Intragene Higher Order Repeats in Neuroblastoma BreakPoint Family Genes Distinguish Humans from Chimpanzees

Vladimir Paar,†1 Matko Glunčić,†1 Marija Rosandić,2 Ivan Basar,1 and Ines Vlahović1

1Faculty of Science, University of Zagreb, Zagreb, Croatia
2Department of Internal Medicine, University Hospital Rebro, University of Zagreb, Zagreb, Croatia

†These authors contributed equally to this work.
*Corresponding author: E-mail: paar@hazu.hr.
Associate editor: James McInerney

Abstract

Much attention has been devoted to identifying genomic patterns underlying the evolution of the human brain and its emergent advanced cognitive capabilities, which lie at the heart of differences distinguishing humans from chimpanzees, our closest living relatives. Here, we identify two particular intragene repeat structures of noncoding human DNA, spanning as much as a hundred kilobases, that are present in human genome but are absent from the chimpanzee genome and other nonhuman primates. Using our novel computational method Global Repeat Map, we examine tandem repeat structure in human and chimpanzee chromosome 1. In human chromosome 1, we find three higher order repeats (HORs), two of them novel, not reported previously, whereas in chimpanzee chromosome 1, we find only one HOR, a 2mer alpoid HOR instead of human alphoid 11mer HOR. In human chromosome 1, we identify an HOR based on 39-bp primary repeat unit, with secondary, tertiary, and quartic repeat units, fully embedded in human hornerin gene, related to regenerating and psoriatic skin. Such an HOR is not found in chimpanzee chromosome 1. We find a remarkable human 3mer HOR organization based on the 1.6-kb primary repeat unit, fully embedded within the neuroblastoma breakpoint family genes, which is related to the function of the human brain. Such HORs are not present in chimpanzees. In general, we find that human–chimpanzee differences are much larger for tandem repeats, in particularly for HORs, than for gene sequences. This may be of great significance in light of recent studies that are beginning to reveal the large-scale regulatory architecture of the human genome, in particular the role of noncoding sequences. We hypothesize about the possible importance of human accelerated HOR patterns as components in the gene expression multilayered regulatory network.

Key words: human brain evolution, chromosome 1, higher order repeats, NBPF genes, human hornerin gene, global repeat map.

Introduction

Human–Chimpanzee Divergence and Regulatory Elements

Identification of genomic differences that set humans apart from chimpanzees is important from evolutionary, medical, and cultural perspectives (Dorus et al. 2004; Enard and Paabo 2004; Hayakawa et al. 2005; Haygood et al. 2007, 2010; Kleinjan and Lettice 2008; Portin 2008; Varki et al. 2008; Woolfe and Elgar 2008; Enard et al. 2009; Taft et al. 2010). The most striking trend in human evolution is the rapid increase in brain size over the past 3–4 My and the associated increase in cognitive capacity and complexity (Jerison 1975; Melkon-Bobrov et al. 2007). It could be expected that the genetic changes, potentially underlying human brain evolution, span a wide range from nucleotide substitutions to large-scale structural alteration of the genome (Vallender 2008; Vallender et al. 2008). Chimpanzees share nearly 99% of human DNA in genes, but given the substantial number of neutral mutations, only a small subset of the observed gene differences is likely to be responsible for the key phenotypic changes between humans and chimpanzees (The chimpanzee sequencing and analysis consortium 2005). Developmental and evolutionary mechanisms underlying the characteristics that set the human brain apart from other human primates are poorly understood (Pollard et al. 2006a).

Recently, computational search was performed for DNA sequences that have changed most since humans and chimpanzees diverged from their last common ancestor (Amadio and Walsh 2006; Pollard et al. 2006b; Prabhakar et al. 2006; Bird et al. 2007; Benjamino et al. 2009). Among such human accelerated regions (HARs) the most pronounced is a 118-bp stretch on human chromosome 20, known as HAR1. It is highly conserved among species during most of vertebrate evolution, while changing significantly after the chimpanzee–human split: The human and chimpanzee HAR1 sequences differ by 18 bp. This rapid burst of base pair substitutions in HAR1 was interpreted as evidence that HAR1 acquired an important new function in humans and in particular may have altered human brain by influencing the activity of a whole network of genes (Pollard et al. 2006a). On the other hand, many highly conserved noncoding sequences have been found (Dermitzakis

Half a century ago, Jacob and Monod (1961) discovered that specific noncoding sequences are required to activate genes in *Escherichia coli* and noted that mutations in these regulatory sequences might play a role in the evolution of organismal traits. Britten and Davidson (1969, 1971) proposed a model that discrete regulatory sequences are abundant in the genomes of higher eukaryotes and that increases in the number of regulatory elements and the complexity of gene regulatory networks drive increased biological complexity in evolution. King and Wilson (1975) proposed that phenotypic differences between humans and chimpanzees are mainly caused by regulatory changes in gene expression. Raff and Kaufman (1983) hypothesized that mutations affecting the function of regulatory genes are likely to underlie many evolutionary changes in morphology: changes in just a handful of regulatory interactions could cause large-scale morphological changes. Recent investigations are progressing along such general framework.

Components of regulatory control in the human genome include proximal regulatory elements and long-range elements that can act across large genomic distances to influence the spatial and temporal distribution of gene expression, with enhancers and repressors of gene activity often located in evolutionary conserved noncoding regions of the genome (Pennacchio and Rubin 2001; Caceres et al. 2003; Cooper and Sidow 2003; Gu JY and Gu X 2003; Nobrega et al. 2003; Wray et al. 2003; Khaitovich et al. 2004; Uddin et al. 2004; Heissig et al. 2005; Kleinjan and van Heyningen 2005; Evans et al. 2006; Maston et al. 2006; Pennacchio et al. 2006; King et al. 2007; Xie et al. 2007; Portin 2008; Visel et al. 2008; Wray and Babbitt 2008; Duret and Galtier 2009; Mercer et al. 2009; Garfield and Wray 2010; Haygood et al. 2010; Noonan and McCallon 2010; Xu et al. 2010). Recent analyses provide evidence of many thousands of distant-acting noncoding sequences, beyond the proximal regulatory sequences and thus are beginning to reveal the large-scale regulatory architecture of the genome. At the present time, however, identifying regulatory elements in the human genome still remains a major challenge (Wooife and Elgar 2008; Noonan and McCallon 2010).

Genomewide studies have shown that human genome in transcription produces many thousands of regulatory noncoding RNAs (Huttenhofer et al. 2005; Johnson et al. 2005; Costa 2007; Kapranov et al. 2007; Amaral et al. 2008; Mattick 2009; Mercer et al. 2009; Van Bakel and Hughes 2009; Taft et al. 2010; Xu et al. 2010). Transcriptome studies indicated that a large proportion of transcription takes place outside of the known gene boundaries (Mercer et al. 2009; Van Bakel and Hughes 2009). Recent results of Xu et al. (2010) showed that close to 40% of all transcripts expressed in the human brain map are within repetitive elements, whereas <10% of the human brain transcriptome corresponds to nonrepetitive intergenic regions.

Morphological and anatomical differences between humans and chimpanzees are mainly caused by differences in the regulation of function of genes arising from noncoding sequences (King and Wilson 1975; Bush and Lahn 2008), and the chromosomal regions in man, that are gene-poor, harbor gene regulatory elements that have the ability to modulate gene expression over very long distances (Lettice et al. 2002).

A possible functional role of noncoding sequences and in particular of repeats has been much discussed. It is of importance to investigate the role of repeat sequences in the regulatory mechanism, related to a general concept that the regulatory system of genomes is encoded in networks of repetitive sequence relationships (Britten and Kohne 1968; Britten and Davidson 1969, 1971; Davidson and Britten 1979). In this sense, the repetitive components may play a major architectonic role in higher order physical structuring. It was argued that in this way one could think about genomes as information storage systems with parallels to electronic information storage systems. From such a perspective, repetitive DNA sequences are important components of genomes, required for formatting coding information so that it can be accurately expressed and for formatting DNA molecules for transmission to new generations of cells. The cooperative nature of protein–DNA interactions provides another fundamental reason why repeated sequence elements are essential to format genomic DNA (Shapiro and von Sternberg 2005). This was accompanied by observation that tandem arrays are often the regions that vary most between related taxi (Shapiro and von Sternberg 2005). A probable importance of noncoding changes in the evolution of human traits was underscored, particularly for cognitive traits (Haygood et al. 2010). Recent studies are beginning to reveal the large-scale regulatory architecture of the human genome (Noonan and McCallion 2010).

**Neuroblastoma BreakPoint Family Genes**

The recently described gene family neuroblastoma break-point family (*NBPF*) with an intricate genomic organization was expanded during recent primate evolution. There are indications that they are brain-related genes: At least one of them might suppress the development of neuroblastoma and possibly of other tumor types (Vandepoele et al. 2005, 2008, 2009; Gregory et al. 2006; Popesco et al. 2006; Vandepoele and van Roy 2007; Diskin et al. 2009).

Previous studies of genomic sequences of human *NBPF* genes using standard bioinformatics tools (Altschul et al. 1990; Benson 1999; Gelfand et al. 2007) have reported tandem repeats in these genes, so-called *NBPF* repeats, based on diverging ~1.5–1.6-kb repeat unit (Vandepoele et al. 2005; Gregory et al. 2006; Gelfand et al. 2007; Warburton et al. 2008). Several reports have shown that the copy number of these genes is variable in humans (Tuzun et al. 2005; Redon et al. 2006). A remarkable reduction of *NBPF* copy number with respect to the human genome was found in other primates, and in mouse, the *NBPF* copies are absent (Vandepoele et al. 2005, 2009; Popesco et al. 2006).

**Higher Order Repeats**

Higher order repeats (HORs) have been extensively studied in human and nonhuman primate chromosomes (Warburton and Willard 1996). The best known prototypes...
of HORs are alpha satellite (alphoid) arrays, located in the centromeric region of all human chromosomes (Manuelidis 1978; Wu and Manuelidis 1980; Willard 1985; Tyler-Smith and Brown 1987; Willard and WAYE 1987a, 1987b; Warburton and Willard 1996; Jurka 2000; Alexandrov et al. 2001; Rosandić et al. 2003, 2006; Rudd and Willard 2004; Jurka et al. 2005; Paar et al. 2005, 2007; Warburton et al. 2008). Alpha satellite arrays consist of primary repeat units (alphoid monomers) of ~171 bp, tandemly arranged in a head-to-tail fashion. Individual alphoid monomers diverge by 20–40% from each other. However, some stretches of alpha satellites are hierarchically organized into HORs, secondary repeat units with highly convergent HOR copies (divergence between HOR copies <5%) (Warburton and Willard 1996). In that case, divergence between HOR copies is much smaller than divergence between monomers within each HOR copy. Figure 1 schematically describes the overall concept of HORs for an illustrative case of 11mer HOR. Most of the reported HORs are based on relatively short primary repeat units, like 171-bp alpha satellite or less. The longest primary repeat unit was reported by Warburton et al. (2008) as a ~3.3-kb repeat forming 18-kb HORs.

Highly homogeneous arrays of HOR alpha satellite monomers are relatively recent additions to human genome (Warburton and Willard 1996; Alexandrov et al. 2001). An explanation for generating HORs involves unequal crossing-over between misaligned HOR units aligned on the register of homologous monomers. Unequal crossing-over, restricted to tandem sequences, explains the generation and local homogenization of HOR units and accounts for large size variation among HORs on homologous chromosomes (Southern 1975; Smith 1976; Willard and WAYE 1987b; Warburton and Willard 1996; Alkan et al. 2004; Rudd et al. 2006). By the process of unequal crossing-over, HORs enable rapid evolutionary development.

Global Repeat Map Analysis of Human–Chimpanzee Divergence

Here, we use our novel robust bioinformatics tool, the Global Repeat Map (GRM [see Methods]), for identification of tandem repeats and HORs in human and chimpanzee chromosome 1. In previous investigations, short stretches of human accelerated sequences (up to several hundreds of base pairs; HARs) have been identified, mostly corresponding to nonhuman multispecies conserved regions. Here, we search for HARs containing large tandem repeat and HOR units. We are asking whether a more complex genomic pattern, involving long-range correlations, is accelerated in brain-related human genes that could characterize the human brain evolution. As a case study, we investigate computationally the genomic sequence of chromosome 1 (National Center for Biotechnology Information [NCBI] Build 36.3 assembly) that contains the NBPF gene family. Here, we discover HORs within NBPF repeats in human chromosome 1, whereas we find no HOR in arrays of NBPF repeats in chimpanzee chromosome 1. This is the first time that HORs are found fully embedded within a gene. Furthermore, these HORs are based on an extraordinary large primary repeat unit. Besides the novel NBPF HOR pattern, highly accelerated in human chromosome 1 (in chimpanzee chromosome 1 this HOR pattern is missing), we find two additional human HORs (one of them novel) and several tandem repeats with large repeat units, all showing significant human–chimpanzee divergence, that is, the human accelerated pattern. We hypothesize on a possible role of this acceleration as possibly important components in the gene expression multilayered regulatory network.

Materials and Methods

Key String Algorithm

In spite of powerful standard computational tools in bioinformatics, there are still difficulties to identify and analyze long repeat units. For example, the Tandem Repeat Finder can identify tandem repeat units up to 2 kb (Benson 1999; Warburton et al. 2008). Here, we use a new robust approach, useful in particular for investigating very long and/or complex repeats. The Key String Algorithm (KSA) framework (Rosandić et al. 2003, 2006; Paar et al. 2005, 2007) is based on the use of a short sequence of nucleotides, referred to as key string, which cuts a given genomic sequence at each location where the key string appears within a given genomic sequence. The ensuing KSA fragments form the KSA length array that could be compared with an array of lengths of restriction fragments resulting from hypothetical complete digestion cutting genomic sequence at recognition sites corresponding to the KSA key string.

Global Repeat Map

The GRM algorithm is an extension of KSA framework and consists of the following steps: 1) GRM-Total module: computes the frequency versus fragment length distribution for a given genomic sequence by superposing results of consecutive KSA segmentations computed for an ensemble of all 8-bp key strings ($4^8 = 65,536$ key strings) (Paar et al. 2007). In the GRM diagram, each pronounced peak corresponds to one or more repeats at that length, tandem or dispersed. GRM computation is fast and can be easily
executed for a human chromosome using PC. 2) GRM-Dom module: determines the dominant key string corresponding to the fragment length for each peak in the GRM diagram from step 1. An 8-bp key string (or a group of 8-bp key strings) that gives the largest frequency for a fragment length under consideration is referred to as a dominant key string. 3) GRM-Seg module: performs segmentation of a given genomic sequence into KSA fragments using dominant key string from the step 2. Any periodic segment within the KSA length array reveals the location of repeats and provides genomic sequences of the corresponding repeat copies. 4) GRM-Cons module: aligns all sequences of repeat copies from step 3 and constructs consensus sequence. 5) NW module: computes divergence between each repeat copy from step 3 and consensus sequence from step 4 using the Needleman–Wunsch (Needleman and Wunsch 1970) algorithm.

Diagram outlining the main steps in the GRM process is shown in figure 2. Code for GRM modules is available upon the request to the authors.

Regarding the 8-bp choice of the key string size, using an ensemble of all 8-bp key strings, the average length of KSA fragments is \( \sim 4^r \). With increasing length of key strings, the overall frequency of large fragment lengths increases. For an ensemble of all 8-bp key strings, from computed GRM diagrams, we can identify the primary and secondary repeat units as large as 100 kb.

Characteristics of GRM are 1) robustness with respect to deviations from perfect repeats, that is, substitutions, insertions, and deletions; 2) straightforward and parameter-free identification of simple repeats (tandem and dispersed), applicable to very large repeat units (as large as several kilobases); 3) straightforward identification of HORs (secondary repeats) for very large primary and/or secondary repeat units; 4) straightforward determination of consensus lengths and consensus sequences for simple repeats and HORs; and 5) modest scope of computation using a PC.

The GRM method is a straightforward method to provide a GRM in a single diagram identifying all pronounced repeats in a given sequence, without any prior knowledge of the sequences structure. Once the size of the repeat is determined, GRM provides in a straightforward way the location of the corresponding repeat arrays and their precise analysis. GRM is particularly useful for precise sequence analysis because the method does not involve any averaging procedure. It is also useful that the method is rather robust with respect to sizeable substitutions and indels. Once the consensus repeat unit is determined using

![Diagram](image-url)
**Fig. 3** GRM diagrams for human and chimpanzee chromosome 1 assemblies, Build 36.3 and Build 2.1, respectively. Horizontal axis: fragment length (length of repeat unit); vertical axis: frequency of appearance. Upper panel: human GRM diagram; lower panel: chimpanzee GRM diagram. (A–D) GRM for the whole chromosome (except some smaller segments in inserts). (A) Insert of enlarged segments of human and chimpanzee GRM diagrams (fragment lengths 160–180 bp) for contigs NT_077389.3 and NW_001230099.1, respectively (peak 171 bp—upper
GRM, in the next step it could be well combined with basic local alignment search tool (Blast) search for dispersed units or their fragments. For very large repeat units, Tandem Repeat Finder has limitations, whereas GRM has no such size limitations.

Results and Discussion

GRM Diagrams and Repeat Patterns for Human and Chimpanzee Chromosomes 1

We compute the GRM diagrams for human chromosome 1 (Build 36.3) and chimpanzee chromosome 1 (Build 2.1) (fig. 3). The GRM analysis of pronounced peaks leads to the identification of pronounced repeats with large repeat units (table 1, supplementary tables S1 and S2, Supplementary Material online).

Of particular interest are three human HORs that distinguish humans from chimpanzees: 1) human 1,410-bp 36mer HOR based on 39-bp primary repeat unit, without a chimpanzee counterpart; 2) human 1,866-bp 11mer alphoid HOR based on ~171-bp alpha satellite primary repeat unit, with a chimpanzee counterpart 340-bp 2mer alphoid HOR; and 3) human ~4,770-bp 3mer HOR based on ~1.6-kb primary repeat unit, without the chimpanzee counterpart.

Furthermore, using GRM, we find eight tandem repeat units and some peaks that correspond to dispersed repeats (Alu, long interspersed nuclear elements [LINE]). In table 1, we compare human and chimpanzee GRM results and previous identification of repeats in human chromosome 1 (Warburton et al. 2008). We find a number of GRM peaks corresponding to novel tandem repeat units. There is a significant difference in the repeat pattern of human and chimpanzee chromosomes 1 (fig. 3). The lengths of sequenced chromosome 1 assemblies of human Build 36.3 and chimpanzee Build 2.1 are similar, which is reflected in a similar level of noise in the upper and lower panels of figure 3.

Human-Specific, 1,410-bp Quartic HOR Repeat Embedded in Human Hornerin Gene

In the GRM diagram for human chromosome 1 (fig. 3), we identify five copies of the 1,410-bp HOR repeat unit (start position in chromosome 1: 150452728, start position in contig NT_00487.18: 2676458) and determine the consensus HOR unit (supplementary table S3A, Supplementary Material online). The average divergence of HOR copies with respect to consensus is ~4%. (These five HOR copies appear as two subsets: average divergence among the first three copies is ~1%, whereas for pairs involving one or two of the remaining two copies, the divergence is sizably larger, 7~12%).

The GRM diagram of 1,410-bp HOR consensus sequence (fig. 3f) reveals its internal repeat structure. First, there is a series of equidistant peaks at multiples of 39 bp ($n \times 39$ bp, $n = 1, 2, 3, \ldots$). This shows that the 39-bp monomer is a primary repeat unit for the 1,410-bp HOR unit (pronounced peak in fig. 3f). The consensus sequence of the 39-bp primary repeat unit is displayed in supplementary table S3A (Supplementary Material online). Among the peaks at multiples of 39 bp in figure 3f, the strongest peak is at ~0.7 kb, and the second strongest is at ~0.35 kb.

This reveals three levels of HOR organization: nine 39-bp primary repeat units are organized into ~0.35-kb secondary repeat unit, two ~0.35-kb secondary repeat units are organized into ~0.7-kb tertiary HOR repeat unit, and finally, two ~0.7-kb tertiary repeat units are organized into the ~1.4-kb HOR quartic repeat units. Thus, the 1,410-bp repeat unit represents the quartic HOR repeat unit (fig. 4).

By using the consensus sequence, we find that the average divergence between neighboring copies is gradually decreasing with increasing level of HOR organization, from ~32% between 39-bp copies (first level—primary repeat unit) to ~4% between 1,410-bp HOR copies (fourth level—quartic repeat unit) (see fig. 4). Such hierarchy of divergences is a signature of HOR organization.

Tandem repeats are of biological interest because they represent a rapidly evolving type of DNA sequences that can show profound differences even between closely related species, as are humans and chimpanzees, and may contribute to phenotypic differences. In the present case of 36mer HOR, the whole HOR array is embedded within the human hornerin (HRNR) gene that produces hornerin protein, expressed in regenerating and psoriatic skin (Takaishi et al. 2005). Comparing with positions within genomic sequence in the NCBI database, this HOR array is positioned within one lengthy exon in HRNR gene (fig. 5).

Using the structure of human hornerin protein, Takaishi et al. (2005) have deduced the corresponding amino acid sequence. The repetitive region was divided into segments. The smallest units of ~39-bp amino acids showed a moderate homology to each other. However, Takaishi et al. (2005) concluded that the analysis was compatible with the notion that unit of ~39 bp was at first amplified 4-fold and then triplicated to form three segments in tandem, these being further amplified 6-fold, implying a [(39 bp) × 4] × 3] × 6 = 2.8-kb organization, which differs from the HOR organization [[(39 bp) × 9] × 2] × 2 = 1.4 kb found here applying GRM to the Build 36.3 genomic assembly. In the corresponding GRM diagram (fig. 3f), we clearly see
Table 1 Large Repeat Units in Human and Chimpanzee Chromosome 1: GRM Computation and Previous Identification for Humans and GRM Computation for Chimpanzees.

<table>
<thead>
<tr>
<th>GRM Human</th>
<th>Previous Identification Human</th>
<th>GRM Chimpanzee</th>
</tr>
</thead>
<tbody>
<tr>
<td>135 bp, Signature of Alu</td>
<td>135 bp, Signature of Alu</td>
<td>135 bp, Signature of Alu</td>
</tr>
<tr>
<td>166 bp, Signature of Alu</td>
<td>166 bp, Signature of Alu</td>
<td>166 bp, Signature of Alu</td>
</tr>
<tr>
<td>~171 bp, PRU—171 bp (alpha satellite, div. ~20%)</td>
<td>~171 bp</td>
<td>171 bp, PRU—alpha satellite</td>
</tr>
<tr>
<td>~310 bp, Signature of Alus in tandem</td>
<td>342 bp, 2 × 171 PRU</td>
<td>342 bp, 2 × 171 PRU</td>
</tr>
<tr>
<td>342 bp, 2 × 171 PRU</td>
<td></td>
<td>342 bp, 2 × 171 PRU</td>
</tr>
<tr>
<td>310 bp, Signature of Alus in tandem</td>
<td>310 bp, Signature of Alus in tandem</td>
<td>310 bp, Signature of Alus in tandem</td>
</tr>
<tr>
<td>310 bp, Signature of Alus in tandem</td>
<td>310 bp, Signature of Alus in tandem</td>
<td>310 bp, Signature of Alus in tandem</td>
</tr>
<tr>
<td>169 bp, PRU—alpha satellite</td>
<td>513 bp, 3 × 171 PRU</td>
<td>513 bp, 3 × 171 PRU</td>
</tr>
<tr>
<td>171 bp, PRU—alpha satellite</td>
<td>684 bp, 4 × 171 PRU</td>
<td>684 bp, 4 × 171 PRU</td>
</tr>
<tr>
<td>171 bp (alpha satellite, div. 2%)</td>
<td>972 bp, PRU = 972 bp (TA, 11 copies)</td>
<td>972 bp, PRU = 972 bp (TA, 11 copies)</td>
</tr>
<tr>
<td>342 bp, 2 × 171 PRU</td>
<td>1,410 bp, HOR 36mer (TA, 5.5 copies, div. <del>3%) (QRU</del>1,410 bp, TRU<del>0.7 kb, SRU</del>0.35 kb, PRU39 bp)</td>
<td>1,410 bp, HOR 36mer (TA, 5.5 copies, div. <del>3%) (QRU</del>1,410 bp, TRU<del>0.7 kb, SRU</del>0.35 kb, PRU39 bp)</td>
</tr>
<tr>
<td>1,410 bp, HOR 36mer (TA, 5.5 copies, div. <del>3%) (QRU</del>1,410 bp, TRU<del>0.7 kb, SRU</del>0.35 kb, PRU39 bp)</td>
<td>~1.5 kb (84 copies in 3 arrays, div. &lt;5%)</td>
<td>~1.5 kb (84 copies in 3 arrays, div. &lt;5%)</td>
</tr>
<tr>
<td>1,866 bp, HOR 11mer alphoid (SRU = 1,866 bp, PRU = 171 bp, div. ~4%)</td>
<td>1,866 bp, HOR 11mer alphoid (SRU = 1,866 bp, PRU = 171 bp)</td>
<td>1,866 bp, HOR 11mer alphoid (SRU = 1,866 bp, PRU = 171 bp)</td>
</tr>
<tr>
<td>2,241 bp, PRU = 2,241 bp (TA, 17 copies, div. ~0.4%)</td>
<td>2.5 kb (TA, 16 copies, div. &lt;1%)</td>
<td>2.5 kb (TA, 16 copies, div. &lt;1%)</td>
</tr>
<tr>
<td>3,183–3,193 bp, 2 × 1.6-kb PRU</td>
<td></td>
<td>3,229 bp, PRU (TA, two copies, div. ~1%)</td>
</tr>
<tr>
<td>~3,416 bp, PRU (TA 3.25 copies, div. ~1%)</td>
<td>3.4 kb (TA, four copies, div. &lt;2%)</td>
<td>3,451 bp, PRU (TA, two copies, div. ~1%)</td>
</tr>
<tr>
<td>~3,732 bp, 2 × 1.866 bp</td>
<td></td>
<td>3,451 bp, PRU (TA, two copies, div. ~1%)</td>
</tr>
<tr>
<td>4,720–4,770 bp, HOR 3mer (SRU = 4,770 bp, PRU~1.6 kb)</td>
<td></td>
<td>4,972 bp, PRU (TA, two copies, div. ~1%)</td>
</tr>
<tr>
<td>6,277–6,305 bp (TA, ~13 copies in 6 arrays, div. ~5%)</td>
<td></td>
<td>5,408 bp, PRU (TA, two copies, div. ~2%)</td>
</tr>
<tr>
<td>7,380 bp, PRU = 7,380 (TA, 4.4 copies, div. ~0.5%)</td>
<td></td>
<td>6,461 bp, PRU (TA, two copies, div. ~2%)</td>
</tr>
<tr>
<td>10,828 bp, PRU (TA, two copies, div. ~5%)</td>
<td></td>
<td>9,734 bp, two 7,295-bp dispersed repeats (div. ~1%)</td>
</tr>
<tr>
<td>13,745 bp, two ~6.1-kb dispersed LINE1 copies (div. ~5%)</td>
<td>10.1 (TA, three copies, div.&lt;9.7%)</td>
<td>11,602 bp, two ~7-kb dispersed repeats (div. ~1%)</td>
</tr>
<tr>
<td>14,264 bp, two ~6-kb dispersed parts of L1HS (div. ~5%)</td>
<td></td>
<td>13,255 bp, two 3,259-bp dispersed repeats (div. ~1%)</td>
</tr>
<tr>
<td>14,618 bp, two ~4.7-kb dispersed parts of L1HS (div. ~5%)</td>
<td>18.6 kb (TA, three copies, div.&lt;3%)</td>
<td></td>
</tr>
<tr>
<td>17,095 bp, PRU 18,555 bp (TA, three copies, div. ~2%)k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18,555 bp, PRU 18,555 bp (TA, three copies, div. ~2%)k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18,555 bp, PRU 18,555 bp (TA, three copies, div. ~2%)k</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTE**—PRU, primary repeat unit; SRU, secondary repeat unit; TRU, tertiary repeat unit; QRU, quartic repeat unit; TA, tandem array.

k Previously reported results for large repeats in chromosome 1 are from tables 1 and 2 in Warburton et al. (2008).

Both the 18,555- and 17,095-bp peaks are generated by a tandem of three ~18.6-bp copies. In the third copy, the first ~1.5-bp bases are deleted, and therefore, the third copy is shortened to 7,095 bp. Thus, the first and second copies generate the peak at 18,555 bp, whereas the second and third are at 17,095 bp.
GRM peaks at 39, ~0.35, ~0.7, and ~1.4-kb (table 2), in accordance with our annotation of 1.4-kb quartic HOR, whereas there is no peak at 2.8 kb predicted by Takaishi et al. (2005).

By using Tandem Repeat Finder, a tandem repeat unit of 1.4 kb was previously detected (Warburton et al. 2008), but the HOR structure was not found. Because the 1.4-kb repeat is within a coding region (fig. 5), it is of evolutionary interest to compare the resultant nucleotide structure of this highly repetitive gene, that is, how variation between repeat copies affects the protein structure. Using the EMBOSS Transeq code (Williams 2002) for translating nucleotide into protein sequence, we find that the variation within the genomic sequence mostly causes nonsynonymous changes, predicting a variable protein domain structure (see supplementary table S3B, Supplementary Material online).

In chimpanzee GRM diagram, we find no counterparts of 1.4-kb human HOR.

Human 11mer and Chimpanzee 2mer Alphoid HORs

For 11mer alphoid HOR in chromosome 1 (Waye et al. 1987), the consensus length obtained using the GRM algorithm is 1,866 bp (fig. 3), in accordance with the precise value from previous bioinformatics studies (Paar et al. 2005; Rosandić et al. 2006; Warburton et al. 2008). The array of alpha satellite monomers constituting HOR gives rise to GRM peaks at multiples of alpha satellite monomer length 171:171 bp, 2 × 171 = 342 bp, 3 × 171 = 513 bp, . . . with decreasing frequency (fig. 3). The next higher alphoid GRM fragment length corresponds to a tandem of two 11mer HOR copies, that is, 2 × 1,866 bp = 3,732 bp (table 1).

For the chimpanzee chromosome 1, we find a different alphoid HOR pattern: Instead of the 11mer HOR, we obtain the 2mer HOR. This 2mer HOR is based on two alpha satellite monomers of the lengths 169 and 171 bp (mutual divergence 25%). Thus, the length of the 2mer HOR unit is 340 bp; this is the length of a peak in the chimpanzee GRM diagram. Accordingly, the next pronounced alphoid GRM fragment length corresponds to a tandem of two 2mer HOR copies (mutual divergence <1%), that is, 2 × 340 = 680 bp, whereas there is no pronounced peak at ~ 3 × 171 = 513 bp. We compare the alphoid HOR units of human chromosome 1 (11mer) and chimpanzee chromosome 1 (2mer) (supplementary table S4, Supplementary Material online); the average divergence between the 169- and 171-bp chimpanzee monomers and m01–m11 human monomers is 27%.

The present result for 2mer HOR in chimpanzee is in accordance with previous observations that some early primate alpha satellites have a dimeric organization (Alexandrov et al. 2001; Schueler and Sullivan 2006). It is not surprising that the alphoid HORs from human and chimpanzee chromosomes 1 are different, because different alphoid HORs can
be found on homologous chromosomes on these species, representing a rapid expansion since speciation.

**Human-Specific 4,770-bp 3mer HOR in NBPF Genes in NT_113799.1**

Performing GRM computation for Build 36.3 genomic assembly of human chromosome 1, we obtain an interesting new result: The two GRM peaks, at 4,770 and 3,193 bp, correspond to HORs based on the ~1.6-kb primary repeat unit in the NBPF genes. The largest contribution to the 4,770-bp peak arises from the contig NT_113799.1 in chromosome 1. The GRM diagram for the 4,770-bp consensus sequence shows two peaks, at ~1.6 and ~3.2 kb (fig. 6A). In a further step, the GRM diagram computed for the 3.2-kb peak shows only one pronounced peak, at ~1.6-kb (fig. 6B). This shows that the ~1.6-kb peak corresponds to the primary repeat unit for the 4,770- and 3,193-bp peaks. Using GRM, we determine a dominant key string CACTGACC for segmentation of the contig NT_113799.1 into a maximum number of the ~1.6-kb fragments. Using this key string, we obtain an array of KSA fragment lengths (supplementary table S5, Supplementary Material online). Aligning the resulting length arrays, we see that the three ~1.6-kb fragments, representing monomers, form a tandemly repeating 3mer: It consists of three ~1.6-kb NBPF monomers, belonging to three monomer families, with consensus lengths 1,623, 1,593, and 1,554 bp, respectively. Consensus length for each family is determined as the most frequent length for primary repeat monomers belonging to that family. Explicit relations between consensus sequence and copies are illustrated in Supplementary figure S13, Supplementary Material online. The corresponding basic consensus monomers are denoted as m01, m02, and m03, respectively (see consensus sequences in supplementary table S6, Supplementary Material online).

Finally, the computed GRM diagram for each of three ~1.6-kb basic monomers is without any pronounced peak, reflecting its monomer character, that is, the absence of any internal repeat structure. The broadened peak consisting of a cluster of scattered ~1.6-kb lengths in the human GRM diagram (fig. 38) is due to cases with the presence of the same key string in both m01 and m02 and/or in both m02 and m03 within the 3mer m01m02m03. A peak at ~3.2 kb appears in the case when the same key string is present in m01 and m03 but not in m02. Table 3 shows the detailed NBPF monomer structure of the 3mer HOR. It is seen that the m03 monomer is deleted from 3mer HOR copies no. 6–8. This monomer deletion, resulting in the appearance of NBPF dimers within the NBPF array, contributes to the ~3.2-kb peak in the GRM diagram.

Divergence between the three consensus monomers is 19% (m01 vs. m02), 15% (m01 vs. m03) and 20% (m02 vs. m03). On the other hand, the average divergence between the 3mer HOR copies is mostly <0.5%, that is, the 3mer copies are highly identical. This substantially smaller divergence between HOR copies than between constituting monomers is a signature of well-developed HOR pattern (see Introduction). Thus, the 4,770-bp repeat unit corresponds to strongly convergent HORs based on three diverging monomers m01, m02, and m03. This HOR array consists of a tandem of 17 highly identical HOR copies (in the interval in chromosome 1: 146618386–146694662). (We find no NBPF monomers outside the HOR array.) This HOR is fully embedded within the NBPF20 gene (NM_001037675.2). (Note that the Build 36.3 sequence is reverse complement to the sequence of NBPF20 gene from NCBI database.) All 17 NBPF HOR copies (horizontal stretches) are displayed by aligning the constituent NBPF monomers and the corresponding regions of exons are shown below HOR copies (vertical bars; fig. 7).

An example of divergence between segments at position of exons in the 15th and 16th HOR copies and between the corresponding protein sequences is shown in supplementary table S9 (Supplementary Material online).

The 4,770-bp 3mer HOR was not reported previously. However, the self-similarity dot plot of the NBPF region can also detect the presence of this HOR.

**Table 2** Scheme of Quartic HOR Organization Based on 39-bp Primary Repeat Unit Embedded within Human Hornerin Gene in Human Chromosome 1.

<table>
<thead>
<tr>
<th>Repeat Unit</th>
<th>Structure</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>Monomer</td>
<td>39 bp</td>
</tr>
<tr>
<td>Secondary</td>
<td>9 × 39 bp</td>
<td>~0.35 kb</td>
</tr>
<tr>
<td>Tertiary</td>
<td>2 × 0.35 kb</td>
<td>~0.7 kb</td>
</tr>
<tr>
<td>Quartic</td>
<td>2 × 0.7 kb</td>
<td>~1.4 kb</td>
</tr>
</tbody>
</table>

![Fig. 6](image_url) **(A)** GRM diagram of the 4,770-bp 3mer HOR consensus sequence (m01m02m03). **(B)** GRM diagram of the 3.2-kb 2mer HOR consensus sequence (one monomer from 3mer HOR is missing).
Most of the \textit{NBPF} repeats in other contigs are organized in HORs; at the edges of \textit{NBPF} regions, increased divergence and/or truncation appears in HOR copies. HOR arrays or their segments are present in 14 distinct arrays in 10 contigs (supplementary table S8, Supplementary Material online). Only three \textit{NBPF} HOR tandem arrays are of significant size, located in contigs NT\_113799.1, NT\_079497.3, and NT\_004434.18: They contain 17, 16, and 12 \textit{NBPF} HOR copies, respectively (fig. 8). All three HOR arrays are highly similar: The average divergence between 17 HOR copies in NT\_113799.1 and 16 HOR copies in NT\_079497.3 is 0.6%, and between 17 HOR copies in NT\_113799.1 and 12 HOR copies in NT\_004434.18, the average divergence is 2%.

Absence of \textit{NBPF} HORs in Nonhuman Primates
To search for \textit{NBPF} HORs in nonhuman primate genomes, we used human \textit{NBPF} consensus monomers m01, m02, m03 (supplementary table S6, Supplementary Material online) for Blast (Altschul et al. 1990) search against the chimpanzee, orangutan, and rhesus macaque genomic sequences (supplementary tables S10–S12, Supplementary Material online). The arrays of \textit{NBPF} repeats in chimpanzees and other nonhuman primates are not organized into HORs. This is shown by the absence of pronounced GRM peaks at \~4,770 bp (fig. 3).

Human and Chimpanzee GRM Diagrams and Tandem Repeats without Higher Order Organization
For most of monomeric tandem repeats, there is a sizeable difference between human and chimpanzee chromosomes 1 (see table 1). For example, pronounced repeat units at 2,241, 7,380, and 18,555 bp in human chromosome 1 are absent in chimpanzee chromosome 1. On the other hand, a sizeable 9,734-bp GRM peak in chimpanzee chromosome 1 is absent in human chromosome 1.

The exception is, for example, a tandem based on the 972-bp repeat unit that is present in both humans and chimpanzees. The corresponding chimpanzee peak in an overall GRM diagram of the whole chromosome 1 is weak, almost immersed into noise (weak peak in the lower panel of fig. 3). To reduce the noise, we additionally compute the GRM diagram for the contig NW\_001229599.1 that contains tandem array based on the 972-bp repeat unit.
Table 3 NBPF Monomer Repeat Structure of 3mer HORs in Contig NT_113799.1 in Chromosome 1.

<table>
<thead>
<tr>
<th>Consensus Monomer</th>
<th>Ch1 Start Position</th>
<th>Contig Start Position</th>
<th>Monomer Length (bp)</th>
<th>Monomer Divergence (%)</th>
<th>HOR Length (bp)</th>
<th>HOR Copy No.</th>
<th>HOR Divergence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m01</td>
<td>146618386</td>
<td>75723</td>
<td>1,608</td>
<td>8.9</td>
<td>4,749</td>
<td>1</td>
<td>3.3</td>
</tr>
<tr>
<td>m02</td>
<td>146619994</td>
<td>77331</td>
<td>1,587</td>
<td>0.7</td>
<td>4,744</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>m03</td>
<td>146621581</td>
<td>78918</td>
<td>1,554</td>
<td>0.1</td>
<td>4,764</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>m01</td>
<td>146623135</td>
<td>80472</td>
<td>1,623</td>
<td>0.1</td>
<td>4,760</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>m02</td>
<td>146624758</td>
<td>82095</td>
<td>1,587</td>
<td>0.6</td>
<td>4,764</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>m03</td>
<td>146626345</td>
<td>83682</td>
<td>1,554</td>
<td>0.1</td>
<td>4,772</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>m01</td>
<td>146627899</td>
<td>85236</td>
<td>1,623</td>
<td>0.1</td>
<td>4,764</td>
<td>5</td>
<td>0.4</td>
</tr>
<tr>
<td>m02</td>
<td>146629522</td>
<td>86859</td>
<td>1,583</td>
<td>0.8</td>
<td>4,764</td>
<td>5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

*Consensus monomers m01 (1,623 bp), m02 (1,593 bp), and m03 (1,554 bp) (see supplementary table S2, Supplementary Material online).

b Length of monomer copy.

c Divergence of monomer copy with respect to consensus monomer.

d Length of HOR copy (consensus length 4,770 bp).

e Ordinal number of HOR copy in tandem array.

f Divergence of HOR copy with respect to consensus HOR.

g Monomer m03 in HOR copies No. 6–8 is missing.

h Monomer m01 in HOR copies no. 7–9 is substituted by hybrid m1–3 (1,600 bp). Divergence of m1–3 with respect to consensus monomers m01, m02, and m03 is 7%, 22%, and 9%, respectively. The m1–3 sequence exhibits a specific hybrid structure: the first 845 bases in m1–3 are identical to the first 841 bases from m03 additionally, a 4-bp insertion AGAG is inserted into m03 after position 581), whereas the last 821-bp segment in m1–3 is identical to the last 821-bp segment from m01. Divergence between m1–3 and m01 (7%) is smaller than divergence between m01 and m03 but larger than the mutual divergence between different copies of m01 or of m03 (<0.5%). Considering the three 1,600-bp copies as belonging to the m01 family, divergence of the corresponding three HOR copies with respect to consensus HOR is still <0.5%, whereas for the m1–3 classification of these three monomer copies, divergence between HOR copies is reduced to 0.5%.
In this case with a shorter sequence, the noise level at the position of the 972-bp peak is reduced (as seen from insert in the lower panel in fig. 3B).

For short repeat units, 20 bp, the corresponding human and chimpanzee GRM peaks are similar. In the length interval from 20 to 100 bp, the most pronounced peak in the human GRM diagram is at 39 and 40 bp, and in chimpanzee, it is at 32 bp. We find weaker peaks at multiples of 39 bp (78, 117 bp, ...), and 40 bp (80, 120 bp, ...), and 32 bp (64, 96 bp, ...), in human and chimpanzee GRM diagrams. The human 39-bp repeat unit is the primary repeat unit for the ~1,410-bp HOR, and 40 bp is a tandem repeat unit of another array.

**GRM Signature of Alu Sequences**

The pronounced GRM peaks at 135, 166, and ~310 bp in the GRM diagram (fig. 3) are signature of Alu sequences. Human Alu elements are dimeric structures of about 300 bp that comprise two similar but nonidentical monomers (Weiner et al. 1986; Britten et al. 1988; Jurka 1995, 2004; Schmid 1996). The longer monomer (on the right) contains an insert of 31 nt that is absent in the shorter monomer (on the left). The two monomers are connected by a short internal poly-A tract. The dimer sequence is followed by a long terminal poly-A tail of variable length. The 135-bp GRM fragment length corresponds to a distance between the starts of the left and right monomers. The 166-bp GRM fragment corresponds to the distance from the start of the segment of right monomer after the 31-bp insert and the start of a similar segment in the left monomer (thus, this distance is 135 + 31 bp = 166 bp). The ~310-bp length corresponds to a tandem of two Alu sequences; because of the variable length of the poly-A tail, this GRM peak is broadened in the interval of about a dozen base pairs, centered at ~310 bp.

**Statistical Significance of Peaks in the GRM Diagram**

An important question can be raised about the statistical significance of the peaks appearing in the GRM diagram, because they arise by applying an ensemble of a large number (4^45 = 65,536) of key strings. We use three strong tests of this significance.

First, we compute the corresponding GRM diagrams for artificial sequences, constructed using a random number generator, of the same nucleotide composition as in the genomic sequences under study. We compute the GRM diagram for a pseudochromosome 1, constructed in such a way (fig. 9). This randomly constructed sequence from a given number of nucleotides does not produce any significant peak above the level of noise in the GRM diagram.

The second test of statistical significance is provided by the steps of the GRM algorithm (see Methods): For each length at position of a peak, we determine the dominant key string and determine the corresponding repeat array giving rise to this peak. In this way, we find explicitly that to each GRM peak corresponds a concrete repeat pattern.

The third test of statistical significance is the identification of previously known HORs in human chromosomes (e.g., alphoid 11mer HOR in chromosome 1, alphoid 18mer HOR in chromosome 7, and so on). In this way, we identify GRM peaks for all previously known human alphoid HORs. Additionally, for the Build 36.3 assembly of human chromosome 1, we find peaks corresponding to all repeats previously identified in human chromosome 1 using standard bioinformatics tools (Warburton et al. 2008) and some novel peaks reported for the first time in this work.

**Human Accelerated NBPF and HRNR HORs—Components of Genomic Regulatory Network?**

Previously, it was found that the number of NBPF monomer repeats is related to the evolutionary level of higher primates, showing a gradual increase with evolutionary development (Vandepoele et al. 2005, 2009; Popesco et al. 2006), but the NBPF HOR pattern was not identified. Our novel GRM results for human chromosome 1 show that the ~1.6-kb NBPF monomers, mutually diverging 20–25%,
are in fact organized into highly identical 3mer HORs (4,770-bp consensus). With respect to previously known HORs in the human genome, this HOR pattern is interesting in several aspects: 1) The repeat unit is much longer than most of the primary repeat units identified so far. 2) Most of the NBPF repeat monomers are organized into 3mer HORs. 3) There are three distinct major NBPF HOR copies organized in tandem repeats (two or more HOR copies in tandem array). 4) NBPF HOR copies are in chromosome 1, whereas eight monomers and two distinct HOR copies are in contigs not assigned to the chromosome.

Another case of HOR being fully embedded within a gene is a 1.4-kb quartic HOR based on 39-bp primary repeat unit. This HOR is fully embedded within a single large exon in the human hornerin gene (HRNR), related to human skin. It is of substantial interest evolutionary to compare the resultant amino acid structure of this highly repetitive gene.

We show that the presence of NBPF HOR and HRNR HOR tandem repeats is an exclusively human specific. We hypothesize that this may be related to the possible tandem repeat role of the HOR pattern. In both of these within-the-gene HORs, we find no HOR counterparts in the chimpanzee genome. These HORs represent a human accelerated pattern. It seems of interest to look systematically for higher order repeats (HORs), particularly in or near brain-related and other genes and for their possible regulatory function.

More generally, it is an intriguing question related to a possible genomic regulatory network (GRN) with various components including known gene regulators and as yet hidden regulatory elements at different levels of coordination and hierarchical organization. In this sense, it is challenging whether some types of accelerated sequence may be functional, acting dynamically within a network of genes, gene regulators, and higher order GRN components.

Small differences among repeat copies, even when these copies exhibit a high degree of mutual similarity, might have a significant impact on the function of the multilayer regulatory network ("circuitry"). Their influence on gene expression may be direct, as in the case of some known HAR regulators, or indirect along a hierarchy of possible higher level organization. In this connection, it should be noted that growing evidence from the field of developmental genetics is supporting the hypothesis that small differences in the time of activation or in the level of activity of even a single gene could have important evolutionary consequences owing to the extensive consequences that changes in regulatory interactions could have on developmental processes (Garfield and Wray 2010). In general, one could argue that because any vanishingly small input applied at critical points in nonlinear dynamical systems might cause a finite response (Otto 1993; Hilborn 1994; Alligood et al. 1997; Zak et al. 1997), a system could be controlled by a microdynamical device that operates by sign strings, as a "genetic code."

Tandem repeats and HORs are the expected result of rapid expansion, given models of unequal crossing-over. They are of interest in the framework of a large-scale regulatory architecture of genome as a rapidly evolving type of DNA sequence that can show a profound difference between closely related species and may contribute to phenotypic differences. This may shed more light on human–chimpanzee evolutionary divergence from the point of view of long-range higher order periodicity.

**Supplementary Material**

Supplementary tables S1–S12 and Supplementary figure S13 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).
Acknowledgments

We thank two anonymous reviewers for valuable comments and suggestions, contributing significantly to the quality of our manuscript. We also thank Đ. Ugarković, J. Brajković, and K. Vlahoviček for useful discussion.

References


