations per generation per base pair)[9] and generation time as for present-day humans (29 years), it was suggested that Chagyrskaya 8 lived ~ 30 ky after Denisova 5, and ~ 30 ky before Vindija 33.19[6]. An analysis of these high-quality genomes revealed significant changes in genes expressed in the striatum of the brain, indicating the potential evolution of unique functions in the Neanderthal brain[6].

Human specific ~ 1.6 kb tandem repeat units in NBPF genes. Neuroblastoma is a solid malignancy that primarily affect children and has been the focus of intense research[10,11]. The NBPF gene family was originally identified by the disruption of one of its members in a neuroblastoma patient[12]. The NBPF genes are located on human chromosome 1 and contain a repetitive structure of ~ 1.6 kb tandem repeat units known as Olduvai domains (also called NBPF domains, NBPF repeats, or DUF1220 domains), which code for Olduvai protein domains (previously called DUF1220)[13–19] involved in human brain evolution. The term Olduvai for these repeat structures is referred to as Sikela–van Roy terminology[19]. Alternatively, in accordance with Willard's terminology used for tandem repeats in centromeric region of human genome[20–22], the repeat units in NBPF sequences were referred to as NBPF monomers[23].

[1]Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia. [2]Algebra University College, 10000 Zagreb, Croatia. [3]University Hospital Centre Zagreb (Ret.), 10000 Zagreb, Croatia. [4]Croatian Academy of Sciences and Arts, 10000 Zagreb, Croatia. ✉email: matko@phy.hr

Studies have found that the copy number of Olduvai domains is correlated with various aspects of brain function and pathology, including brain size, cortical neuron number, IQ scores, cognitive aptitude, autism, schizophrenia, microcephaly, macrocephaly, and neuroblastoma[13,15,18,24–33]. The association between HLS Olduvai domain copy number and the human brain evolution with increased cognitive function was suggested by Sikela et al.[15,19].

Interestingly, the copy number of NBPF domains in nonhuman species generally decreases with increasing phylogenetic distance from humans, with humans having the highest number (~ 300 copies) followed by great apes (~ 38–97 copies), monkeys (~ 48–75 copies), and non-primate mammals (~ 1–8 copies), while these domains are mostly absent in non-mammalian species[15,17,18,27,29,34,35].

In our previous research, we found that the human specificity of NBPF copy number variation is significantly more pronounced in tandemly organized NBPF 3mer higher order repeats (HORs) than in individual NBPF ~ 1.6 kbp monomer copies[23]. Specifically, we observed a copy number of 47 HORs in humans, whereas chimpanzees, gorillas, orangutans, and rhesus macaques showed zero copy number. Recent computations using higher quality sequencing, specifically the hg38.p14 human reference genome and NC_036879.1 chimpanzee ensemble, have yielded similar results[36]. Based on these findings, we have hypothesized that the tandemly organized ~ 4.8 kb NBPF 3mer HOR copy number may provide an additional evolutionary signature, in conjunction with the individual ~ 1.6 kb NBPF primary repeat/Olduvai domain copy number effect, potentially leading to a coherent overall effect.

In this study, we aim to compare the copy number of NBPF tandemly repeated HORs in Neanderthals and humans, as well as in comparison to the human complete T2T-CHM13 assembly[37–39] and chimpanzee reference Pan troglodytes NHGRI_mPanTro3-v1.1-hic.freeze_pri (CM054434.1 Pan troglodytes isolate AG18354 chromosome 1, whole genome shotgun sequence). Our findings could shed light on the role of NBPF genes in human evolution, as well as on the genetic differences between Neanderthals and modern humans.

The ~ 171 bp alpha satellite monomers and alpha satellite nmer higher order repeats (HORs) in centromeres of human genomes: a HOR prototype.

Most pronounced tandem repeats in human genome are alpha satellites located in the centromeric regions of chromosomes. These repeats consist of ~ 30–50% diverged ~ 171 bp alpha satellite monomer units, which serve as the primary repeats[40,41]. Frequently, these alpha satellites are organized into higher order repeats (HORs)[22]. HORs are composed of monomers arranged in multimeric repeat HOR copies that are tandemly positioned. The level of divergence between HOR copies is very small, often less than 5%, which is an order of magnitude smaller than the divergence observed between neighboring monomers[21,22,37,38,41–49]. These primary repeats and tandem HOR repeats are commonly referred to as Willard's terminology. The repetitive structure of alpha satellite repeats and the organization of HOR arrays have significant implications in the study of chromosome biology, genome instability, evolution, and human disease. They serve as a source of genetic and epigenetic variation, contributing to the dynamic nature of the genome[38,50].

NBPF 3mer HORS/Olduvai triplets.

In 2011, we applied our robust HOR-searching algorithm, GRM (Global Repeat Map algorithm), to the Build 36.3 human genome assembly for chromosome 1, which led to the discovery of tandemly organized ~ 4.8-kb 3mer HOR copies in NBPF genes[23]. The GRM algorithm revealed that each 3mer HOR copy is composed of three ~ 1.6 kb NBPF monomers, denoted m1, m2 and m3, respectively[23].

In 2012, the same pattern was recognized through another method, which involved analyzing the similarity between the ~ 1.6 kb Olduvai domains present in NBPF genes (Fig. 1). This analysis revealed that the ~ 1.6 kb Olduvai domains are predominantly organized in repeating triplets with minimal divergence between the triplets[16,17]. Previously referred to as HLS/DUF1220 triplets in Sikela–van Roy terminology, and consist of three domains designated as HLS1, HLS2, and HLS3, these triplets have recently been renamed Olduvai triplets[19]. Table 1 presents the correspondence between different names for the same repeat pattern obtained through different computational methods.

Willard terminology was used generally for *n* monomers in *n*mer HOR copies with *n* distinct monomers of types m1, m2, … m*n*[21,22]. For the NBPF repeats, Willard terminology is applied here to the special case of n = 3, with three types of ~ 1.6 kb NBPF monomers m1, m2, and m3. In the computational method for HOR identification, Willard terminology is characterized by identification of ~ 4.8 kb NBPF 3mer HOR copies (i.e., Olduvai triplets) in the first step of computation, and their constituent monomers m1, m2 and m3 are identified in the second step within HORs[23]. On the other hand, Sikela–van Roy terminology is characterized by identification of Olduvai domains (i.e., 3mer NBPF monomers) in the first step of computation, and their corresponding Olduvai triplets (i.e., 3mer HOR copies) identified on the basis of divergence among Olduvai domains.

Additionally, it is worth noting that, besides the NBPF 3mer HOR, the same HOR-searching GRM computation also identified two others prominent HORs in human chromosome 1[23]. In hornerin genes, novel quartic HORs were discovered, consisting of primary, secondary, tertiary, and quartic repeats with lengths of approximately ~ 39 bp, ~ 0.35 kb, ~ 0.7 kb, and ~ 1.4 kb, respectively[23].

Moreover, the pronounced well known centromeric HOR pattern in the centromeric region of chromosome 1 is the canonical alpha satellite 11mer HOR[20,44,46,52] with suprachromosomal assignment SF2[52]. Recently, a canonical 6mer HOR with suprachromosomal assignment SF1 was identified[37,48]. Here, using the GRM algorithm both the 11mer and 6mer canonical HORs were identified in hg38. In the GRM result, the divergence between HOR copies in 11mer HOR array is a few times smaller than between 6mer copies, corresponding to homogene and divergent HORs, respectively.

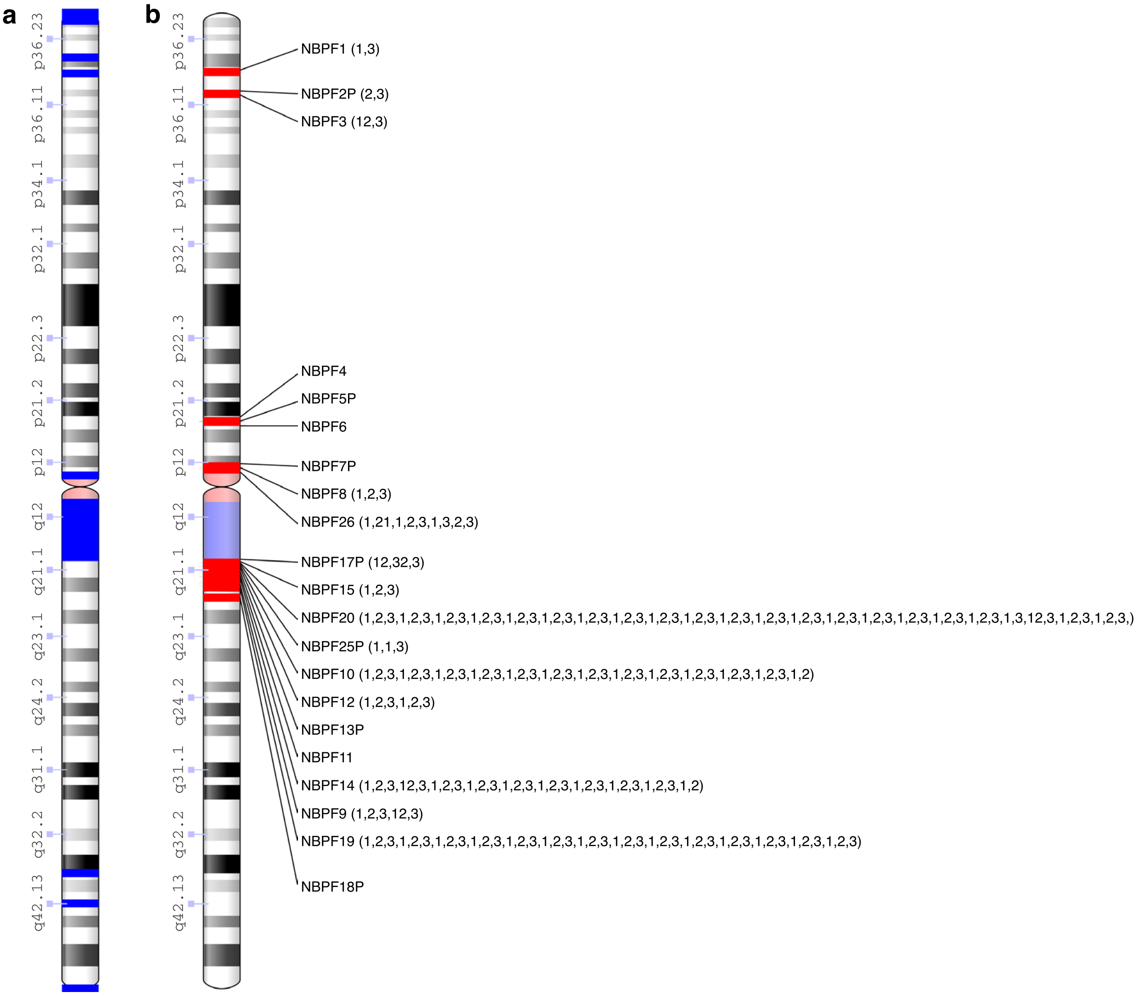
a
b
p36.23
p36.11
p34.1
p32.1
p22.3
p21.2
p12
q12
q21.1
q23.1
q24.2
q31.1
q32.2
q42.13
NBPF1 (1,3)
NBPF2P (2,3)
NBPF3 (12,3)
NBPF4
NBPF5P
NBPF6
NBPF7P
NBPF8 (1,2,3)
NBPF26 (1,21,1,2,3,1,3,2,3)
NBPF17P (12,32,3)
NBPF15 (1,2,3)
NBPF20 (1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,3,12,3,1,2,3,1,2,3,)
NBPF25P (1,1,3)
NBPF10 (1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2)
NBPF12 (1,2,3,1,2,3)
NBPF13P
NBPF11
NBPF14 (1,2,3,12,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2)
NBPF9 (1,2,3,12,3)
NBPF19 (1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3)
NBPF18P

Figure 1. Maps of gaps, NBPF gene members, and their content of Olduvai repeats in the human chromosome 1 genome hg38 assembly. (**a**) Positions of gaps (blue rectangles) (indicated in the assembly as 'N's) larger than 2000 bp. (**b**) Positions of NBPF gene members (red rectangles) (according to the National Center for Biotechnology Information) and arrangement of Olduvai domain subtypes[51]. Only HLS1 (1), HLS2 (2), and HLS3 (3) Olduvai domains subtypes are indicated. Domains consisting of two different HLS types are denoted with two connected numbers (e.g., the second domain in NBPF26 consists of half HLS2 and half HLS1, is indicated as 21). It is evident that the hg38 assembly has a gap at the position of the NBPF1 and NBPF7P genes.

Sikela-van Roy terminology	↔	Willard terminology
DUF1220 domains/Olduvai domains	↔	NBPF monomers
HLS-1	↔	m1
HLS-2	↔	m2
HLS-3	↔	m3
DUF1220 triplets/Olduvai triplets	↔	NBPF 3mer HOR copies

Table 1. Correspondence of Sikela-van Roy terminology and Willard terminology for NBPF repeats.

Results and discussion

Exclusively human-specific GRM diagrams. In the first step, the GRM diagrams for chromosome 1 in Neanderthal's AltaiNea.hg19 assembly[4] and in recent complete human T2T assembly T2T-CHM13[37–39,53] are computed (Fig. 2b,c, respectively). These results are then compared to GRM diagrams for hg38.p14 (NC_000001.11) assembly of the human chromosome 1 and other nonhuman assemblies from Ref.[36], namely the chimpanzee assembly Clint_PTRv2 (RefSeq sequence NC_036879.1) (Fig. 2a), gorilla assembly Kamilah_GGO_v0 (RefSeq sequence NC_044602.1), orangutan assembly Susie_PABv2 (RefSeq sequence NC_036903.1) and rhesus macaque assembly Mmul_10 (RefSeq sequence NC_041754.1). The GRM peak at ~ 4.8 kb corresponds to

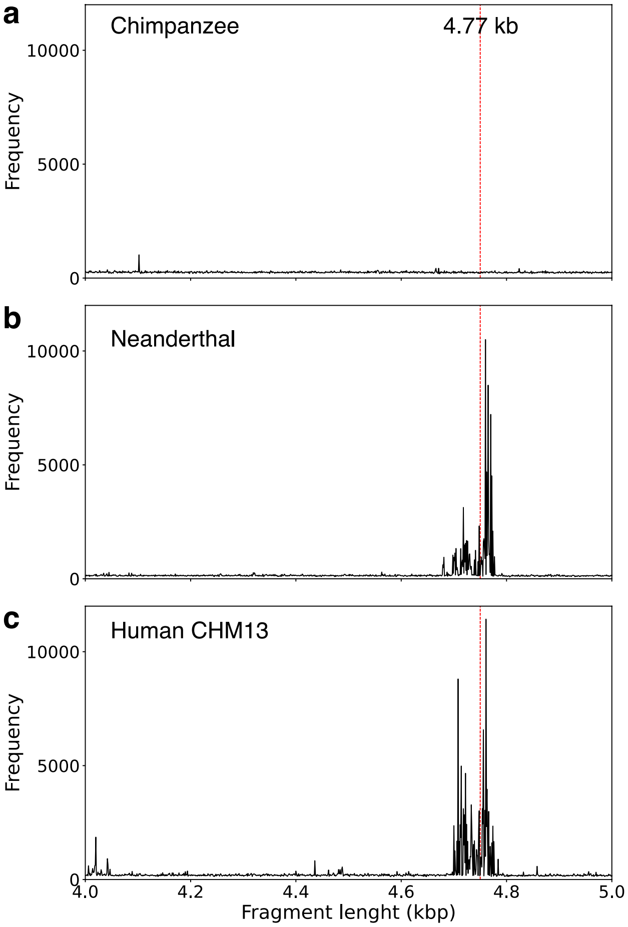

Figure 2. GRM diagrams for 140–150 Mb segment of chromosome 1: (**a**) Chimpanzee *NHGRI_mPanTro3-v1.1*; (**b**) Neanderthal AltaiNea.hg19, and (**c**) Complete human T2T CHM13 assembly. GRM diagrams have pronounced GRM peaks at ~ 4.8 kb for Neanderthal and human genomes, while for chimpanzee the peak at ~ 4.8 kb is absent.

the ~ 4.8 kb tandemly organized canonical NBPF 3mer HOR copies (m1m2m3), which are based on the ~ 1.6 kb NBPF monomers of types m1, m2, and m3. Consensus NBPF monomers m1, m2 and m3 for Neanderthal, hg38 human and T2T-CHM13 human assemblies are given in the Supplementary Table 1.

NBPF 3mer HOR copy aligned schemes. Figure 3 compares the Global Repeat Map (GRM) results for aligned NBPF 3mer higher order repeat (HOR) copies, including canonical and variant copies, in chromosome 1 across different genome assemblies. The panels show the results for the chimpanzee NHGRI_mPanTro3-v1.1 assembly (1st panel), Neanderthal AltaiNea.hg19 assembly (2nd panel), human hg38.p14 assembly (3rd panel), and human complete T2T-CHM13 assembly (4th panel). To construct these aligned NBPF HOR schemes, we first present the aligned NBPF 3mer HOR copies in the human hg38.p14 assembly (3rd panel), which are similar to previous results computed for HLS domains[16] and GRM results[23,36]. Some small differences are due to variations in the computational methods used to identify primary repeats of ~ 1.6 kb and/or ~ 4.8 kb secondary HOR repeats, as well as the quality of the sequenced genome. In this presentation, the NBPF monomers are horizontally grouped into HOR copies (i.e., within each 3mer HOR copy—canonical or variant), which are then aligned vertically. A blank space is inserted between any two neighboring groups of tandemly organized HOR copies, treating individual isolated monomers of types m1, m2 or m3 as single-monomer HOR copies.

Human hg38.p14 panel. The human hg38.p14 (3rd panel) is characterized by presence of four pronounced tandemly organized arrays of NBPF 3mer HOR copies (canonical and variant). These arrays include a 22-copy array, 13-copy array, 10-copy array, and 14-copy array, respectively, (Table 2) with at least two neighboring canonical HOR copies arranged in tandem (Fig. 3). These arrays correspond to the four prominent NBPF genes found on human chromosome 1: NBPF20, NBPF14, NBPF10, and NBPF19, respectively. Combined, these four NBPF genes encompass a total of 59 NBPF 3mer HOR copies (52 of which are canonical). It is important to mention that the order of these four tandem arrays is reversed compared to the ordering specified in Ref.[16]. To facilitate visual clarity, each tandem array is color-coded: green, red, blue, and orange, respectively. Beyond

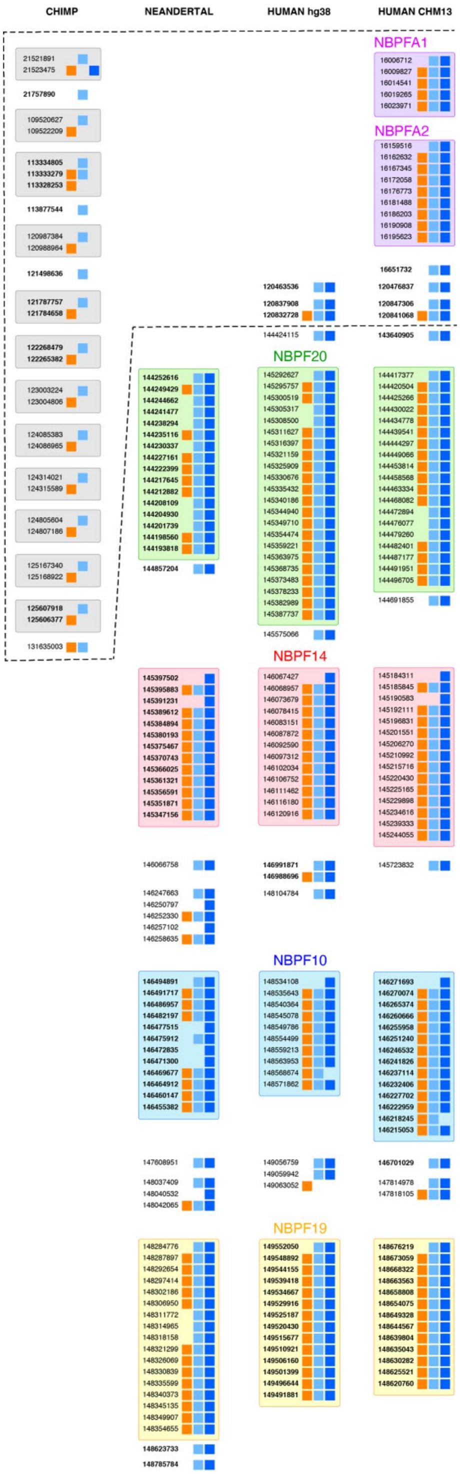

Figure 3. Comparison of NBPF monomer alignment scheme for NBPF 3mer HOR copies for chromosome 1 in chimpanzee *NHGRI_mPanTro3-v1.1* (1st panel), *Neanderthal AltaiNea.hg19* (2nd panel), *human hg38.p14* (3rd panel) and complete *human T2T-CHM13* (4th panel). Boxes represent three types of NBPF monomers, denoted m1 (orange), m2 (light blue) and m3 (blue). Each row of boxes (three in canonical HOR copy, two or one in variant HOR copies) represents an NBPF HOR copy. In front of each row (HOR copy) its start position in the genomic sequence is given. Initial positions that are in bold indicate monomers that appear in the original sequence in the reverse-complement orientation. Each array of tandemly arranged HOR copies is separated from neighboring arrays and/or isolated individual monomers by blank space. Consensus monomers determined by GRM algorithm for human and Neanderthal genome are almost the same (average divergence less than 4%). A characteristics of chimpanzee NHGRI_*mPanTro3-v1.1* assembly is the absence of canonical HOR copies. Moreover, except in the second variant HOR copy m1m3 (at position 21,523,475 bp), the m3 NBPF monomer is missing. A pronounced additional segment, identified only for the T2T-CHM13 assembly, is the appearance of two novel tandem arrays of NBPF 3mer HOR copies, the 5-copy and 9-copy arrays, interpreted as belonging to the novel genes named NBPFA1 and NBPFA2 genes.

	No. of canonical and variant HORs gene NBPF						Total no. of canonical HORs
Assembly	20	14	10	19	A1	A2	
Chimpanzee NHGRI_mPanTro3-v1.1							0
Neanderthal AltaiNea.hg19	16	13	12	17			39
Human hg38.p14	22	13	8	13			52
Human T2T-CHM13	19	15	14	13	5	9	64

Table 2. Number NBPF 3mer HOR copies in tandemly organized 3mer HOR copies (canonical and variant) in chromosome 1 of Neanderthal AltaiNea.hg19, human hg38.p14 and complete human T2T-CHM13 assemblies (deduced from Fig. 3).

these four HOR copy arrays, there are an additional seven smaller groups of scattered HOR copies. Three of these scattered groups are positioned above the 22-copy array corresponding to the NBPF20 gene.

Horizontal alignment of 2nd, 3rd and 4th panels. In the next step, we adjusted the relative positions of the four prominent NBPF 3mer HOR tandems in panels 2, 3, and 4 of Fig. 3. This was done by aligning the first rows of each NBPF gene horizontally while maintaining the distances within the tandem arrays. By employing this spreading technique, the first rows corresponding to the NBPF20 gene in Neanderthal AltaiNea.hg19, human hg38.p14, and human T2T-CHM13 were horizontally aligned. Similarly, the other three notable NBPF genes (NBPF14, NBPF10, and NBPF19) were horizontally aligned as well. As a result of this spreading based on the positions of the four significant NBPF genes within each panel, the top sections of the hg38.p14 (3rd panel) and CHM13 (4th panel) are positioned above the NBPF20 gene (green area). The top section of the 3rd panel (for hg38.p14) exhibits three distinct sets of HOR copies (2-monomer variant HOR copy, canonical HOR copy + 2-monomer variant HOR copy, 2-monomer variant HOR copy).

Neanderthal AltaiNea.hg19 panel. The Neanderthal AltaiNea.hg19 (2nd panel) exhibits four distinct and tandemly organized arrays of NBPF 3mer HOR copies (canonical plus variant) in a top-to-bottom manner: 16-copy, 13-copy, 12-copy, and 17-copy, respectively (Fig. 3, Table 2). These arrays correspond to four prominent NBPF genes located in human chromosome 1: NBPF20, NBPF14, NBPF10, and NBPF19. Altogether, these four NBPF genes contribute to a total of 58 NBPF 3mer HOR copies (39 canonical).

When comparing the alignment of NBPF 3mer HORs between Neanderthal AltaiNea.hg19 and human hg38.p14, the tandem HOR groups from the corresponding NBPF genes appear to have a similar order. However, it is important to consider that the Neanderthal genome was sequenced based on a comparison with an earlier human assembly, and despite its relatively high-quality sequencing, there were still gaps in the genome sequences. In Fig. 1a the positions of sequence gaps discussed in the region of NBPF genes are presented. Notably, human hg38.p14 has a sequencing gap of 18 Mb, which is located in proximity to the position of the NBPF20 gene. Additionally, there are several other sequencing gaps of approximately 50 kb that are further away from the NBPF genes. In the Neanderthal AltaiNea.hg19 assembly, there is a 21 Mb gap near the positions of NBPF genes, along with several other gaps of around 150 kb, 100 kb, and 50 kb close to the region of NBPF genes. Consequently, while comparing the results between Neanderthal AltaiNea.hg19 and human hg38.p14 is reasonable, it is possible that some HOR copies have been missed in the sequencing process due to the presence of more gaps in the region near the four 3mer HOR copy-rich arrays in Neanderthal.

Human T2T-CHM13 panel. The complete human T2T-CHM13 assembly, represented in the 4th panel, displays four prominent and tandemly organized NBPF 3mer HOR copies that roughly align with the pattern observed in hg38.p14 (3rd panel): a 19-copy array (green area), a 15-copy array (red), a 14-copy array (blue), and a 13-copy array (orange) (Fig. 3, Table 2). This ordering of tandemly organized NBPF HOR-copy arrays roughly corresponds to the NBPF genes 20, 13, 10 and 19, respectively, similar to the hg38.p14 case. Additionally, we identified eight scattered small groups of HOR copies outside the four major HOR copy arrays in the T2T-CHM13 sequence. Notably, above the 19-copy array (green area), there are three small, scattered groups exhibiting the same HOR pattern as observed in hg38.p14 (2-monomer variant HOR copy, canonical HOR copy + 2-monomer variant HOR copy, 2-monomer variant HOR copy). However, at the top of the 4th panel, two additional tandemly organized arrays of NBPF 3mer HOR copies are present: a 5-mer array and a 9-mer array (violet area). We tentatively assign these NBPF tandem arrays to two genes, designated as NBPFA1 and NBPFA2, respectively. These arrays are positioned relatively close to the telomeric region, located more than 100 Mb away from the known NBPF genes. The total number of constituting NBPF 3mer HOR copies (canonical plus variant) in the T2T-CHM13 assembly is 75 (64 canonical). All NBPF genes with tandemly organized Olduvai triplets (canonical NBPF 3mer HOR copies) are located in the 1q region[16]. However, several NBPF genes are also known to be located also in the 1p region, but they lack tandemly organized Olduvai triplets. For instance, in the hg38.p14 assembly, the gene NBPF1 in 1p36.13 contains no Olduvai triplet, while the gene NBPF8 in 1p11.2 contains only one Olduvai triplet[16]. The two novel NBPF genes with tandemly organized Olduvai triplets, discovered here in the CHM13 assembly and absent in hg38.p14, are referred to as NBPFA1 and NBPFA2. As the Neanderthal AltaiNea.hg19 assembly was sequenced based on a comparison to the previous human reference genome similar to hg38.p14, it is expected that these two novel NBPF genes are also absent in the Neanderthal AltaiNea.hg19 assembly.

Conclusion

It is shown that the abundant tandemly organized NBPF 3mer HOR copies in complete human assembly T2T-CHM13 and in Neanderthal AltaiNea.hg19 assembly of chromosome 1 are exclusively human (homo sapiens and Neanderthal) specific, because they are completely absent in chimpanzee chromosome 1. In addition, we identified two new tandemly organized arrays of NBPF 3mer HOR copies in the complete human T2T-CHM13 assembly which we assign to two new genes. On the other hand, Sikela and collaborators have shown that the number of NBPF repeats, which is only about twice as high in humans compared to chimpanzees, correlates with the gradual increase in primate cognitive abilities[24,26,27]. We hypothesize that the increase of cognitive abilities is coherently increased by tandemly organized NBPF higher order repeats (HORs) which are highly present in human and Neanderthal genomes and absent in the chimpanzee genome. In conclusion, our findings highlight the need for an additional sequencing of the Neanderthal genome, utilizing the T2T-CHM13 human genome as a reference. This proposed approach holds the potential to provide more accurate and detailed information, which would play a crucial role in advancing our understanding of the evolution of cognitive abilities in both human and Neanderthal populations. Particularly intriguing is the contrast with chimpanzees and other primates, as they entirely lack tandemly organized NBPF HORs, further underscoring the significance of conducting comparative genomic analysis. By shedding light on the genetic factors that contributed to cognitive development in our evolutionary history, this research could offer valuable insights into the shared and distinct traits between humans and Neanderthals. Ultimately, such knowledge may enrich our comprehension of the complexities of human evolution and contribute to broader discussions in the field of evolutionary biology.

It is interesting to note that a sophisticated phenomenon of cognitive development is related to the underlying tandem HOR pattern, which is based on the concept of DNA symmetries. This pattern represents an evolutionary trajectory characterized by symmetries, resulting in increased order and a reduction in information entropy.[36,54,55].

Materials and methods

The NBPF HORs were identified in the NCBI assembly (2023) of human, pan troglodytes and Neanderthal genomes using the GRM algorithm[56–58]. The GRM algorithm is an efficient and robust method specifically designed to detect and analyze very large repeat units, such as HORs, within genomic sequences. It effectively reduces computational noise associated with detecting longer and more complex HOR repeat units, ensuring the accurate identification of peaks corresponding to HOR copies. Unlike other methods, the GRM approach directly maps symbolic DNA sequences into the frequency domain using a complete K-string ensemble, avoiding the need for statistical adjustments and local optimization of individual K-strings. This unique feature allows for straightforward identification of DNA repeats in the frequency domain without the need for mapping symbolic DNA sequences to numerical sequences. The GRM algorithm demonstrates robustness in handling deviations from ideal repeats, making it suitable for repeats with substitutions, insertions, and deletions. Additionally, it provides parameter-free identification of repeats, enabling the determination of consensus lengths and consensus sequences for primary repeats and HORs. The GRM method generates a global repeat map in a GRM diagram, identifying all prominent repeats in a given sequence without any prior knowledge of the repeats. Furthermore, once the consensus repeat unit is determined using GRM, it can be further combined with a search for dispersed HOR copies or individual constituting monomers.

Specifically, NBPF HORs in this study were identified through the following steps:

(i) Using GRMapp (the GRM graphical user interface application is freely available at http://genom.hazu.hr/tools.html), NBPF monomers were identified within the entire human hg38, human CHM13, Neandertal and chimpanzee chromosome 1 assemblies. GRMapp provides all tandem repeats (TRs) in the analyzed assembly as its output. From the list of all TRs, those with lengths of ~ 1.6 kb in the region of the NBPF genes were selected and subjected to GRM diagram analysis within GRMapp. To be classified as NBPF monomers, the GRM diagram must exhibit peaks at ~ 1.6 kb and multiples at ~ 3.2 kb and ~ 4.8 kb, indicating the existence of higher-order structures (HORs).

(ii) The extracted NBPF monomers were compared to each other, and a divergence matrix was created. From the divergence matrix, monomer families were identified, encompassing all monomers that differ from each other by less than 5%. In this manner, three monomer families, m1, m2, and m3, were obtained.

(iii) For each monomer family, a consensus sequence was generated using the stand-alone tool for multiple-sequence alignment, pyabPOA (pyabpoa 1.0.0a0), available at https://github.com/yangao07/abpoa. The consensus sequences for m1, m2, and m3 are provided in Supplementary Table 1.

(iv) Each chromosome 1 assembly (chimp, Neanderthal, human hg38, and human T2T-CHM13) was searched with all three consensus sequences, m1, m2, and m3, using the Edlib open-source C/C++ library for exact pairwise sequence alignment[59]. The search was conducted base by base for the entire chromosome, considering both the direct and reverse complement consensus sequences. The resulting tables for each organism are provided in Supplementary Table 2.

(v) The results of the search in step (iv) were presented graphically (Fig. 3) in a way that all monomers of the same family (m1, m2, m3) are located in the same column and colored with the same color.

Data availability

All genomic sequences are freely available at the National Center for Biotechnology Information (NCBI) website https://www.ncbi.nlm.nih.gov. The GRM graphical user interface application (JAR file) is freely available at our project's website http://genom.hazu.hr/tools.html. It can be run on any platform which have Java Runtime Environment (JRE) installed (freely available at https://www.oracle.com/java/technologies/javase-downloads.html).

Received: 24 May 2023; Accepted: 28 August 2023
Published online: 02 September 2023

References

1. Paabo, S. *et al.* Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38**, 645–679. https://doi.org/10.1146/annurev.genet.37.110801.143214 (2004).
2. Noonan, J. P. & McCallion, A. S. Genomics of long-range regulatory elements. *Annu. Rev. Genom. Hum. Genet.* **11**, 1–23. https://doi.org/10.1146/annurev-genom-082509-141651 (2010).
3. Kelso, J. & Prufer, K. Ancient humans and the origin of modern humans. *Curr. Opin. Genet. Dev.* **29**, 133–138. https://doi.org/10.1016/j.gde.2014.09.004 (2014).
4. Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49. https://doi.org/10.1038/nature12886 (2014).
5. Prufer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658. https://doi.org/10.1126/science.aao1887 (2017).
6. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 15132–15136. https://doi.org/10.1073/pnas.2004944117 (2020).
7. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722. https://doi.org/10.1126/science.1188021 (2010).
8. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060. https://doi.org/10.1038/nature09710 (2010).
9. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449. https://doi.org/10.1038/nature13810 (2014).
10. Maris, J. M. & Matthay, K. K. Molecular biology of neuroblastoma. *J. Clin. Oncol.* **17**, 2264–2279. https://doi.org/10.1200/JCO.1999.17.7.2264 (1999).
11. Van Roy, N. *et al.* The emerging molecular pathogenesis of neuroblastoma: Implications for improved risk assessment and targeted therapy. *Genome Med.* **1**, 74. https://doi.org/10.1186/gm74 (2009).
12. Vandepoele, K. *et al.* 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. https://doi.org/10.1371/journal.pone.0002207 (2008).
13. Vandepoele, K., Van Roy, N., Staes, K., Speleman, F. & van Roy, F. A novel gene family NBPF: Intricate structure generated by gene duplications during primate evolution. *Mol. Biol. Evol.* **22**, 2265–2274. https://doi.org/10.1093/molbev/msi222 (2005).
14. Fortna, A. *et al.* Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* **2**, E207. https://doi.org/10.1371/journal.pbio.0020207 (2004).
15. Popesco, M. C. *et al.* Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* **313**, 1304–1307. https://doi.org/10.1126/science.1127980 (2006).
16. O'Bleness, M. *et al.* Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genom.* **15**, 387. https://doi.org/10.1186/1471-2164-15-387 (2014).
17. O'Bleness, M. S. *et al.* Evolutionary history and genome organization of DUF1220 protein domains. *G3* **2**, 977–986. https://doi.org/10.1534/g3.112.003061 (2012).
18. Heft, I. E. *et al.* The driver of extreme human-specific Olduvai repeat expansion remains highly active in the human genome. *Genetics* **214**, 179–191. https://doi.org/10.1534/genetics.119.302782 (2020).
19. Sikela, J. M. & van Roy, F. Changing the name of the NBPF/DUF1220 domain to the Olduvai domain. *F1000Research* **6**, 2185. https://doi.org/10.12688/f1000research.13586.2 (2017).
20. Waye, J. S. & Willard, H. F. Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: A survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res.* **15**, 7549–7569 (1987).
21. Willard, H. F. Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.* **37**, 524–532 (1985).
22. Warburton, P. E. & Willard, H. F. *Human Genome Evolution* 121–145 (BIOS Scientific Publisher, 1996).
23. Paar, V., Gluncic, M., Rosandic, M., Basar, I. & Vlahovic, I. Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. *Mol. Biol. Evol.* **28**, 1877–1892. https://doi.org/10.1093/molbev/msr009 (2011).
24. Dumas, L. & Sikela, J. M. DUF1220 domains, cognitive disease, and human brain evolution. *Cold Spring Harb. Symp. Quant. Biol.* **74**, 375–382. https://doi.org/10.1101/sqb.2009.74.025 (2009).
25. Dumas, L. J. *et al.* DUF1220-domain copy number implicated in human brain-size pathology and evolution. *Am. J. Hum. Genet.* **91**, 444–454. https://doi.org/10.1016/j.ajhg.2012.07.016 (2012).
26. Davis, J. M. *et al.* DUF1220 dosage is linearly associated with increasing severity of the three primary symptoms of autism. *PLoS Genet.* **10**, e1004241. https://doi.org/10.1371/journal.pgen.1004241 (2014).
27. Keeney, J. G., Dumas, L. & Sikela, J. M. The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion. *Front. Hum. Neurosci.* **8**, 427. https://doi.org/10.3389/fnhum.2014.00427 (2014).
28. Andries, V. *et al.* NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest. *BMC Cancer* **15**, 391. https://doi.org/10.1186/s12885-015-1408-5 (2015).
29. Zimmer, F. & Montgomery, S. H. Phylogenetic analysis supports a link between DUF1220 domain number and primate brain expansion. *Genome Biol. Evol.* **7**, 2083–2088. https://doi.org/10.1093/gbe/evv122 (2015).
30. Quick, V. B., Davis, J. M., Olincy, A. & Sikela, J. M. DUF1220 copy number is associated with schizophrenia risk and severity: Implications for understanding autism and schizophrenia as related diseases. *Transl. Psychiatry* **6**, e735. https://doi.org/10.1038/tp.2016.11 (2016).
31. Astling, D. P., Heft, I. E., Jones, K. L. & Sikela, J. M. High resolution measurement of DUF1220 domain copy number from whole genome sequence data. *BMC Genom.* **18**, 614. https://doi.org/10.1186/s12864-017-3976-z (2017).
32. Mitchell, C. & Silver, D. L. Enhancing our brains: Genomic mechanisms underlying cortical evolution. *Semin. Cell Dev. Biol.* **76**, 23–32. https://doi.org/10.1016/j.semcdb.2017.08.045 (2018).
33. Fiddes, I. T., Pollen, A. A., Davis, J. M. & Sikela, J. M. Paired involvement of human-specific Olduvai domains and NOTCH2NL genes in human brain evolution. *Hum. Genet.* **138**, 715–721. https://doi.org/10.1007/s00439-019-02018-4 (2019).
34. Dumas, L. *et al.* Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res.* **17**, 1266–1277. https://doi.org/10.1101/gr.6557307 (2007).

35. Andries, V., Vandepoele, K. & van Roy, F. The NBPF gene family. In *Neuroblastoma—Present and Future* (ed. Shimada, H.) 185–214 (InTech, 2012).
36. Gluncic, M., Vlahovic, I., Rosandic, M. & Paar, V. Tandemly repeated NBPF HOR copies (Olduvai triplets): Possible impact on human brain evolution. *Life Sci. Alliance* **6**, 306. https://doi.org/10.26508/lsa.202101306 (2023).
37. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, 4178. https://doi.org/10.1126/science.abl4178 (2022).
38. Miga, K. H. & Alexandrov, I. A. Variation and evolution of human centromeres: A field guide and perspective. *Annu. Rev. Genet.* **55**, 583–602. https://doi.org/10.1146/annurev-genet-071719-020519 (2021).
39. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53. https://doi.org/10.1126/science.abj6987 (2022).
40. Manuelidis, L. Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma* **66**, 23–32 (1978).
41. Willard, H. F. & Waye, J. S. Chromosome-specific subsets of human alpha satellite DNA: Analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.* **25**, 207–214 (1987).
42. Mellone, B. G. & Fachinetti, D. Diverse mechanisms of centromere specification. *Curr. Biol.* **31**, R1491–R1504. https://doi.org/10.1016/j.cub.2021.09.083 (2021).
43. Paar, V., Basar, I., Rosandic, M. & Gluncic, M. Consensus higher order repeats and frequency of string distributions in human genome. *Curr. Genom.* **8**, 93–111 (2007).
44. Paar, V. *et al.* 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. https://doi.org/10.1093/bioinformatics/bti072 (2005).
45. Vlahovic, I., Gluncic, M., Rosandic, M., Ugarkovic, E. & Paar, V. Regular higher order repeat structures in beetle *Tribolium castaneum* genome. *Genome Biol. Evol.* **9**, 2668–2680. https://doi.org/10.1093/gbe/evw174 (2017).
46. Warburton, P. E. *et al.* Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genom.* **9**, 533. https://doi.org/10.1186/1471-2164-9-533 (2008).
47. Sullivan, L. L., Chew, K. & Sullivan, B. A. Alpha satellite DNA variation and function of the human centromere. *Nucleus* **8**, 331–339. https://doi.org/10.1080/19491034.2017.1308989 (2017).
48. Uralsky, L. I. *et al.* 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. https://doi.org/10.1016/j.dib.2019.103708 (2019).
49. Feliciello, I., Pezer, Z., Kordis, D., Bruvo Madaric, B. & Ugarkovic, D. Evolutionary history of alpha satellite DNA repeats dispersed within human genome Euchromatin. *Genome Biol. Evol.* **12**, 2125–2138. https://doi.org/10.1093/gbe/evaa224 (2020).
50. Rudd, M. K., Wray, G. A. & Willard, H. F. The evolutionary dynamics of alpha-satellite. *Genome Res.* **16**, 88–96. https://doi.org/10.1101/gr.3810906 (2006).
51. Sikela, J. M. & Searles Quick, V. B. Genomic trade-offs: Are autism and schizophrenia the steep price of the human brain? *Hum. Genet.* **137**, 1–13. https://doi.org/10.1007/s00439-017-1865-9 (2018).
52. Rosandic, M. *et al.* CENP-B box and pJalpha sequence distribution in human alpha satellite higher-order repeats (HOR). *Chromosome Res.* **14**, 735–753. https://doi.org/10.1007/s10577-006-1078-x (2006).
53. Miga, K. H. The promises and challenges of genomic studies of human centromeres. *Prog. Mol. Subcell. Biol.* **56**, 285–304. https://doi.org/10.1007/978-3-319-58592-5_12 (2017).
54. Rosandic, M. & Paar, V. Standard genetic code vs supersymmetry genetic code—Alphabetical table vs physicochemical table. *Biosystems* **218**, 104695. https://doi.org/10.1016/j.biosystems.2022.104695 (2022).
55. Rosandic, M., Vlahovic, I., Pilas, I., Gluncic, M. & Paar, V. An explanation of exceptions from Chargaff's second parity rule/strand symmetry of DNA molecules. *Genes* **13**, 929. https://doi.org/10.3390/genes13111929 (2022).
56. Gluncic, M. & Paar, V. Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. *Nucleic Acids Res.* **41**, e17. https://doi.org/10.1093/nar/gks721 (2013).
57. Gluncic, M., Vlahovic, I. & Paar, V. Discovery of 33mer in chromosome 21-the largest alpha satellite higher order repeat unit among all human somatic chromosomes. *Sci. Rep. U. K.* **9**, 2. https://doi.org/10.1038/s41598-019-49022-2 (2019).
58. Vlahović, I. *et al.* 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. In *43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia* 237–242. https://doi.org/10.23919/MIPRO48935.2020.9245278 (2020).
59. Sosic, M. & Sikic, M. Edlib: A C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* **33**, 1394–1395. https://doi.org/10.1093/bioinformatics/btw753 (2017).

Acknowledgements

This work was supported by the QuantiXLie Centre of Excellence, a project cofinanced by the Croatian Government and European Union through the European Regional Development Fund—the Competitiveness and Cohesion Operational Programme (Grant KK.01.1.1.01.0004), and the Grant IP-2019-04-2757 from Croatian Science Foundation. The authors thank to Janet Kelso for discussion on Neanderthal genomics.

Author contributions

M.G. and I.V. performed computations and analysis of results. M.R. performed analysis of results. V.P. initiated the project and M.G. guided the work. All authors participated in discussions and writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-41517-3.

Correspondence and requests for materials should be addressed to M.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.