Large Tandem, Higher Order Repeats and Regularly Dispersed Repeat Units Contribute Substantially to Divergence Between Human and Chimpanzee Y Chromosomes

V. Paar[†],[*] M. Glunčić[†], I. Basar, and M. Cvitković
Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia

M. Rosandić
Department of Internal Medicine, University Hospital Rebro, University of Zagreb, 10000 Zagreb, Croatia

P. Paar
Faculty of Electrical Engineering and Computing, 10000 Zagreb, Croatia

(Dated: December 21, 2010)

Comparison of human and chimpanzee genomes has received much attention, because of paramount role for understanding evolutionary step distinguishing us from our closest living relative. In order to contribute to insight into Y chromosome evolutionary history, we study and compare tandems, higher order repeats (HORs), and regularly dispersed repeats in human and chimpanzee Y chromosome contigs, using robust Global Repeat Map algorithm. We find a new type of long–range acceleration, human–accelerated HOR regions. In peripheral domains of 35mer human alphoid HORs, we find riddled features with ten additional repeat monomers. In chimpanzee, we identify 30mer alphoid HOR. We construct alphoid HOR schemes showing significant human–chimpanzee difference, revealing rapid evolution after human–chimpanzee separation. We identify and analyze over 20 large repeat units, most of them reported here for the first time as: chimpanzee and human ∼1.6 kb 3mer secondary repeat unit (SRU) and ∼23.5 kb tertiary repeat unit (∼0.55 kb primary repeat unit, PRU); human 10848, 15775, 20309, 60910, and 72140 bp PRUs; human 3mer SRU (∼2.4 kb PRU); 715mer and 1123mer SRUs (5mer PRU); chimpanzee 5096, 10762, 10853, 60523 bp PRUs; and chimpanzee 64624 bp SRU (10853 bp PRU). We show that substantial human–chimpanzee differences are concentrated in large repeat structures, at the level of as much as ∼70% divergence, sizably exceeding previous numerical estimates for some selected noncoding sequences. Smeared over the whole sequenced assembly (25 Mb) this gives ∼14% human-chimpanzee divergence. This is significantly higher estimate of divergence between human and chimpanzee than previous estimates.

Keywords: Human genome, Chimpanzee genome, Y chromosome, Male–specific region, Higher order repeat, Tandem repeat, Alpha satellite, Global Repeat Map, Evolution genetics, Long–range regulatory elements

I. INTRODUCTION

A. Atypical Structure of Human Y Chromosome

One of challenging problems in genomics is related to the evolutionary development of Y chromosome. The Y chromosome has a unique role in human population genetics with properties that distinguish it from all other chromosomes (Jobling and Tyler–Smith 2003; Mitchell et al. 1985; Skaletsky et al. 2003). Prevailing theory is that X and Y chromosomes evolved from a pair of autosomes (Graves 1995; Lahn and Page 1999; Marshall Graves 2006; Muller 1914; Ohno 1967). Lack of recombination between nonrecombining parts of X and Y chromosomes was thought to be responsible for decay of the Y–linked genes, the pace of which slows over time, eventually leading to a paucity of genes. Identification of distinct palindromes harboring several distinct gene families unique to the long arm of Y chromosome, frequent gene conversion, and multiplication have raised some doubt about progressive decay of the Y chromosome (Ali and Hasnian 2003; de Knijff 2006; Kuroda–Kawaguchi et al. 2001; Rozen et al. 2003; Skaletsky et al. 2003). It was shown that the Y chromosome has acquired a large number of testis specific genes during the course of evolution, including those essential for spermatogenesis (Saxena et al. 1996; Silber and Repping 2002; Skaletsky et al. 2003).

Considerations of atypical structure of human Y chromosome were largely focused on the gene–related content. On the other hand, however, the human Y chromosome is replete with many pronounced repetitive sequences,

[†]V. Paar and M. Glunčić contributed equally to this work.

[*]Electronic address: paar@hazu.hr

and multicopy gene arrays are embedded in palindromes (Cooper et al. 1993a; Kirsch et al. 2008; Oakey and Tyler–Smith 1990; Perry et al. 2007; Rozen et al. 2003; Skaletsky et al. 2003; Tyler–Smith 1985; Tyler–Smith and Brown 1987; Wolfe et al. 1985).

B. Alphoid Higher Order Repeats

Alphoid arrays in centromeres of human and other mammal chromosomes consist of tandem repeats of AT–rich alpha satellites (Alexandrov et al. 2001; Choo 1997; Maio 1971; Manuelidis and Wu 1978; Mitchell et al. 1985; Romanova et al. 1996; Rudd et al. 2006; Tyler–Smith 1985; Tyler–Smith and Brown 1987; Warburton and Willard 1996; Warburton et al. 1996; Waye and Willard 1987; Willard 1985). Stretches of alpha satellites lacking any higher–order periodicity mutually diverge by∼20-35% and are referred to as monomeric(Warburton and Willard 1996).

Higher order repeats (HORs) are defined as higher order periodicity pattern superimposed on the approximately periodic tandem of alpha monomers: if an array of n monomers denoted by $1, 2, \ldots, n$ is followed by the next array of monomers denoted by $n+1, n+2, \ldots, 2n$, where the monomer 1 is almost identical (more than 95%) to the monomer $n+1$, the monomer 2 to the monomer $n+2$, and the monomer n to the monomer $2n$, these arrays belong to the nmer HOR (Warburton and Willard 1996). The HOR copies from the same locus diverge from each other by $< 5\%$, while the alpha satellite copies within any HOR copy diverge from each other by $\sim 20 - 35\%$ (Warburton and Willard 1996).

Alphoid HORs are chromosome–specific (Choo 1997; Haaf and Willard 1992; Jorgensen et al. 1986; Warburton and Willard 1996; Willard 1985; Willard and Waye 1987). A type of polymorphism found in alphoid arrays involves HOR units that differ by an integral number of monomers (monomer insertion or deletion), but nonetheless closely related in sequence (Haaf and Willard 1992; Warburton and Willard 1996).

Investigations using restriction endonuclease digestion have revealed a major block of alphoid DNA in the centromeric region of human Y chromosome (Cooper et al. 1993a,b; Mitchell et al. 1985; Tyler–Smith 1985; Tyler–Smith and Brown 1987; Wolfe et al. 1985). The size of this alphoid block was found to be polymorphic, widely varying between different individuals (Oakey and Tyler–Smith 1990; Tyler–Smith and Brown 1987). Initially, a 5.7 kb HOR unit was reported as a major variant of secondary periodicity and 6.0 kb HOR unit as a minor variant. These HOR units were associated with 34mer and 36mer, respectively (Tyler–Smith and Brown 1987). In a more recent study, a 5941 bp secondary periodicity (35 alphoid repeat units) was reported (Skaletsky et al. 2003).

The alpha satellite DNA can be considered as a paradigm for processes of concerted evolution in tandemly repeated DNA families (Warburton and Willard 1996; Willard 1991; Willard and Waye 1987).

C. Bioinformatics Studies of Alphoid HORs

During the last decade sequence contigs spanning the junction at the edges of the centromere DNA array are available for bioinformatics analyses (Nusbaum et al. 2006; Paar et al. 2005, 2007; Rosandić et al. 2003a,b, 2006; Ross et al. 2005; Rudd and Willard 2004; Rudd et al. 2003; Skaletsky et al. 2003). However, major gaps still remain at the centromeric region of chromosomes (Henikoff 2002; Rudd and Willard 2004; Schueller et al. 2001). Mostly, only peripheral HOR copies are accessible, at the edges of centromeric region. Previously, Rudd and Willard (2004) analyzed the Build 34 assembly, using a combination of BLAST (Altschul et al. 1990) and DOTTER (Sonnhammer and Durbin 1995), and reported the presence of HORs. Recently, using Tandem Repeat Finder (TRF) (Benson 1999) and other standard bioinformatics tools, Gelfand et al. (2007) and Warburton et al. (2008) studied human HORs in more details.

In a different approach, we have shown that the Key String Algorithm (KSA) and an extension Global Repeat Map (GRM) are effective in identification and analysis of intrinsic structure of HORs (Paar et al. 2005, 2007; Rosandić et al. 2003a,b, 2006). Applying KSA and GRM to the NCBI human genome assembly, detailed structure of known and some new human alphoid HORs was determined.

D. Comparison of Human and Chimpanzee Genome Sequences

To understand the genetic basis of unique human features, the human and chimpanzee genomes have been compared in a number of studies (Bailey and Eichler 2006; Boffelli et al. 2003; Chen and Li 2001; Cheng et al. 2005; Ebersberger et al. 2007; Fujiyama et al. 2002; Haaf and Willard 1997, 1998; Kehrer–Sawatzki and Cooper 2007; Khaitovich et al. 2005; King and Wilson 1975; Kuroki et al. 2006; Laursen et al. 1992; Liu et al. 2009; Mikkelsen et al. 2005; Newman et al. 2005; Olson and Varki 2003; Patterson et al. 2006; Pennacchio and Rubin 2001; Perry et al. 2008; Sibley and Ahlquist 1987; Varki and Altheide 2005; Varki et al. 2008; Watanabe et al. 2004; Webster et al. 2003). Large variation in sequence divergence was often seen among genomic regions. For example, the last intron of the ZFY gene showed only 0.69% divergence between human and chimpanzee (Dorit et al. 1995), whereas for the OR1D3P pseudogene a divergence of 3.04% was found (Glusman et al. 2000). Thus, to have reliable estimates of the average divergences between hominoid genomes, it was concluded that sequence data from many genomic regions are needed (Chen and Li 2001). Estimates of divergence due

to nucleotide substitutions were about 1.24% between selected intergenic nonrepetitive DNA segments in humans and chimpanzees, substantially lower than previous ones, of about 3%, which included repetitive sequences (Chen and Li 2001; Ebersberger et al. 2002; Fujiyama et al. 2002; Mikkelsen et al. 2005). A greater sequence divergence (1.78%) was obtained between reported finished sequence of the chimpanzee Y chromosome (PTRY) and the human Y chromosome (Kuroki et al. 2006). Comparing the DNA sequences of unique, Y–linked genes in chimpanzee and human, evidence was found that in the human lineage all such genes were conserved, and in the chimpanzee lineage, by contrast, several genes have sustained inactivating mutations (Hughes et al. 2005).

On the other hand, the overall sequence divergence by taking regions of indels into account was estimated to be approximately 5% (Britten 2002, 2003; Cheng et al. 2005; Gibbs et al. 2007). In some short stretches of human and chimpanzee genomes, so called human–accelerated regions, significant increase of substitution divergence was found (Pollard 2009; Pollard et al. 2006a,b; Popesco et al. 2006; Prabhakar et al 2006). On the other hand, based on phylogenetic analysis of large number of DNA sequence alignments from human and chimpanzee it was found that for a sizeable fraction of our genome we share no immediate genetic ancestry with chimpanzee (Ebersberger et al. 2007).

Experimental evidence suggests that a progenitor of suprachromosomal alphoid family 3 was established and dispersed to chimpanzee chromosomes homologous to human chromosomes 1, 11, 17 and X prior to the human–chimpanzee split (Baldini et al. 1991; Durfy and Willard 1990; Warburton et al. 1996; Willard 1991). Notably, the alphoid HOR organization in the X chromosome has been conserved (Durfy and Willard 1990); only the localization of the suprachromosomal family (SF) 3 alpha satellite is substantially conserved. It was concluded that the lack of sequence or HOR conservation among human and chimpanzee indicates that most alpha satellite sequences do not evolve orthologously.

In a recent publication, Hughes et al. (2010) have shown by sequence comparison of human and chimpanzee MSY that humans and chimpanzees differ radically in sequence structure and gene content. It was concluded that, since the separation of human and chimpanzee lineages, sequence gain and loss have been far more concentrated in the MSY than in the balance of the genome, indicating accelerated structural remodeling of the MSY in the chimpanzee and human lineages during the past six million years.

The previously reported 35mer alphoid HOR in human Y chromosome (Skaletsky et al. 2003; Tyler–Smith and Brown 1987; Warburton and Willard 1996) involves the largest alphoid HOR unit found in human genome and it is of particular interest to look for divergence between alphoid HOR in human and chimpanzee Y chromosome. Alphoid HOR in chimpanzee Y chromosome was not yet reported.

Having in mind possibly important information regarding the evolutionary role of human and chimpanzee Y chromosomes and availability of their genomic sequences (Mikkelsen et al. 2005; Skaletsky et al. 2003) and a demanding task of studying bioinformatically such long HOR units, we perform here an extensive study applying novel robust bioinformatics tools GRM. We investigate the major alphoid HOR from Build 37.1 assembly of human Y chromosome and determine detailed monomer scheme and consensus sequence, finding a riddling pattern not reported previously. In the chimpanzee Y chromosome, for the first time, we identify and analyze alphoid HOR. We find that the human and chimpanzee HORs are sizeably different, both in size and composition of HOR units and in the constituting monomer structure.

Furthermore, we identify and investigate in human and chimpanzee Y chromosomes more than 20 other tandems, HORs and regularly dispersed repeats based on large repeat units, showing sizeable human–chimpanzee divergence. Most of these repeats are reported here for the first time.

II. MATERIALS AND METHODS

A. Key String Algorithm

In spite of powerful standard bioinformatics tools, there are still difficulties to identify and analyze large repeat units. For example, the detection limit of TRF is 2 kb (Gelfand et al. 2007; Warburton et al. 2008). Here, we use a new approach useful in particular for very long and/or complex repeats.

The KSA framework is based on the use of a freely chosen short sequence of nucleotides, called a key string, which cuts a given genomic sequence at each location of the key string within genomic sequence. Going along genomic sequence, the lengths of ensuing KSA fragments form KSA length array. Such array could be compared to an array of lengths of restriction fragments resulting from a hypothetical complete digestion, cutting genomic sequence at recognition sites corresponding to KSA key string. Any periodicity appearing in the KSAlength array enables identification and location of repeat in a given genomic sequence. Analysis of repeat sequences at position of any periodicity in the KSA length array gives consensus repeat unit and divergence of each repeat copy with respect to consensus. Any presence of higher order periodicity in the KSA length array reveals the presence of HOR at that location and enables determination of consensus HOR repeat unit and divergence of each HOR copy with respect to consensus.

Similarly, with a proper choice of key string, the KSA fragments a given tandem repeat into monomers, as for example cutting Alu sequence at two identical positions providing identification of Alu sequences, cuts a palindrome providing identification of large palindrome sequences and their substructure, and so on. KSA pro-

vides a straightforward ordering of KSA fragments, regardless of their size (from small fragments of a few bp to as large as tens of kilobasepairs). KSA provides high degree of robustness and requires only a modest scope of computations using PC. Due to its robustness, KSA is effective even in cases of significant deletions, insertions, and substitutions, providing detailed HOR annotation and structure, consensus sequence, and exact consensus length in a given genomic sequence even if it is highly distorted, intertwined and riddled (segmentally fuzzy repeats). Using a HOR consensus sequence, in the next step KSA computes finer characteristics, as for example the SF classification and CENP–B box/pJα distributions.

B. Global Repeat Map

The GRM program is an extension of KSA framework. GRM of a given genomic sequence is executed in five steps.

Step 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 ($4^8 = 65536$ key strings). In 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 human chromosome using PC.

Step 2 *GRM–Dom module* Determines dominant key string corresponding to fragment length for each peak in the GRM diagram from the step 1. A particular 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 dominant key string.

Step 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 repeat and provides genomic sequences of the corresponding repeat copies.

Step 4 *GRM–Cons module* Aligning all sequences of repeat copies from step 3 constructs the consensus sequence.

Step 5 *NW module* Computes divergence between each repeat copy from step 3 and consensus sequence from step 4 using Needleman–Wunsch algorithm (Needleman and Wunsch 1970).

Regarding the 8 bp choice of key string size: using an ensemble of r–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. We tested that the 8 bp key string ensemble is suitable for identification of repeat units in a wide range of lengths, from ∼10 bp to as much as ∼100 kb. However, from GRM construction it follows that fully reliable results are obtained for key string lengths not exceeding the repeat length under study.

In summary, the characteristics of GRM are:

– robustness of the method with respect to deviations from perfect repeats, i.e., substitutions, insertions, and deletions;

– use of ensemble of all 8 bp key strings as a starting point of algorithm, thus avoiding the need to choose a particular key string for any repeat structure;

– straightforward identification of repeats (tandem and dispersed), applicable to very large repeat units, as large as tens of kilobasepairs;

– easy identification of HORs and determination of consensus lengths and consensus sequences.

III. RESULTS AND DISCUSSION

Using GRM algorithm we have identified and analyzed tandem repeats, HORs and regularly dispersed repeats with large repeat units in human and chimpanzee Y chromosomes (Build 37.1 and Build 2.1 assemblies, respectively). Summary of all large repeat units identified and analyzed in this article and the human–chimpanzee comparison are given in Tables I, II, and III.

A. Alphoid Higher Order Repeat Units in Human and Chimpanzee Y Chromosome

1. Riddled HOR Scheme with 45 Distinct Alphoid Monomers in Human Y Chromosome

The largest repeat array in human Y chromosome assemblies studied here is the major alphoid HOR array and, as will be shown here, strongly diverges from the chimpanzee alphoid HOR. For this reason, we first present our results for alphoid HORs. In the contig NT_087001.1 in centromere of human chromosome Y and in NT_011878.9 in the pericentromeric region on the proximal side of p arm (DYZ3 locus), we identify the peripheral segments of the major block of alphoid HOR array. In the spacing between these two contigs lies a large central section of this HOR array. This spacing of ∼3 Mb was not sequenced so far in the Build 37.1 assembly. The GRM results for alphoid monomer structure of the two peripheral HOR segments are shown in Fig. 1 and Supplementary Table 1. In Fig. 1, we use a method of schematic presentation described by Rosandić et al. (2006).

TABLE I Tandem repeats, HORs and dispersed repeats with large repeat units in contigs of human Y chromosome.

Repeat unit (bp)	Structure	Character	Contig	Chr Y start position	Chr Y end position	Length
∼171	PRU[a]	Alpha satellite	NT_011878.9	10083775	13131913	3048138
35mer(45mer[c])	SRU[a]	Alphoid HOR	NT_087001.1			
125	PRU	Tandem	NT_011875.12	22216726	22513032	296306
∼545	PRU[a]	Tandem	see Table V			12577
∼1641[c]	SRU[a]	Regularly dispersed				
∼23541[c]	TRU[b]	Third order tandem	NT_011903.12	24023693	24070760	47067
				24312159	24333896	21737
				24544818	24566560	21742
			NT_011875.12	23654663	23713744	59081
∼2385	PRU[a]	Tandem	NT_011903.12	25298078	25312458	14380
∼4757[c]	SRU[a]	2mer HOR		25376692	25424719	48027
∼7155[c]	SRU[a]	3mer HOR		26929417	26948531	19114
				27001927	27038009	36082
5	PRU[a]	Tandem	NT_025975.2	58819393	58917657	98264
∼3579	SRU[a]	715mer HOR				
5	PRU[a]	Tandem	NT_113819.1	13690637	13747836	57199
∼5607[c]	SRU[a]	1123mer HOR				
∼5096[c]	PRU[b]	Dispersed	NT_011875.12	20121395	20126501	5106
				20003268	20008374	5106
		Dispersed	NT_011903.12	26206614	26211701	5087
				27750731	27755818	5087
∼10848	PRU[b]	Tandem	NT_011903.12	25312733	25341062	28329
				26984151	27001645	17494
∼15766[c]	PRU[b]	Dispersed	NT_011875.12	23167813	23183579	15766
				23209651	23225434	15783
∼15775[c]	PRU[b]	Tandem	NT_011896.9	6543373	6574923	31550
	PRU[b]	Dispersed	NT_011651.17	14540408	14556183	15775
∼20309	PRU[b]	Tandem	NT_011878.9	9293306	9374535	81229
	PRU[b]	Tandem	NT_086998.1	9170808	9241328	70520
∼60910[c]	PRU[b]	Dispersed	NT_011875.12	19697222	19759044	60917
				20420735	20482553	60909
∼72140[c]	PRU[b]	Dispersed	NT_011875.12	19829682	19900381	70699
				20279397	20350098	70701

PRU primary repeat unit, *SRU* secondary repeat unit, *TRU* tertiary repeat unit, *dispersed* dispersed at random spacings, *regularly dispersed* dispersed at regular spacings

[a]Described in text

[b]Described in Supplementary text

[c]For the first time reported in this work

In each of these two segments we identify 45 distinct alphoid monomers, denoted $m01, \ldots, m45$, arranged head–to–tail in the same orientation and mutually diverging by ∼20%. The consensus length of this 45mer HOR is 7662 bp. Here, an alphoid monomer is assigned as constituent of HOR if it appears in at least two HOR copies at a very low mutual divergence. Consensus sequences of monomers forming HOR are shown in Supplementary Table 2. In both the contigs, the consensus sequences of monomers constituting HOR are equal, reflecting the fact that they are two peripheral segments of the same HOR array (Table IV).

Divergence between monomers in individual HOR copies and the corresponding consensus monomers is very low (on the average 0.3%). However, the HOR structure is characterized by some pronounced monomer deletions and insertions, giving a riddled pattern (Table IV) due to a variety of lengths of HOR copies (Fig. 1). We find monomer deletions in seven HOR copies, monomer insertions in two, and nonalphoid insertions of 0.2 to 0.3

TABLE II Tandem repeats, HORs and dispersed repeats with large repeat units in contigs of chimpanzee Y chromosome.

Repeat unit (bp)	Structure	Character	Contig	Chr Y start position	Chr Y end position	Length
∼171	PRU[a]	Alpha satellite	NW_001252921.1	7108946	7151404	42458
30mer[c]	SRU[a]	Alphoid HOR				
∼550[c]	PRU[a]	Tandem See Table IV				30832
∼1652[c]	SRU[a]	Regularly dispersed				
∼23578[c]	TRU[a]	Third order tandem	NW_001252921.1	7707476	7728531	21055
				8130226	8160370	30144
				8433315	8464030	30715
				8866559	8897264	30705
				9166900	9197050	30150
				9598779	9628923	30144
∼2383	PRU[a]	Tandem	NW_001252917.1	3256815	3278585	21770
				3406716	3428486	21770
		Tandem	NW_001252922.1	11224963	11256302	31339
				11298117	11327074	28957
∼5096[c]	PRU[b]	Tandem	NW_001252916.1	1956128	1981606	25478
				2082379	2092569	10190
		Dispersed	NW_001252920.1	5633270	5638363	5093
		Dispersed	NW_001252924.1	12174453	12179546	5093
				12280382	12285475	5093
∼10762[c]	PRU[b]	Tandem	NW_001252919.1	276373	308349	31976
		Tandem	NW_001252921.1	2823896	2845204	21308
		Dispersed	NW_001252925.1	1219588	1230035	10447
∼10853[c]	PRU[a]	Tandem	NW_001252917.1	1130756	1160123	29367
∼64624[c]	SRU[a]			1174942	1224747	49805
∼60523[c]	PRU[b]	Tandem	NW_001252918.1	3827479	3948523	121044
	PRU[b]	Dispersed	NW_001252922.1	10310933	10371414	60481
	PRU[b]	Dispersed	NW_001252919.1	5301324	5361771	60447
∼71778[c]	PRU	Dispersed	NW_001252925.1	12394038	12465796	71758
	PRU	Dispersed	NW_001252915.1	1775843	1847647	71804
	PRU	Dispersed	NW_001252917.1	2201698	2273485	71787
	PRU	Dispersed	NW_001252919.1	5440228	5505887	65659
∼72140[c]	PRU	Tandem	NW_001252923.1	11947703	12091619	143916

For description see Table I

kb in three HOR copies. (In some HOR copies there are multiple insertions and/or deletions.)

Two out of ten HOR copies contain the 10–alphoid–monomer subsequence $m24, \ldots, m33$ (Fig. 1). These ten monomers are positioned between the monomers $m23$ and $m34$. Distance between the two highly identical 10–alphoid–monomer subsequences is ∼3 Mb.

The other 35 alphoid monomers from 45 distinct alphoid monomers in the peripheral region of major alphoid HOR form a subsequence, consisting of two segments, additionally riddled at some positions. Each of these 35 alphoid monomers appears in three or more HOR copies (Fig. 1). If we delete the 10–alphoid–monomer subsequence from the 45mer, we obtain a 5957 bp 35mer, which is similar to the secondary periodicity sequence of 5941 bp reported in (Skaletsky et al. 2003).

Discussing relationship of the initially reported 5.7 and 6.0 kb repeat units, Tyler–Smith and Brown proposed that one HOR unit is derived from the other, although more complex explanations, with both units derived from a third unknown HOR unit were considered as possible (Tyler–Smith and Brown 1987). It was considered as very unlikely that the 6.0 kb unit arose from a 5.7 kb unit by addition of two alphoid monomers, because results excluded the possibility that the two additional alphoid monomers in the 6.0 kb unit are duplications of any monomers contained in the 5.7 kb unit (Tyler–Smith and Brown 1987). Therefore, the favored hypothesis was that the shorter, 5.7 kb HOR unit arose from the longer 6.0 kb HOR unit by deletion of two alpha monomers. Extending

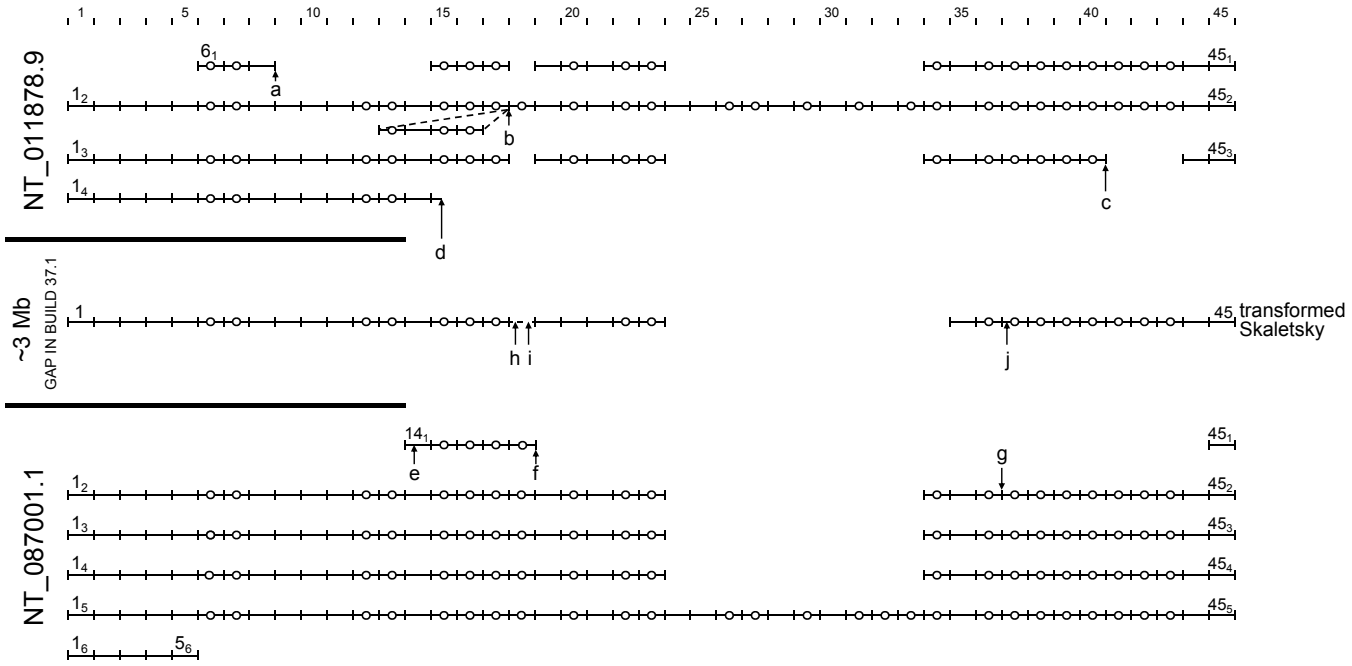

FIG. 1 Schematic presentation of aligned monomer structure of 45mer alphoid HOR (consensus length 7662 bp) in human chromosome Y (Build 37.1). This method of schematic presentation of HOR sequences is self–evident if one compares Fig. 1 and Supplementary Table 1. *Top* enumeration of columns corresponding to 45 constituent consensus monomers (enumerated Nos. 1 to 45) in consensus HOR. (For simplicity, only every fifth number is shown.) Each HOR copy is presented by a *bar* in the corresponding column numerated at the *top*. Monomers from different HOR copies corresponding to the same monomer from consensus HOR are presented by *bars* in the same column corresponding to its enumeration at the *top*. For example, in the first HOR copy the first monomer corresponds to monomer No. 6 in consensus HOR and is presented by a bar at position of 6th column (denoted by 6_1), the second monomer in the first HOR copy corresponds to monomer No. 7 in consensus HOR and is presented by a *bar* at the position of 7th column..., the fourth monomer in the first HOR copy corresponds to monomer No. 15 in consensus HOR and is presented by a *bar* at the position of 15th column..., and the last monomer in the first HOR copy (the 23rd) corresponds to the monomer No. 45 in consensus HOR and is presented by a *bar* at the position of 45th column. *Upper panel*: HOR copies in contig NT_011878.9. *Lower panel*: HOR copies in contig NT_087001.1. *Middle panel*: The 5941 bp secondary periodicity sequence from Skaletsky et al. (2003) mapped into alphoid monomers $\{m\}$. For mapping of $\{w\}$–monomers from Skaletsky et al. (2003) into $\{m\}$–monomers, see the text and Supplementary Tables 24. Open circle: pJα motif (essential part) in alpha monomers. The $m05$ monomer from the last incomplete HOR copy (5_6) in contig NT_087001.1 is followed by alpha satellite monomeric region (not shown here). *a* After $m08$: 210 bp insertion (no similarity to HOR monomers); *b* after $m13$–$m16$ duplication (inserted after $m17$) there are two insertions: 170 bp insertion (differing in 19 bases from $m24$ and $m34$ as the closest monomers from HOR) and 168 bp insertion (differing in 20 bases from $m28$ as the closest monomer from HOR); *c* after $m40$: 278 bp insertion (no similarity to HOR monomers); *d* after the first 34 bases from $m15$: end of the contig NT_011878.9; *e* the last 166 bases of $m14$: start of the contig NT_087001.1; *f* after $m17$: 311 bp insertion (no similarity to HOR monomers); *g* after $m36$: 171 bp insertion (differing in 13 bases from $m23$ as the closest monomer from HOR); *h*, *i* two deletions in $w20$; *j* 53 bp nonalphoid insertion in $w29$

similar considerations to the present case, the 35mer in internal centromere region could be considered as arising from 45mer by deletion of ten alphoid monomers which are all distinct from the monomers in 35mer. This is consistent with a general view (Warburton and Willard 1996) that a type of polymorphism found in alphoid arrays can be related to HOR units that differ by an integral number of alphoid monomers.

Divergence pattern provides an additional evidence that ten additional alphoid monomers $m24, \ldots, m33$ are constituents of major HOR. Mutual divergence between these ten monomers is similar to their mean divergence with respect to the other 35 monomers (Table V).

2. Suprachromosomal Family Assignment of Monomers in 45mer HOR

Studies of sequence comparison of alpha satellite monomers in human chromosomes revealed 12 types of monomers, forming five suprachromosomal families (SFs), which descend from two basic subsets of monomers, A and B: to the subset A belong the SF types J1, D2, W4, W5, M1, and R1, and to the subset B belong J2, D1, W1, W2, W3, and R2 (Alexandrov et al. 2001; Romanova et al. 1996; Warburton and Willard 1996). We determine the SF assignments of monomers constituting alphoid HOR by pairwise comparison between every monomer from HOR to every of 12 SF consensus

TABLE III Correspondence of large repeat and HOR units in Y chromosome contigs of human and chimpanzee.

Human	Chimpanzee
125 bp PRU	-
∼171 bp PRU	∼171 bp PRU
35mer/45mer SRU	30mer SRU
∼545 bp PRU	∼550 bp PRU
∼1641 bp SRU	∼1652 bp SRU
∼23541 bp TRU	∼23578 bp TRU
∼2385 bp PRU	∼2383 bp PRU
∼4757 bp SRU	-
∼7155 bp SRU	-
5 bp PRU	5 bp PRU
∼3579 bp SRU	-
∼5096 bp PRU (dispersed)	∼5096 bp PRU
5 bp PRU	5 bp PRU
5607 bp SRU	-
∼10.8 kb PRU (within ∼20309 bp PRU)	∼10762 bp PRU
∼10848 bp PRU	∼10853 bp PRU
-	∼64624 bp SRU
∼15766 bp PRU (dispersed)	Dispersed fragments
∼15775 bp PRU	-
∼60910 bp PRU (dispersed)	∼60523 bp PRU
∼72140 bp PRU (dispersed)	∼72140 bp PRU
	∼71778 bp PRU (dispersed)

PRU primary repeat unit, *SRU* secondary repeat unit, *TRU* tertiary repeat unit

TABLE IV Riddled pattern with variety of number of monomers in human alphoid HOR copies (Build 37.1 assembly).

HOR copy no.	No. of monomers	
	Counting distinct monomers	Counting all monomers
1[a]	23	23
2	45	51
3	31	31
4[a]	14	14
5[a]	6	6
6	35	36
7	35	35
8	35	35
9	45	45
10[a]	5	5

[a]Truncated at the start or end of the contig. Copies No. 1-4 are from contig NT_011878.9. Copies No. 5–10 are from contig NT_087001.1

TABLE V Average divergence between two subsets of alphoid monomers from 45mer HOR copies.

Monomer comparison	Divergence (%)
10 vs. 10	∼19
10 vs. 35	∼20
35 vs. 35	∼21

10 denotes the subset of ten new monomers $m24, \ldots m33$

35 denotes the subset of 35 monomers $m01, \ldots m23$ and $m34, \ldots m45$

monomers from Romanova et al. (1996). A 45×12 divergence matrix is constructed between 45 monomers from HOR and 12 SF consensus monomers from Romanova et al. (1996). To each monomer from HOR we assign the SF classification of the most similar SF consensus monomer. In this way we find that, out of forty–five monomers from HOR, forty monomers are of M1 type (in most cases the second lowest divergence corresponds to R2, and in three cases the M1 and R2 divergences are equal), and five are of R2 type (in these cases the second lowest divergence corresponds to M1 type).

The differences between A and B subsets are, in general, concentrated in a small region which matches functional protein binding sites for pJα in subset A and for CENP–B in subset B (Romanova et al. 1996). Analyses of human genome have indicated that a CENP–B box appears in the subset B monomers (in about 60% of B–type monomers) and is absent in the subset A monomers; while the pJα motif would occur only in some of monomers from the subset A and not in the subset B monomers (Romanova et al. 1996).

After determining the SF classification of monomers in consensus HOR, we investigate the appearance of CENP–B box and pJα motif in these monomers. We find that the pJα motif (essential part) is present in 55% of ten new alphoid monomers and similarly, in 57% of the other 35 monomers, while the CENP–B box is completely absent (Fig. 1). Consensus HOR has a robust pJα distribution, containing 25 pJα motif copies. All alphoid monomers in consensus HOR are significantly more similar to pJα motif than to the CENP–B box: themean deviation is 0.6 bp for the pJα motif and 4.7 bp for the CENP–B box, reflecting that the absence of pJα motif in some of monomers from 45mer HOR can be attributedmostly to a single nucleotide mutation within an initially pJα motif.

Since the pJα motif is essential for protein binding, an interesting question is whether the monomers with and without pJα motif have different sequence divergences. In this respect, pairwise divergence among 45 monomers shows no dependence on the presence or absence of the pJα motif.

It should be noted that HOR copies in chromosome Y are the only reported case where pJα motif is present and CENP–B box absent.

In this connection, we note a unique case of 13mer HOR (2214 bp consensus length) in chromosome 5, which contains neither CENP–B box nor pJα motif (Rosandić et al. 2006).

3. Alignment of Peripheral and Internal Human HOR Copies

Let us now compare our consensus HOR for the peripheral parts of major HOR alphoid block (DYZ3 locus) (Supplementary Table 2) to the 5941 bp secondary periodicity sequence in its internal part reported by Skaletsky et al. (2003) which corresponds to the sequence gap between the contigs NT_011878.9 and NT_087001.1 in the Build 37.1 assembly.

First, we fragment the 5941 bp sequence from Skaletsky et al. (2003) into 35 constituent alpha monomers, denoted $w01, \ldots, w35$ (Supplementary Table 3). We find a peculiar feature of this secondary periodicity sequence: two of its constituent monomers, $w20$ and $w29$, exhibit sizeable length deviation from the alpha satellite consensus length of 171 bp: the alphoid monomer $w20$ has a length of 104 bp (i.e., 67 nucleotides are deleted with respect to consensus alpha monomer length) while the monomer $w29$ is 224 bp long, containing a 53 bp nonalphoid insertion with respect to consensus alpha monomer.

To align the internal monomer sequence $\{w\}$ (Supplementary Table 3) to the peripheral monomer sequence $\{m\}$ (Supplementary Table 2), we shift the start position of alpha monomers $m01, m02, \ldots, m45$, obtaining the sequence denoted by $n01, n02, \ldots, n45$ (Table VI). The 35 alphoid monomers from the sequence $\{w\}$ are aligned to 35 out of 45 monomers $\{n\}$ (Table VI and Supplementary Table 4). The sequences $n26, \ldots, n35$ have no counterpart in the $\{w\}$ sequence which corresponds to internal part of major alphoid HOR from Skaletsky et al. (2003).

TABLE VI Transformation between monomer sets $\{m\}$ and $\{n\}$ and alignment between alphoid monomer sets $\{w\}$ and $\{n\}$

Transformation
$n01(169) = m44(\ldots 113) + m45(056\ldots)$
$n02(166) = m45(\ldots 110) + m01(056\ldots)$
$\ldots$
$n45(170) = m43(\ldots 114) + m44(056\ldots)$
Alignment
$w01 = n01$
$\ldots$
$w25 = n25$
$w26 = n36$
$\ldots$
$w35 = n45$

For definition of monomers $\{n\}$ and $\{w\}$ see Supplementary Tables 3 and 4. In the transformation from $\{m\}$ to $\{n\}$ the notation $m44(\ldots 113)$ denotes the last 113 bases in $m44$, $m45(056\ldots)$ denotes the first 56 bases in $m45$, and so on (Supplementary Table 4). In alignment between $\{n\}$ and $\{w\}$ the 35 alphoid monomers from the sequence $\{w\}$ are aligned to 35 out of 45 monomers from the sequence $\{n\}$. Here, the only significant differences appear between $w20$ and $n20$ (due to the presence of deletion in $w20$), and between $w29$ and $n39$ (due to presence of insertion in $w29$). The monomers $n26, \ldots n35$ have no counterpart in the set $\{w\}$ which corresponds to the internal part of major alphoid HOR

4. Global Repeat Map for Riddled Alphoid HOR and Characteristic HOR–Signature in Human Chromosome Y

To investigate more closely the major alphoid HOR array in human chromosome Y, we compute the GRM diagram for genomic sequence of Y chromosome (Fig. 2). The most pronounced peaks in this diagram correspond to following tandem repeats in chromosome Y: the alphoid repeats (GRM peaks at multiples of the ∼171 bp repeat unit), the 125 bp repeats (GRM peaks at multiples of the 125 bp repeat unit), GRM peaks at multiples of 5 bp repeat unit and GRM peaks corresponding to ∼20.3 kb repeat unit. In addition, there are nine pronounced GRM peaks at repeat lengths above 2000 bp.

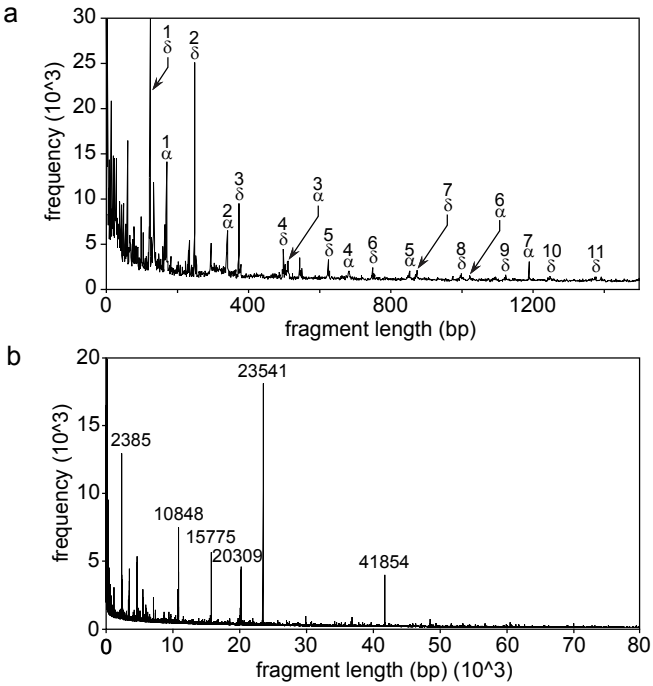

FIG. 2 GRM diagram for Build 37.1 genomic assembly of human chromosome Y for the intervals of fragment lengths: **a** 0-1500 bp. There are two pronounced tandem arrays with repeat units below 1.5 kb: the alphoid tandem repeat with alpha satellite repeat unit of 171 bp and the overlapping tandem repeat with repeat unit of 125 bp. The peaks at multiples of alphoid monomer repeat unit 171 bp, $n{\cdot}171$ bp, are denoted by $n\alpha$. The peaks at multiples of 125 bp repeat unit, $n{\cdot}125$ bp, are denoted by $n\delta$. **b** 0-80000 bp. Pronounced peaks above 2 kb are denoted by the corresponding fragment lengths. The most pronounced peaks are approximately at 2385, 10848, 15775, 20309, 23541, and 41584 bp. *Arrow i*: peak corresponding to 715mer. *Arrow j*: peak corresponding to 1123mer. For description of peaks see the text

Here, we perform detailed study for alphoid HOR repeat sequence. Analyzing partial contributions to GRM diagram of chromosome Y from individual contigs we find that the largest frequency contributions to alphoid HOR peaks are arising from the contigs NT_011878.9 and NT_087001.1. The relevant intervals of fragment lengths

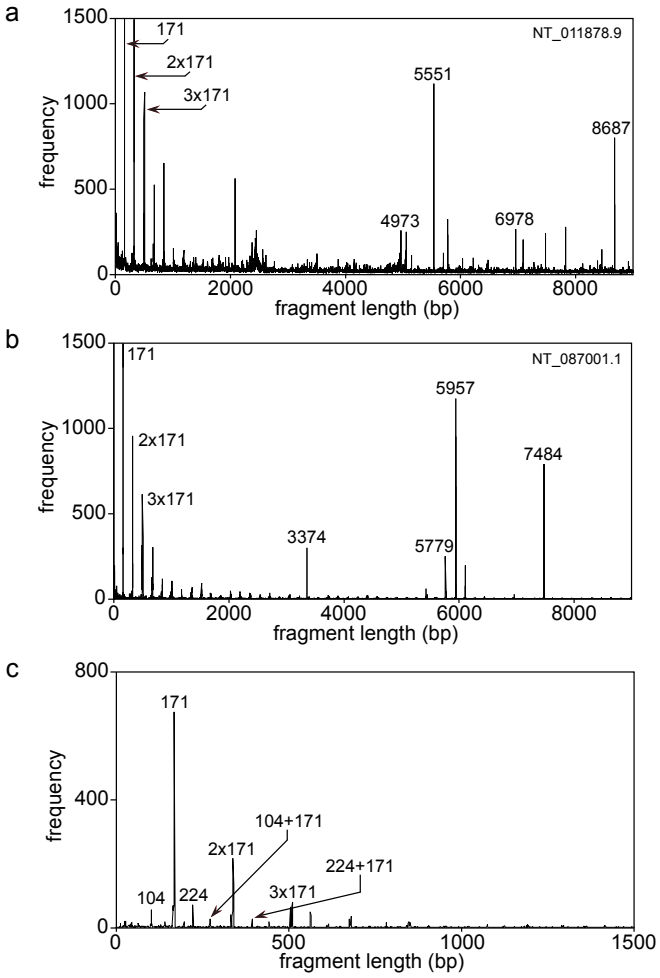

FIG. 3 GRM diagrams for sequences in contigs containing alphoid HOR in chromosome Y: **a** NT_011878.9, **b** NT_087001.1, and **c** secondary periodicity sequence for internal part of major interior alphoid HOR block (genomic sequence from Skaletsky et al. (2003))

for these two contigs are shown in Fig. 3a and b, respectively. In both the figures peaks at approximate multiples of basic repeat length ∼171 bp are decreasing with increasing multiple orders. That is a natural trend for tandem repeats. However, we do not find a peak corresponding to the HOR length, which for regular HORs in other chromosomes appears at their consensus lengths. This is because the Build 37.1 assembly of chromosome Y encompasses only peripheral tails of major HOR array and those exhibit sizeable riddling in both relevant contigs, as shown in the monomer structure of peripheral HOR copies in Fig. 1. For these riddled HOR copies there is no dominating consensus length and therefore no peak corresponding to consensus length is present. Instead, the GRM diagram shows more intricate HOR–related peaks which characterize riddled alphoid HOR copies. These peaks will be referred to as GRM HOR–signature. Most pronounced GRM HOR–signature peaks of riddled HOR pattern in peripheral regions of major alphoid HOR in chromosome Y are at the lengths shown in Fig. 3a, b. These characteristic fragment lengths are fully consistent with the riddled HOR structure from Fig. 1.

As an example, let us consider the largest GRM HORsignature peak at 5551 bp, characterizing HOR pattern in NT_011878.9. This peak arises from approximate repeat of the 1_3–14_3 subsequence at the position of the 1_4–14_4 subsequence. The distance l between the corresponding bases in these two subsequences (Table VII) is equal to a distance between monomers 1_3 and 1_4 (Fig. 1 and Supplementary Table 1).

TABLE VII Contributions to the fragment length 5551 bp alphoid GRM HOR–signature peak for human Y chromosome

Length (bp)	Distance
2896	$1_3 - 17_3$
848	$19_3 - 23_3$
1194	$34_3 - 40_3$
278	Nonalphoid insertion
335	$44_3 - 45_3$
$\sum$ 5551	

Therefore, the GRM diagram shows a pronounced peak at the 5551 bp fragment length, reflecting the riddling structure of HORs. Similarly, we interpret all the other HOR–signature peaks which characterize riddling in HOR copies from Fig. 1.

In addition to GRM computation for Build 37.1 sequence of chromosome Y, let us comment on the GRM HOR–signature related irregularity (monomers $w20$ and $w29$) in the interior region of major alphoid HOR array in chromosome Y (Supplementary Tables 3, 4). Figure 3c displays GRM diagram computed for the 5941 bp secondary periodicity sequence from Skaletsky et al. (2003). Here again, we see the main pattern of monomer multiples ∼171, $\sim 2 \times 171$, $\sim 3 \times 171$ bp,... with decreasing frequencies for increasing multiples. In addition, we obtain two weak subsequences of peaks, at fragment lengths ∼104 bp, ∼ (104+171 bp), ∼ (104+2×171 bp), ... and at ∼224 bp, ∼(224 + 171 bp), ∼(224 + 2 × 171 bp),... These two additional weak subsequences are due to two distorted monomers in the 35mer periodicity (HOR) sequence that we deduced from the HOR genomic sequence in Skaletsky et al. (2003): the alphoid monomer $w20$ has a length of 104 bp (i.e., 67 nucleotides are deleted with respect to consensus monomer) while the monomer $w29$ has the length 224 bp, containing a 53 bp nonalphoid insertion with respect to consensus monomer. Such deletions/insertions in two distant alphoid monomers within HOR are absent in the peripheral regions of major HOR array in chromosome Y, i.e., they are absent in Build 37.1 assembly. Therefore, GRM diagrams of these regions (Fig. 3a, b) do not have these two additional weak subsequences of peaks. This actualizes the interest for future extension of Build assembly to the region of sequence gap of ∼3 Mb between the contigs NT_011878.9

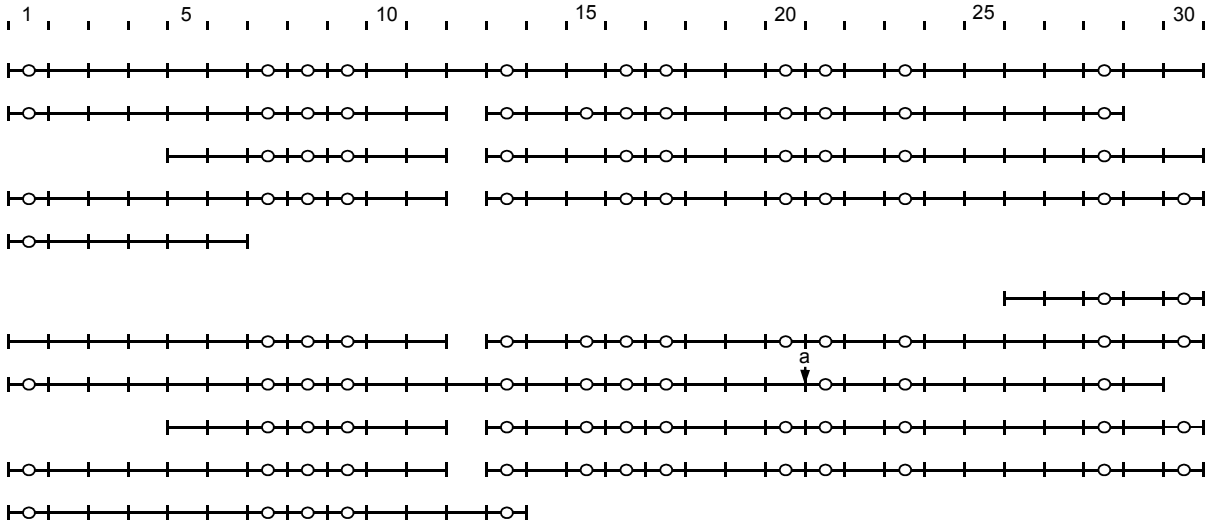

FIG. 4 Schematic presentation of aligned monomer structure of 30mer alphoid HOR (consensus length 5066 bp) in chimpanzee chromosome Y (Build 2.1, contig NW_001252921.1). *Top row* enumeration of 30 constituent alpha monomers from consensus HOR. *Upper panel*: HOR copies in interval 264-20019. *Lower panel* reverse complement of HOR copies in interval from 20618-42459. After monomer No. 20 (label *a*): 41 bp insertion (no similarity to monomers in 30mer). For comparison with human alphoid HOR see Fig. 1. *Open circle* pJα motif (essential part) in alpha monomers

and NT_087001.1.

5. Riddled 30mer HOR Scheme in Chimpanzee Chromosome Y

Applying GRM to the chimpanzee chromosome Y, we find two 30mer HOR arrays in chimpanzee contig NW_001252921.1 (NCBI Build 2.1), positioned one after another (with a gap of 599 bp in between) at the front part of the contig. The first HOR, truncated at the start of the contig is referred to as direct. In fact, it seems to be a truncated tail of a major HOR block positioned in unsequenced domain in front of the contig NW_001252921.1. We find that the reverse complement of the second HOR array is highly identical to the first HOR array, and therefore this second HOR array is referred to as reverse complement. This indicates that the direct and reverse complement HOR arrays are positioned on the opposite arms of a palindrome.

Our results for detailed monomer scheme of these two peripheral HOR arrays, which are reverse complement to each other, are shown in Fig. 4 and Supplementary Table 5. The consensus length of 30mer HOR unit is 5066 bp (consensus sequence in Supplementary Table 6).

In GRM diagram of the whole chimpanzee Y chromosome (Fig. 5), the peak at 5066 bp fragment length is much weaker than the near–lying 5096 bp peak of another repeat structure (see Tables II, III) and is therefore overshadowed. For this reason, we compute the GRM diagram selectively for alphoid HOR–containing section of genomic sequence at the start of contig NW_001252921.1 (positions 1–20019) (Fig. 6). In Fig. 6, in the length interval between 0.1 and 1 kb there are pronounced peaks approximately at multiples of alphoid monomer repeat unit 171 bp (Fig. 6a), in analogy to Fig. 5a for the whole chimpanzee chromosome Y. Furthermore, the HOR–signature peaks are clearly seen in Fig. 6b as pronounced peaks at 5066 bp ($\sim 30 \times 171$ bp, denoted as 30α), 4895 bp ($\sim 29 \times 171$ bp, denoted as 29α), 3884 bp ($\sim 23 \times 171$ bp, denoted as 23α), and 8777 bp ($\sim 52 \times 171$ bp, denoted as 52α). These HOR–signature peaks can be also deduced directly from HOR structure from Fig. 4 and Supplementary Table 5.

For example, the 8777 bp (52α) HOR–signature peak arises from the approximate repeat of the 1_2–4_2 subsequence at position of the 1_4–4_4 subsequence (the 1_3–4_3 subsequence is missing due to riddling) (Table VIII). Distance between the corresponding bases in these two subsequences is equal to the distance between monomers 1_2 and 1_4.

TABLE VIII Contributions to fragment length 8777 bp alphoid GRM HOR–signature peak for chimpanzee Y chromosome

Length (bp)	Distance
1868	$1_2 - 11_2$
2683	$13_2 - 28_2$
1198	$5_3 - 11_3$
3028	$13_3 - 30_3$
$\sum 8777$	

Similarly, we interpret all the other pronounced HORsignature peaks in Fig. 6b. The frequencies of these peaks are sizably smaller than of peaks arising from some other tandem repeats and therefore are overshadowed in Fig. 5 for the whole chimpanzee Y chromosome. We note that the HOR–signature peaks at 3884, 4895, 5066, and

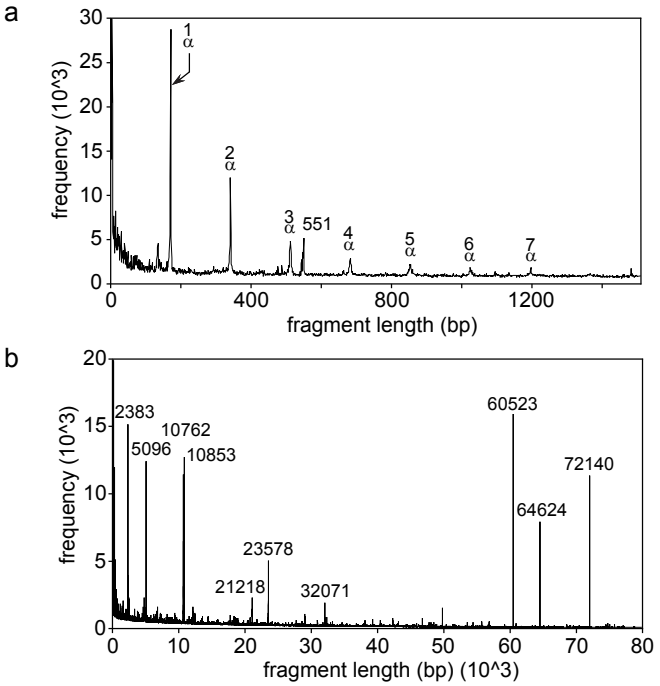

FIG. 5 GRM diagram for Build 2.1 genomic assembly of chimpanzee chromosome Y for intervals of fragment lengths: **a** 0-1500 bp. There is only one pronounced tandem array with repeat units in the interval between 0.1 and 1.5 kb: the alphoid tandem repeat with alpha satellite repeat unit of 171 bp. The peaks at multiples of alphoid monomer repeat unit 171 bp, $n{\cdot}171$ bp, are denoted by $n\alpha$. **b** 0-80000 bp. Pronounced peaks above 2 kb are denoted by the corresponding fragment lengths. The most pronounced peaks above 1.5 kb are approximately at 2383, 5096, 10762, 10853, 21218, 23578, 32071, 60523, 64624, and 72140 bp. For description of peaks see the text

8777 bp are the only significant GRM peaks above 1.5 kb in Fig. 6b.

Some peaks from GRM diagram for the whole chromosome Y (Fig. 5) are missing in GRM diagram for the HOR section in Fig. 6a. For example, the peak at 551 bp from Fig. 5a is missing in Fig. 6a, because the repeat unit of 551 bp is positioned outside of the HOR–section of genomic sequence included in Fig. 6a.

In addition to the equidistant multiple alphoid peaks, in the GRM diagram in Fig. 6a there is a family of weaker equidistant peaks at fragment length 118, $118+\alpha$, $118+2\alpha$, $118+3\alpha$,... (like in Fig. 5, here α, 2α, 3α,... denote multiples of alpha monomer length $\sim$171 bp). This weak equidistant family of repeat lengths is based on the 118 bp peak. The origin of this peak is that one of monomers within HOR, $m25$, is truncated, with size reduced from the standard value $\sim$171 to 118 bp. (Observe that we find an analog appearance of additional bands based on monomers of irregular length, 104 and 224 bp, for two human monomers in 35mer alphoid HOR in the interior part of HOR array.)

6. Comparison of Alpha Satellite Monomers in Human 45mer and Chimpanzee 30mer HORs

Computing divergence between 45 human consensus alpha monomers from consensus 45mer HOR and 30 chimpanzee consensus alpha monomers from consensus 30mer HOR (Supplementary Table 7) we see that due to scattering of divergences and the absence of any small divergence, none of chimpanzee monomers can be assigned to a particular human monomer (Supplementary Table 8). In the whole human–chimpanzee divergence matrix the lowest divergence value is 12%, appearing in a few cases only (Table IX). The mean value of the lowest

TABLE IX Illustration of divergences of human monomers $m01$ and $m24$ with respect to 30 chimpanzee monomers

Human monomer	No. of chimpanzee monomers	Divergence (%)
$m01$	Two	21
$m01$	Four	22
$m01$	Three	23
$m24$	Three	12
$m24$	Three	13
$m24$	One	14

This means, for example, that the lowest divergences between $m01$ (human) monomer and each of 30 chimpanzee monomers is 21% (with respect to two chimpanzee monomers), 22% (with respect to four chimpanzee monomers), etc.

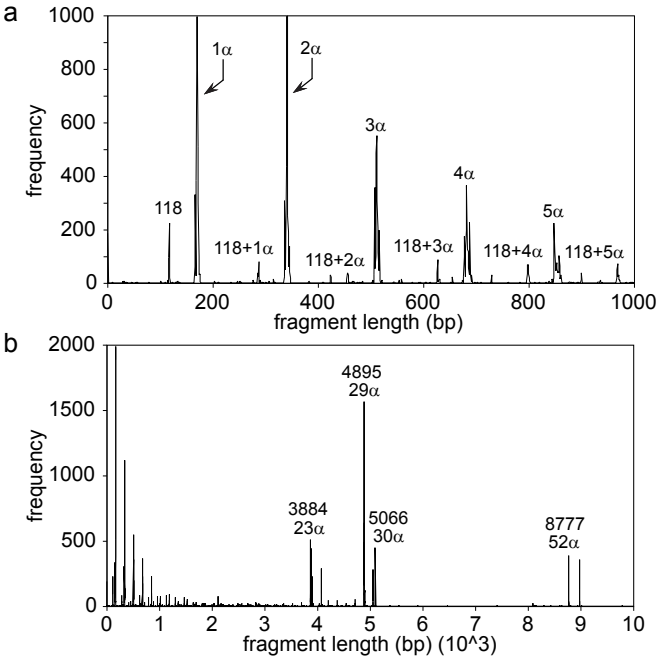

FIG. 6 GRM diagram for HOR containing section from positions 1-20019 bp in the chimpanzee contig NW_001252921. Intervals of fragment lengths: **a** 0-1000 bp, **b** 0-10000 bp. For description of peaks see the text

human–chimpanzee divergence for each human monomer is 17% (Supplementary Table 8). The absence of identity between particular human and chimpanzee monomers from alphoid HORs is also seen from the mean values of divergences in Table X.

TABLE X Comparison of mean values of human and chimpanzee consensus monomer divergences

	Divergence (%)
45 Human vs. 45 human	19
30 Chimpanzee vs. 30 chimpanzee	21
45 Human vs. 30 chimpanzee	23

45 human denotes the set of consensus alpha monomers from human 45mer HOR, and 30 chimp from chimpanzee 30mer HOR

On the other hand, we find that alpha monomers in 30mer HORs in chimpanzee Y chromosome are predominantly of M1 SF type, similarly as alpha monomers in 35mer/45mer HORs in human chromosome Y. Accordingly, similarly as for human Y chromosome, monomers in chimpanzee Y chromosome are also characterized by the presence of pJα motif and the absence of CENP–B box (Fig. 4). As already noted, the human Y chromosome was the only known case where pJα motif is present and CENP–B box absent and now we see that the chimpanzee Y chromosome shares this feature.

As to the degree of riddling, the human HOR is more riddled than the chimpanzee HOR. In particular, the human HOR has more insertions than the chimpanzee HOR, which is reflected in their respective GRM HOR signature.

7. Peculiarities of Alphoid HOR in Human Y Chromosome

We show that HOR structure in the peripheral regions of the major alphoid block in human chromosome Y is more complex than the previously reported structure for the internal region. In this computational study, we identify and fully characterize the peripheral region, in particular finding ten new monomers constituting alphoid HOR copies, different from the known 35 constituent-monomers, giving evidence for the presence of 45mer in the peripheral region of HOR array. Furthermore, while 33 out of 35 constituting alphoid monomers in HOR copies in the interior HOR region are highly homologous to the corresponding monomers in the peripheral region, we find that the remaining two monomers in the interior region have a sizeable deletion and nonalphoid insertion, respectively, with respect to the corresponding monomers from the peripheral region. The study of these riddled HOR copies may be valuable for understanding possible sources of genomic diversity, but also has the potential to provide useful markers for medical, population, and forensic genetic studies, and may give a route for identifying mechanisms of DNA sequence evolution.

Some peculiarities studied in this work regarding the major alphoid HOR that may shed some new light at the mysteries of human Y chromosome are:

The 33 consensus monomers from the peripheral HOR structure are highly identical to the aligned 33 monomers of previously reported secondary periodicity sequence from Skaletsky et al. (2003). On the other hand, we find peculiar differences: the 10mer alphoid sequence, inserted in the peripheral HOR structure, is absent in the reported internal structure; and in the previously reported internal secondary periodicity structure one constituent alphoid monomer has a sizeable deletion (67 bp) and the other a sizeable nonalphoid insertion (53 bp) accompanied by clustered substitutions of 11 bases with respect to the peripheral HOR structure.

The highly identical alphoid 10mer insert appears in both peripheral regions of major HOR, but was not reported so far in the internal centromere region between the two peripheral regions.

The peripheral regions of major HOR alphoid block reveal coexistence: on one hand, very low divergence between the aligned constituent alpha monomers from different HOR copies (average divergence$\sim$0.3%) and, on the other hand, pronounced riddling due to deletions and insertions of alpha monomers and/or due to insertions of nonalphoid segments. The HOR copies in chromosome Y are the only known case where the pJα motif is present and CENP–B box absent.

The major alphoid HOR in Y chromosome exhibits more deletions and insertions of alphoid monomers and highly distorted insertions than HORs in other chromosomes.

8. Difference Between Humans and Chimpanzees Alphoid HOR Repeat Units

The number of different monomers constituting HOR in human Y chromosome (45 monomers in the peripheral sections of major HOR array, and 35 monomers in the interior section) is different than in the chimpanzee genome (30 monomers).

HOR pattern in the sequenced domain in Build 37.1 assembly (peripheral region) is characterized by substantial riddling, which is more pronounced in human than in chimpanzee genome.

All alpha satellite monomers constituting major human 35/45mer HOR are different from monomers constituting chimpanzee 30mer HOR by$\sim$20%, which is comparable to divergence between monomers within a single HOR copy.

The lengths of major alphoid HOR arrays in human and chimpanzee are widely different, $\sim$3 and $\sim$1 Mb, respectively.

B. Other Human and Chimpanzee Tandem, HOR and Regularly Dispersed Repeat Arrays Based on Large Repeat Units

Besides the alphoid HOR, in human Build 37.1 and chimpanzee Build 2.1 Y chromosome assemblies we find over 20 other large repeat units (Tables I, II, III). Some of large repeat units appear both in human and in chimpanzee genomic assembly, and some in human only or in chimpanzee only. We describe here some pronounced repeats identified from GRM diagrams (labeled a in Tables I, II). The remaining repeats (denoted b in Tables I, II) are described in Supplementary information.

1. Chimpanzee ∼550 bp Primary Repeat Unit, ∼1652 bp 3mer HOR Secondary Repeat Unit, and ∼23578 bp Tertiary Repeat Unit

In the GRM diagram for chimpanzee Y chromosome in the length interval between 100 and 1500 bp, besides the major peaks associated with alphoid HOR and tandem repeat based on the 125 bp repeat unit, there is additional pronounced peak at ∼550 bp (Fig. 5a). Using GRM, we find that this peak arises due to the appearance of 3mer HOR copies constituted from three∼550 bp monomers, denoted *mc*01 *mc*02 and *mc*03. These monomers are mutually diverging by ∼8%, while different 3mer HOR copies mutually diverge by only ∼1%. About eight times smaller divergence between 3mer copies then between individual monomers within each 3mer are a signature of HOR. However, these HOR copies are not in tandem, in contrast to previously known HOR structures; instead, they are dispersed with rather regular spacings. Consensus sequences of three monomers *mc*01 *mc*02 and *mc*03, determined from NW_001252921.1 (using key string AGGTACTG) are given in Supplementary Table 9. The main contributions to the ∼550 bp GRM peak arise from the array of ∼550 bp monomers within each 3mer copy.

Performing the GRM analysis we find 20 dispersed HOR copies (Table XI). In addition, in four HOR copies in NW_001252921.1 one of three ∼550 bp monomers is deleted. In NW_001252921.1, we find dispersed highly identical 3mer HORs, direct and reverse complement. HOR copies after the first one are grouped into five pairs of 3mers:

D S D
R S R
D S D
R S R
D S D

where D is the direct 3mer copy, R is the reverse complement 3mer copy, and S is the spacing of ∼24 kb (see Table XI). (Three of 3mer copies in these pairs of 3mer copies are truncated from three to two monomers.) Since the two 3mer copies in each pair are separated by spacing S, there is no GRM peak at ∼1.65 kb. Instead, this gives rise to a tertiary repeat unit, with a ∼24 kb peak (more precisely ∼23578 bp) in the GRM diagram.

TABLE XI Dispersed 3mer HOR copies based on ∼550 bp monomer in chimpanzee Y chromosome

Contig	HOR copy start position	Direction	Monomers in HOR copy
NW_001252921.1	618296	RC	*mc*03 *mc*02 *mc*01
	1021470	D	*mc*01 *mc*03
	1044498	D	*mc*01 *mc*02 *mc*03
	1330195	RC	*mc*03 *mc*02 *mc*01
	1353792	RC	*mc*03 *mc*02 *mc*01
	1757803	D	*mc*01 *mc*02 *mc*03
	1781391	D	*mc*01 *mc*02 *mc*03
	2063781	RC	*mc*03 *mc*02 *mc*01
	2087364	RC	*mc*03 *mc*01
	2490023	D	*mc*01 *mc*03
	2513051	D	*mc*01 *mc*02 *mc*03
	2798724	RC	*mc*03 *mc*01
NW_001252926.1	232825	D	*mc*01 *mc*02 *mc*03
	516623	RC	*mc*03 *mc*02 *mc*01
	540165	RC	*mc*03 *mc*02 *mc*01
NW_001252919.1	328953	D	*mc*01 *mc*02 *mc*03
	574371	RC	*mc*03 *mc*02 *mc*01
NW_001252925.1	922846	D	*mc*01 *mc*02 *mc*03
	1197882	RC	*mc*03 *mc*02 *mc*01
NW_001252915.1	955834	RC	*mc*03 *mc*02 *mc*01

RC denotes a HOR copy having reverse complement sequence with respect to HOR copy defined as direct (D). In reverse complement HOR copy each monomer is reverse complement with respect to direct monomer sequence

We find even an approximate next higher pattern, three copies of quartic repeat unit:

R S_2 D S D S_1 R S R S_2 D S D S_1 R S R S_2 D S D S_1 R

where S_2 is spacing of ∼0.40 MB, and S_1 spacing of ∼0.28 Mb (see Table XI). The length of this unit is ∼0.73 Mb. In NW_001252921.1, we find an array of three such quartic repeat units. This would give rise to a GRM peak at ∼0.74 Mb fragment length (computation is performed here up to 100 kb fragment lengths).

We note that in NW_001252926.1 we find a D S_1 R S R subsection of the above pattern.

2. Human ∼545 bp Primary Repeat Unit, ∼1641 bp 3mer HOR Secondary Repeat Unit, and ∼23541 bp Tertiary Repeat Unit

The GRM peak at 545 bp is due to the ∼545 bp monomers, organized in dispersed 3mer HOR copies of ∼1641 bp (Table XII). The distance between start positions of two 3mer copies is again ∼24 kb, similar as in the chimpanzee Y chromosome, giving rise to the appearance of ∼23541 bp peak in GRM diagram.

TABLE XII Dispersed 3mer HOR copies based on ∼545 bp monomer in human Y chromosome

Contig	HOR copy start position	Direction	Monomers in HOR copy
NT_011903.12	76992	RC	$m03$ $m02$ $m01$
	100533	RC	$m03$ $m02$ $m01$
	365459	RC	$m03$ $m02$ $m01$
	609306	D	$m01$ $m02$ $m03$
NT_011875.12	9862260	D	$m01$ $m02$ $m03$
	9885800	D	$m01$ $m02$ $m03$
	9909341	D	$m01$ $m02$ $m03$
NT_086998.1	185824	D	$m01$ $m03$

RC denotes a HOR copy having reverse complement sequence with respect to HOR copy defined as direct (D). In reverse complement HOR copy each monomer is reverse complement with respect to direct monomer sequence

The 23541 bp repeat unit corresponds to previously reported 23.6 kb repeat units containing RMBY genes, but previously it was not related to the 545 bp PRU (Skaletsky et al. 2003; Warburton et al. 2008).

As seen, the human HOR pattern of sequenced Y chromosome contains fewer copies than chimpanzees and is less symmetrically organized. The human ∼545 bp monomers (denoted $m01$ $m02$ $m03$) are similar to the chimpanzee ∼550 bp monomers (denoted $mc01$ $mc02$ $mc03$): divergence between the human 3mer HORs $m01$, $m02$, and $m03$ and the chimpanzee 3mer HORs is ∼4%, while the divergence between off–diagonal monomers (i.e., $m01$ vs. $mc02$, $m01$ vs. $mc03$,...) is ∼8%. Only a small subsection of ∼24 kb encompassing each human HOR copy is similar to the corresponding section encompassing each chimpanzee HOR copy (divergence less than 10%), while the remaining part of large spacings, of total length ∼2 Mb, strongly diverges between human and chimpanzee. This gives a substantial contribution to the overall human–chimpanzee divergence. Furthermore, the subsequences of ∼24 kb human sequence are scattered in various parts of chimpanzee Y chromosome.

3. Human ∼2385 bp Primary Repeat Unit and ∼7155 bp 3mer HOR Secondary Repeat Unit

The DAZ gene family, located in the AZFc region of Y chromosome, is organized into two clusters and contains a variable number of copies (Fernandes et al. 2006; Glaser et al. 1998; Saxena et al. 2000; Seboun et al. 1997). A ∼2.4 kb repeat unit in DAZ genes was reported by (Skaletsky et al. 2003; Warburton et al. 2008). Accordingly, the GRM peak at 2385 bp (Fig. 2b) is due to tandem repeats with ∼2.4 bp PRU in DAZ genes. Human DAZ repetitions are located in contig NT_011903.12 (positions 1346649 to 1361029, 1425263 to 1473290, 2977988 to 2997102, and 3050498 to 3086580), i.e., from position 25.3 to 27 Mb within the human Y chromosome.

Using GRM we classify the assembly of ∼2.4 kb monomers into five monomer families (consensus sequences in Supplementary Table 10). The average divergence between monomers of the same family is below 1%, while the average divergence between monomers from different families is ∼11%. The monomer family with highest frequency of appearance has consensus length 2385 bp, which determines the length of the 2385 bp GRM peak. This monomer family forms a highly homologous monomeric tandem repeat, which is present in DAZ2 and DAZ4 genes.

We find that the GRM peak at 7155 bp corresponds to 3mer HOR composed of three variants of ∼2.4 kb DAZ repeat monomers, denoted $m01$, $m02$, and $m03$ (the first three consensus sequences from Supplementary Table 10). Computing the GRM diagram of any of the 7155 bp copies we obtain two pronounced peaks, at ∼2.4 and ∼4.8 kb, revealing the 3mer character. We find that these 3mer HOR copies are present in all four DAZ1–DAZ4 genes. Human DAZ genes contain 12 DAZ HOR copies organized into four tandem arrays (DAZ1–DAZ4).

The ∼4757 bp peak in GRM diagram corresponds to the 2mer HOR copies arising from 3mer HOR by deletion of one monomer from the 7155 bp secondary 3mer HOR unit. In GRM diagram of the 4757 bp repeat copies, we obtain only one pronounced GRM peak, at ∼2.4 kb, showing the 2mer character of 4757 bp repeat copies. We find that such 2mer HOR copies are present in all four DAZ1–DAZ4 genes.

4. Chimpanzee ∼2383 bp Primary Repeat Unit and Absence of Tandem of Higher Order Repeats

The GRM peak at ∼2383 bp is due to tandem repeats with ∼2.4 bp repeat unit in DAZ genes in chimpanzee Y chromosome. Chimpanzee DAZ repetitions are located in contigs NW_001252917.1 (positions 1109191 to 1130961 and 1259092 to 1280862) and NW_001252922.1 (positions 997017 to 1028356 and 1070171 to 1099128) that is at chromosome positions from ∼3.2 to 3.4 Mb and from ∼11.2 to 11.3 Mb. Positions of the corresponding subsequences widely differ in human and chimpanzee chromosomes. Divergence between human and chimpanzee consensus sequences is ∼5%.

We find that the chimpanzee Y chromosome contains 3mer and 2mer HOR copies, similar to those for human Y chromosome, but with one pronounced distinction: chimpanzee DAZ genes contain four DAZ HOR copies, which are, unlike the case of human Y chromosome, not organized into tandem but into dispersed HOR copies. Therefore, there are no GRM peaks corresponding to HORs.

The presence of tandem of DAZ HOR copies in human and absence of such tandem in chimpanzee Y chromosome provides an interesting evolutionary distinction between human and chimpanzee Y chromosomes.

5. Human ∼3579 bp 715mer HOR Unit and 5 bp Primary Repeat Unit

The GRM peak at ∼3579 bp is due to a tandem of 28 repeat copies in NT_025975.2. These copies differ in lengths from 3544 to 3589 bp. The length 3579 bp has the highest frequency and is equal to consensus length. Other copy lengths appear due to deletion or insertion of 5 bp subsequences. Average divergence of copies with respect to consensus sequence is ∼1%. Due to differences in lengths of copies, the GRM peak at ∼3579 bp is broadened (Fig. 2b).

In the next step, we find a strong peak at the fragment length 5 bp in GRM diagram for the 3579 bp consensus sequence. A dominant key string for segmentation of the 3579 bp consensus sequence into 5 bp fragments is ATTCC, which is the consensus sequence of 5 bp primary repeat copies. Thus the 3579 bp repeat unit is a 715mer HOR based on ATTCC primary consensus repeat unit. Here 34% of primary repeat 5 bp copies are equal to consensus, 38% differ from consensus by one base, 21% by two, 6% by three and 1% by four bases.

This 3579 bp HOR corresponds to the previously reported 3584 bp HOR (Skaletsky et al. 2003).

6. Absence of Chimpanzee HOR Unit Corresponding to Human 3579 bp 715mer HOR Unit

In the Build 2.1 assembly for chimpanzee Y chromosome we find no analog of the human 3579 bp 715mer HOR unit.

7. Human ∼5607 bp 1123mer HOR Unit and 5 bp Primary Repeat Unit

The 5607 bp peak corresponds to a new HOR, with 5607 bp SRU (5 bp GGAAT PRU). The main contribution to this peak is from contig NT_113819.1. We identify a tandem of 11 copies, from position 496682 to 553881 (Supplementary Table 11) and determine the 5607 bp consensus sequence (Supplementary Table 12).

To investigate the structure of 5607 bp repeat unit, we compute the GRM diagram of its consensus sequence. Using 8 bp key string ensemble, we obtain the GRM diagram characterized by a set of GRM peaks at fragment lengths of 5 bp and its multiples (Supplementary Fig. 1a), revealing the underlying 5 bp PRU. However, the reciprocal distribution of GRM peaks shows deviation from the exponential distribution expected due to random mutations of fragments of multiple orders at KSA recognition sites. This deviation is due to the fact that the length of key strings in the ensemble is larger than the repeat unit. This is shown by computing the GRM diagram by using the 3 bp key string ensemble, shorter than the 5 bp PRU (Supplementary Fig. 1b). In that case the reciprocal distribution of GRM peaks corresponding to the 5607 bp consensus sequence indeed follows exponential distribution, as expected.

The 5607 bp HOR consensus unit consists of 1123 pentamer copies. Out of these copies, 353 are identical to GGAAT which is the primary repeat consensus. The mean divergence between 5 bp consensus GGAAT and pentamer copies that are not identical to consensus is ∼30%. Differences are mostly due to substitutions. There are only a few indels: two copies have 1–base insertion, one has 2–base insertion, ten have 1–base deletion and one has 2–base deletion.

8. Absence of Chimpanzee HOR Unit Corresponding to the Human 5607 bp HOR Unit

In the Build 2.1 assembly for chimpanzee Y chromosome we find no repeat unit corresponding to human 5607 bp HOR unit.

9. Chimpanzee 10853 bp Primary Repeat Unit and 64624 bp Secondary Repeat Unit

The GRM peak at 10853 bp is due to a tandem in NW_001252917.1 (eight copies), with repeat unit consensus length 10853 bp. The 10853 bp consensus sequence is given in Supplementary Table 15. The third copy in this tandem is distorted: truncated after the first 6399 bases and followed by a large insertion, so that the total length of truncated third copy and neighboring insertion amount to the combined length of 21218 bp. The structure of the eighth copy is distorted similarly as the third copy, leading again to a ∼21 kb combined length.

Distance between the corresponding bases in neighboring copies (except those involving the third copy) is ∼10853 bp, giving rise to the 10853 bp GRM peak.

Distance between the start of the 6399 bp subsection of the third copy and the start of the fourth copy is 21218 bp, giving rise to the 21218 bp GRM peak. Distance from the end of the second copy (which has no counterpart in the truncated third copy) to the end of the fourth copy is $10853 + 21218$ bp $= 32071$ bp, giving rise to the 32071 bp GRM peak.

The copies No. 1, 2, and 4–7 are identical up to 1%, while the copies No. 3 and 8 have similar truncation and additional insertion. Therefore, the copies No. 1–5 form a secondary repeat HOR copy of the approximate length $2 \times 10853 + 21218 + 2 \times 10853$ (precise value 64624 bp). The last three copies in tandem, No. 6–8, represent the first three copies belonging to the second 64624 bp HOR copy.

The insertion after the truncated third copy in chimpanzee tandem repeat with 10853 PRU 21218–6399 bp = 14819 bp is also present in the human Y chromosome as a tandem of two repeat units (divergence ∼ 4%) in contig NT_011903.12. Because these repetitive units are mutually reverse complement, GRM diagram for human chromosome Y does not show this peak.

C. Summary of Human–Chimpanzee Divergence Due to Repeats Based on Large Repeat Units

We determine approximately the number of bases which are different in repeat arrays of human and chimpanzee Y chromosome using a simple formula:

$$d = \sum_i d_i = \sum_i \Big(\min(l_{i,\mathrm{hum}}, l_{i,\mathrm{chimp}}) \cdot p_i + l_i \Big). \quad (1)$$

Here, $l_{i,\mathrm{hum}}$ and $l_{i,\mathrm{chimp}}$ are sums of lengths over all copies of the $i$ths human and chimpanzee repeat unit, respectively; $\min(l_{i,\mathrm{hum}}, l_{i,\mathrm{chimp}})$ is the smaller of two lengths $l_{i,\mathrm{hum}}$ and $l_{i,\mathrm{chimp}}$; $l_i = |l_{i,\mathrm{hum}} - l_{i,\mathrm{chimp}}|$; and p_i is divergence between human and chimpanzee repeat unit i. In this way, we include contributions to human–chimpanzee divergence both from substitutions and indels.

For example, in the case of alphoid HOR in Y chromosome (repeat No. 1 from Tables I, II, III) we have: $l_{1,\mathrm{hum}} = 3048138$ bp, $l_{1,\mathrm{chimp}} = 1042459$ bp, $l_1 = 2005679$ bp, $p_1 = 0.20$, giving $d_1 = 2.214.171$ bp (Fig. 7). With respect to the sequence of larger alphoid HOR, of the length $l_{1,\mathrm{hum}}$, this corresponds to an approximate divergence $100 \cdot d_1/l_{1,\mathrm{hum}} = 72.6\%$.

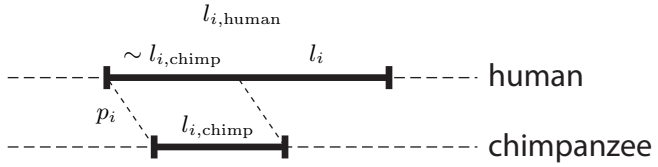

FIG. 7 Schematic presentation of applying the formula for calculation of human-chimpanzee divergence for the case of a large repeat unit (major alphoid HOR)

Summing over all repeats ($i = 1, 2, \ldots$) from Tables I, and III, we obtain a summary number of different bases between human and chimpanzee large repeats: $d \sim 3.4$ Mb (3378539 bp). The corresponding divergence with respect to all repeats from Tables I, II, and III is:

$$\mathrm{div(rep)} = 100 \cdot \frac{d}{L}, \quad (2)$$

where the summary length of all repeats from Tables I, II, and III is $L = 4848892$ bp.

Thus, we obtain divergence with respect to repeat sequences included in Tables I, II, and III:

$$\mathrm{div(rep)} \approx 70\%. \quad (3)$$

If we smear out divergence over the whole Build sequence of length $L_{\mathrm{as}} = 25$ Mb, we obtain the overall divergence with respect to assembly length:

$$\mathrm{div(Build)} = 100 \cdot \frac{d}{L_{\mathrm{as}}} \quad (4)$$

$$\mathrm{div(Build)} \approx 14\%. \quad (5)$$

This estimate of overall divergence due to repeats based on large repeat units should be additionally increased due to overall estimates of approximately 1–2% divergence for nonrepeat sequences.

Both the human and the chimpanzee Y chromosome sequences are still incomplete; in human chromosome ∼25 Mb out of total length of ∼59 Mb was sequenced. Thus, a greater contiguity at several genomic regions is desired to reach more precise conclusions regarding human–chimpanzee divergence. However, the main body of results will probably stand, because, in general, nonsequenced gaps are rich in repeat structures. It should be noted that a whole–genome comparison of chimpanzee and human revealed an increased divergence in the terminal 10 Mb of the corresponding chromosomes, consistent with general association between increased divergence rates and location near the chromosome ends (Mikkelsen et al. 2005; Pollard et al. 2006a). In general, and in accordance with Gibbs et al. (2007), it can be expected that unsequenced regions of repeat elements, that are difficult to align, might for the whole Y chromosome somewhat increase the presently estimated divergence of 14% for the sequenced part. Definitive studies of genome evolution will require high–quality finished sequences (Mikkelsen et al. 2005).

An interesting question is how much the observed sizeable divergence can be generalized to the whole genome. In this sense, we have started a systematic study of human–chimpanzee divergence due to large repeats in other chromosomes.

We see a tendency that large repeat units in humans are on average larger and copy numbers greater than those in chimpanzees. This is in accordance with previous observation that microsatellites in humans are on average longer than those in chimpanzees (Vowles and Amos 2006).

We identify large repeat units which contribute substantially to divergence between humans and chimpanzees. Our results indicate that alphoid HOR and most of characteristic tandem repeats with large repeat units (some present only in human and not in chimpanzee Y chromosome, or some vice versa) have been created after the human–chimpanzee separation, while only a smaller number of tandems with large repeat units (present both in human and in chimpanzee Y chromosome at low mutual divergence) originate from a common ancestor that predated the human–chimpanzee separation. This is in accordance with previous observations in some other chromosomes that alpha satellite subsets found in great apes and humans are in general not located on their corresponding homologous chromosomes (Jorgensen et al. 1992; Warburton et al. 1996); for example, the alpha satellite subset on human chromosome 5 is a member of SF 1, while the homologous chimpanzee chromosome belongs to SF 2 (Haaf and Willard 1997, 1998). It was pointed out that this implies that the human–chimpanzee sequence divergence has not arisen from a common ancestral repeat, but instead represents initial

amplification and homogenization of distinct repeats on homologous chromosomes (nonorthologous evolution).

Haaf and Willard (1997) discussed the propositions for homogenization of alpha satellites. Homogenization processes appear to proceed in localized, short–range fashion that leads to formation of large domains of sequence identity (Durfy and Willard 1989; Tyler–Smith and Brown 1987; Warburton and Willard 1990). Genomic turnover mechanisms (molecular drive; (Dover 1982, 1986)) must be at work that spread and homogenize individual variant repeat units throughout arrays and throughout populations (Haaf et al. 1995). However, the mechanisms by which this concerted evolution occurs seem unclear, although several genomic turnover mechanisms such as unequal crossing over between repeats of sister chromatids (Smith 1976), sequence conversion (Baltimore 1981), sequence transposition (Calos and Miller 1980), translocation exchange (Krystal et al. 1981), and disproportionate replication (Hourcade et al. 1973; Lohe and Brutlag 1987; Spradling 1981) have been observed to be active in certain genomes.

Previous FISH studies support the conclusion that the localization of SF 3 alpha satellite is substantially conserved, while alpha satellite sequences belonging to families 1 and 2 are not shared by the corresponding chimpanzee homologs (Archidiacono et al. 1995; D'Aiuto et al. 1993). Here we find that, although the SF 4 which is composed of M1 alpha satellite monomers constituting human and chimpanzee alphoid HORs in Y chromosomes is conserved, both the alpha satellite monomers in human and chimpanzee HORs and the HOR lengths are widely different.

It was pointed out that it is not known whether evolutionary important mutations predominantly occurred in regulatory sequences or coding regions (Carroll 2003; King and Wilson 1975; McConkey 2002; McConkey et al. 2000; Olson and Varki 2003). Preliminary data suggested that gene expression patterns of human brain might have evolved rapidly (Caceres et al. 2003; Dorus et al. 2004; Enard et al. 2002; Uddin et al. 2004).

Comparative genomic analyzes strongly indicated that the marked phenotypic differences between humans and chimpanzees are likely due more to changes in gene regulations then to modifications of genes themselves (King and Wilson 1975; Pollard et al. 2006a, b; Popesco et al. 2006; Prabhakar et al. 2006)(King and Wilson 1975; Pollard et al. 2006a,b; Popesco et al. 2006; Prabhakar et al 2006). The gene regulatory evolution hypothesis proposes that the striking differences between humans and chimpanzees are due to gene expression: the change of pattern and timing of turning genes on and off.

Pollard et al. (2006b) identified ∼100 bp short genomic regions that are highly conserved in vertebrates, but show significantly accelerated substitution rates on human lineage relative to chimpanzee (Pollard et al. 2006a,b). Many of these Human Accelerated Regions (HARs), characterized by dense clusters of nucleotide substitutions, are associated, in particular, with the nervous system, reproductive system, and immune system.

Detailed studies have indicated that forces other than selection for random mutations that increase fitness in specific functional elements may be at play in strongly accelerated regions (Pollard et al. 2006a). There is a possibility that changes in the accelerated regions result from a combination of multiple evolutionary processes, perhaps including biased gene conversion and a selection–based process (Pollard et al. 2006a).

Here, we find another type of accelerated regions: for some repeat arrays we find dramatic evolutionary acceleration of repeat pattern, from monomeric arrays in chimpanzee to HOR organization of repeat arrays in human Y chromosome, i.e., the rapid onset of unequal crossing over in human lineage. Such region of accelerated evolution of HOR pattern will be referred to as human accelerated HOR region (HAHOR).

The hallmark of evolutionary shift of function is sudden change in a region of genome that previously has been conserved (Pollard et al. 2006b). The function of sets of genomic regulatory sequences has been previously compared to electronic microprocessing: they process the information contained in a set of regulatory elements into the corresponding pattern of gene expression. It was noted that one of basic ways how the regulatory genomic features are related to evolutionary processes is the recruitment of existing regulatory pathways into newly evolving context (Gierer 1998; Pires–da Silva and Sommer 2003; Tautz 2000). These processes follow the rules of nonlinear interactions. These, in turn, allow for sudden or very fast changes resulting from the accumulation of rapidly succeeding small steps with self–enhancing features. Furthermore, mechanisms of bifurcation and de novo pattern formation may lead, for instance, to strikingly different developments in parts of an initially near–uniform area. Thus, in general, small causes can result in big effects (Gierer 2004). Finally we note a possibility that accelerated large repeat units and HAHORs could have a functional role of new categories of long–range regulatory elements (Noonan and McCallion 2010).

IV. CONCLUSION

In this study, we identify and analyze tandem repeats, HORs and regularly dispersed repeats in chimpanzee and human. For the first time we report a dozen new large repeats in chimpanzee and several new large repeats in human genome. Comparing the corresponding repeats based on large repeat units in human and chimpanzee we find substantial contribution to the human–chimpanzee divergence from these repeats, approximately 70% divergence with respect to repeat arrays based on large repeat units. Smearing out these differences in large repeats over the whole sequenced assemblies, human Build 37.1 and chimpanzee Build 2.1, i.e., by neglecting divergence between other segments of genome sequences, we obtain an overall human–chimpanzee divergence between

sequenced assemblies of approximately 14%. This numerical estimate far exceeds the available earlier numerical estimates for human–chimpanzee divergence.

Our results are in accordance with recent publication by Hughes et al. (2010) where it was shown by overall comparison that the human and chimpanzee MSYs differ radically.

We explicitly identify, analyze, and compare a dozen of large repeats which give a substantial contribution to human–chimpanzee divergence.

We find in humans several HAHORs on human lineage relative to chimpanzee, containing HOR structures, in particular the alphoid HORs, the ∼2.4 kb DAZ repetitions and the ∼15.8 kb repetitions. On the other hand, in chimpanzee genome we find a chimpanzee–accelerated HOR region (CAHOR) based on ∼550 bp PRU.

While the HARs discovered previously (Pollard 2009; Pollard et al. 2006a,b; Popesco et al. 2006; Prabhakar et al 2006) were HARs characterized by short dense clusters of nucleotide substitutions, the HAHORs found in this work are characterized by higher–order organization extended over larger genomic stretches.

Our results show explicitly that large repeat units and HORs provide substantial contribution to the human–chimpanzee divergence.

V. GRM ANALYSIS

GRM analysis was performed using novel GRM code, which is available upon request.

Acknowledgments

Authors are grateful to Martin Kreitman and Chris Tyler–Smith for very helpful comments and suggestions.

References

Alexandrov IA, Kazakov A, Tumeneva I, Shepelev V, Yurov Y (2001) Alpha–satellite DNA of primates: old and new families. Chromosoma 110:253–266

Ali S, Hasnian SE (2003) Genomics of the human Y chromosome1 association with male infertility. Gene 321:25–37

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

Archidiacono N, Antonacci R, Marzella R, Finelli P, Lonoce A, Rocchi M (1995) Comparative mapping of human alphoid sequences in great apes using fluorescence in situ hybridization. Genomics 25:477–484

Bailey JA, Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity and diesease. Nat Rev Genet 7:552–564

Baldini A, Miller DA, Miller OJ, Ryder OA, Mitchell AR (1991) A chimpanzee–derived chromosome–specific alpha satellite DNA sequence conserved between chimpanzee and human. Chromosoma 100:156–161

Baltimore D (1981) Gene conversion: some implications for immunoglobulin genes. Cell 24:592–594

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391–1394

Britten RJ (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. Proc Natl Acad Sci USA 99:13633–13635

Britten RJ, Rowen L, Williams J, Cameron RA (2003) Majority of divergence between closely related DNA samples is due to indels. Proc Natl Acad Sci USA 100:4661–4665

Caceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C (2003) Elevated gene expression levels distinguish human from non–human primate brains. Proc Natl Acad Sci USA 10:13030–13035

Calos MP, Miller JH (1980) Transposable elements. Cell 20:579–595

Carroll SB (2003) Genetics and the making of Homo sapiens. Nature 422:849–857

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444–456

Cheng Z, Ventura M, She XW, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S, Rocchi M, Eichler EE (2005) A genome–wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88–93

Choo KHA (1997) The Centromere. Oxford University Press, Oxford

Cooper KF, Fisher RB, Tyler–Smith C (1993a) The major centromeric array of alphoid satellite DNA on the human Y chromosome is non–palindromic. Hum Mol Genet 2:1267–1270

Cooper KF, Fisher RB, Tyler–Smith C (1993b) Structure of the sequences adjacent to the centromeric alphoid satellite DNA array on the human Y chromosome. J Mol Biol 230:787–799

D'Aiuto L, Antonacci R, Marzella R, Archidiacono N, Rocchi M (1993) Cloning an comparative mapping of a human chromosome 4–specific alpha satellite DNA sequence. Genomics 18:230–235

de Knijff P (2006) The human Y chromosome is not dead (yet). Heredity 97:377–378

Dorit RL, Akashi H, Gilbert W (1995) Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 268:1183–1185

Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, Mahowald M, Wyckoff GJ, Malcom CM, Lahn BT (2004) Accelerated evolution of nervous system genes in the origin of Homo sapiens. Cell 119(7):1027–1040

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–116

Dover G (1986) Molecular drive in multigene families: how biological novelties arise, spread and are asasimilated. Trends Genet 2:159–165

Durfy SJ, Willard HF (1989) Patterns of intra– and interarray sequence variation in alpha satellite from the human X chromosome: evidence for short–range homogenization of tandemly repeated DNA sequences. Genomics 5:810–821

Durfy SJ, Willard HF (1990) Concerted evolution of primate alpha satellite DNA. Evidence for ancestral sequence shared by gorilla and human X chromosome alpha satellite. J Mol Biol 216:555–566

Ebersberger I, Metzler D, Schwarz C, Paabo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet 70:1490–1497

Ebersberger I, Galgoczy P, Taudien S, Taenzer S, Platzer M, von Haeseler A (2007) Mapping human ancestry. Mol Biol Evol 24:2266–2276

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869–872

Fernandes AT, Fernandes S, Goncalves R, Sa R, Costa P, Rosa A, Ferras C, Sousa M, Brehm A, Barros A (2006) DAZ gene copies: evidence of Y chromosome evolution. Mol Hum Reprod 12:519–523

Fujiyama A, Watanabe H, Toyoda A, Taylor TD, Itoh T, Tsai SF, Park HS, Yaspo ML, Lehrach H, Chen Z et al (2002) Construction and analysis of a human–chimpanzee comparative clone–map. Science 295:131–134

Gelfand Y, Rodriguez A, Benson G (2007) TRDBthe tandem repeats database. Nucleic Acids Res 35:D80–D87

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK et al (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234

Gierer A (1998) Networks of gene regulation, neural development and the evolution of general capabilities, such as human empathy. Z Naturforsch C 53(7–8):716–722

Gierer A (2004) Human brain evolution, theories of innovation, and lessons from the history of technology. J Biosci 29:235

Glaser B, Yen PH, Schempp W (1998) Fibre–fluorescence in situ hybridization unravels apparently seven DAZ genes or pseudogenes clustered within a Y chromosome region frequently deleted in azoospermic males. Chromosome Res l:481–486

Glusman G, Sosinsky A, Ben–Asher E, Avidan N, Sonkin D, Bahar A, Rosenthal A, Clifton S, Roe B, Ferraz C, Demaille J, Lancet D (2000) Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. Genomics 63:227–245

Graves JA (1995) The origin and function of the mammalian Y chromosome and Y–borne genes – an evolving understanding. Bioessays 17:311–320

Haaf T, Willard HF (1992) Organization, polymorphism, and molecular cytogenetics of chromosome–specific alpha–satellite DNA from the centromere of chromosome 2. Genomics 13:122–128

Haaf T, Willard HF (1997) Chromosome specific alpha satellite DNA from the centromere of chimpanzee chromosome 4. Chromosoma 106:226–232

Haaf T, Willard HF (1998) Orangutan alpha satellite monomers are closely related to the human consensus sequence. Mamm Genome 9:440–447

Haaf T, Matera AG, Wienberg J, Ward DC (1995) Presence and abundance of CENP–B box sequences in great ape subsets of primate–specific α–satellite DNA. J Mol Evol 41:487–491

Henikoff S (2002) Near the edge of a chromosomes black hole. Trends Genet 18:165–167

Hourcade D, Dressler D, Wolfson J (1973) The amplification of ribosomal RNA genes involves a rolling circle intermediate. Proc Natl Acad Sci USA 70:2926–2930

Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, Rozen S, Wilson RK, Page DC (2005) Conservation of Y–linked genes during human evolution revealed by comparative sequencing in chimpanzee. Nature 437:101–104

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C et al (2010) Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature 463:536–539

Jobling MA, Tyler–Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598–612

Jorgensen AL, Bostock CJ, Bak AL (1986) Chromosome–specific subfamilies within human alphoid repetitive DNA. J Mol Biol 187:185–196

Jorgensen AL, Laursen HB, Jones C, Bak AL (1992) Evolutionarily different alphoid repeat DNA on homologous chromosomes in human and chimpanzee. Proc Natl Acad Sci USA 89:3310–3314

Kehrer–Sawatzki H, Cooper DN (2007) Structural divergence between the human and chimpanzee genomes. Hum Genet 120:759–778

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309:1850–1854

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188:107–116

Kirsch S, Muench C, Jiang Z, Cheng Z, Chen L, Batz C, Eichler EE, Schempp W (2008) Evolutionary dynamics of segmental duplications from human Y–chromosomal euchromatin/heterochromatin transition regions. Genome Res 18:1030–1042

Krystal M, D'Eustachio P, Ruddle FH, Arnheim N (1981) Human nucleolus organizers on non–homologous chromosomes can share the same ribosomal gene variants. Proc Natl Acad Sci USA 78:5744–5748

Kuroda–Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Silber S, Oates R, Rozen S et al (2001) The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet 29:279–286

Kuroki Y, Toyoda A, Noguchi H, Taylor TD, Itoh T, Kim DS, Choi SH, Kim IC, Choi HH, Kim YS et al (2006) Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. Nature Genet 38:158–167

Lahn BT, Page DC (1999) Four evolutionary strata on the San Francisco human X chromosome. Science 286:964–967

Laursen HB, Jorgensen AL, Jones C, Bak AL (1992) Higher rate of evolution of X chromosome alpha repeat DNA in human than in great apes. EMBO J 11:2367–2372

Liu GE, Alkan C, Jiang L, Zhao S, Eichler EE (2009) Comparative analysis of Alu repeats in primate genomes. Genome Res 19:876–885

Lohe AE, Brutlag DL (1987) Adjacent satellite DNA segments in Drosophila. Structure of junctions. J Mol Biol 194:171–179

Maio JJ (1971) DNA strand reassociation and polyribonucleotide binding in the African green monkey Cercopithecus aethiops. J Mol Biol 56:579–595

Manuelidis L, Wu JC (1978) Homology between human and simian repeated DNA. Nature 276:92–94

Marshall Graves JA (2006) Sex chromosome specialization and degeneration in mammals. Cell 124:901–914

McConkey EA (2002) A project on gene expression during primate development is urgently needed. Trends Genet 18:446

McConkey EH, Fouts R, Goodman M, Nelson D, Penny D, Ruvolo M, Sikela J, Stewart CB, Varki A, Wise S (2000) Proposal for a human genome evolution project. Mol Phylogenet Evol 15:1–4

Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang SP, Enard W, Hellmann I, Lindblad–Toth K, Altheide TK et al (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87

Mitchell AR, Gosden JR, Miller DA (1985) A cloned sequence, p82H, of the alphoid repeated DNA family found at the centromeres of all human chromosomes. Chromosoma 92:369–377

Muller HJ (1914) A gene for the fourth chromosome of Drosophila. J Exp Zool 17:325–336

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M, McGrath SD, Rocchi M, Eichler EE (2005) A genome–wide survey of structural variation between human and chimpanzee. Genome Res 15:1344–1356

Noonan JP, McCallion AS (2010) Genomics of long–range regulatory elements. Annu Rev Genomics Hum Genet 11:1–23

Nusbaum C, Mikkelsen TS, Zody MC, Asakawa S, Taudien S, Garber M, Kodira CD, Schueler MG, Shimizu A, Whittaker CA et al (2006) DNA sequence and analysis of human chromosome 8. Nature 439:331–335

Oakey R, Tyler–Smith C (1990) Y chromosome haplotyping suggests that most European and Asian men are descended from one or two males. Genomics 7:325–330

Ohno S (1967) Sex chromosomes and sex linked genes. Springer, New York

Olson MV, Varki A (2003) Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet 4:20–28

Paar V, Pavin N, Rosandić M, Glunčić M, Basar I, Pezer R, Durajlija Žinić S (2005) ColorHORnovel graphical algorithm for fast scan of alpha satellite higher–order repeats and HOR annotation for GenBank sequence of human genome. Bioinformatics 21:846–852

Paar V, Basar I, Rosandić M, Glunčić M (2007) Consensus higher order repeats and frequency of string distributions in human genome. Curr Genomics 8:93–111

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108

Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2:100–109

Perry GH, Tito RY, Verelli BC (2007) The evolutionary history of human and chimpanzee Y–chromosome gene loss. Mol Biol Evol 24:853–859

Perry GH, Yang F, Marques–Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler–Smith C et al (2008) Copy number variation and evolution in humans and chimpanzees. Genome Res 18:1698–1710

Pires–da Silva A, Sommer RJ (2003) The evolution of signalling pathways in animal development. Nat Rev Genet 4:39–49

Pollard KS (2009) What makes us human? Sci Am 300:32–37

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R et al (2006a) Forces shaping the fastest evolving regions in the human genome. PLoS Genet 2:1599–1611

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A et al (2006b) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443:167–172

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

Prabhakar S, Noonan JP, Paabo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. Science 314:786

Romanova LY, Deriagin GV, Mashkova TD, Tumeneva IG, Mushegian AR, Kisselev LL, Alexandrov IA (1996) Evidence for selection of alpha satellite DNA: the central role of CENP–B/pJα binding region. J Mol Biol 261:334–340

Rosandić M, Paar V, Basar I (2003a) 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

Rosandić M, Paar V, Glunčić M, Basar I, Pavin N (2003b) Key–string algorithm – Novel approach to computational analysis of repetitive sequences in human centromeric DNA. Croat Med J 44:386–406

Rosandić M, Paar V, Basar I, Glunčić M, Pavin N, Pilaš I (2006) CENP–B box and pJα sequence distribution in human alpha satellite higher–order repeats (HOR). Chromosome Res 14:735–753

Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP et al (2005) The DNA sequence of the human X chromosome. Nature 434:325–337

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423:873–876

Rudd MK, Willard HF (2004) Analysis of the centromeric regions of the human genome assembly. Trends Genet 20:529–533

Rudd MK, Schueller MG, Willard HF (2003) Sequence organization and functional annotation of human centromeres. Cold Spring Harb Symp Quant Biol 68:141–149

Rudd MK, Wray GA, Willard HF (2006) The evolutionary dynamics of alpha–satellite. Genome Res 16:88–96

Saxena R, Brown LG, Hawkins T, Alagappan RK, Skaletsky H, Reeve MP, Reijo R, Rozen S, Dinulos MB, Disteche CM, Page DC (1996) The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. Nature Genet 14:292–299

Saxena R, De Vries JWA, Repping S, Alagappan RK, Skaletsky H, Brown LG, Ma P, Chen E, Hoovers JMN, Page DC (2000) Four genes in two clusters found in the AZFc region of the human Y chromosome. Genomics 67:256–267

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF (2001) Genomic and genetic definition of a functional

human centromere. Science 294:109–115
Seboun E, Barbaux S, Bourgeron T, Nishi S, Agulnik A, Agasshira M, Nikkawa N, Bishop C, Fellous M, McElreavey K et al (1997) Gene sequence, localization, and evolutionary conservation of DAZL1 A, a candidate male sterility gene. Genomics 41:227–235
Sibley CG, Ahlquist JE (1987) DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. J Mol Evol 26:99–121
Silber SJ, Repping S (2002) Transmission of male infertility to future generations: lessons from the Y chromosome. Hum Reprod Upd 8:217–229
Skaletsky H, Kuroda–Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T et al (2003) The male–specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423:825–837
Smith GP (1976) Evolution of repeated DNA sequences by unequal crossing over. Science 191:528–535
Sonnhammer ELL, Durbin R (1995) A dot–matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1–GC10
Spradling AC (1981) The organization and amplification of two chromosomal domains containing Drosophila chorion genes. Cell 27:193–201
Tautz D (2000) Evolution of transcriptional regulation. Curr Opin Genet Dev 10:575–579
Tyler–Smith C (1985) Structure of repeated sequences in the centromeric region of the human Y chromosome. Development 101:93–100
Tyler–Smith C, Brown WRA (1987) Structure of the major block of alphoid satellite DNA on the human Y chromosome. J Mol Biol 195:457–470
Uddin M, Wildman DE, Liu G, Xu W, Johnson RM, Hof PR, Kapatos G, Grossman LI, Goodman M (2004) Sister grouping of chimpanzees and humans as revealed by genome–wide phylogenetic analysis of brain gene expression profiles. Proc Natl Acad Sci USA 101:2957–2962
Varki A, Altheide TK (2005) Comparing human and chimpanzee genomes: searching for needles in a haystack. Genome Res 15:1746–1758
Varki A, Geschwind DH, Eichler EE (2008) Explaining human uniqueness: genome interactions with environment, behavior and culture. Nature Genet 9:749–763
Vowles EJ, Amos W (2006) Quantifying ascertainment bias and species–species length differences in human and chimpanzee microsatellites using genome sequences. Mol Biol Evol 23:598–607
Warburton PE, Willard HF (1990) Genomic analysis of sequence variation in tandemly repeated DNA. Evidence for localized homogeneous sequence domains within arrays of α–satellite DNA. J Mol Biol 216:3–16
Warburton PE, Willard HF (1996) Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In: Jackson M, Strachan T, Dover G (eds) Human Genome Evolution. BIOS Scientific, Oxford, pp 121–145
Warburton PE, Haaf T, Gosden J, Lawson D, Willard HF (1996) Characterization of a chromosome–specific chimpanzee alpha satellite subset: evolutionary relationship to subsets on human chromosomes. Genomics 33:220–228
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
Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, Kuroki Y, Noguchi H, BenKahla A, Lehrach H, Sudbrak R et al (2004) DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature 429:382–388
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
Webster MT, Smith NG, Ellegren H (2003) Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol Biol Evol 20:278–286
Willard HF (1985) Chromosome–specific organization of human alpha satellite DNA. Am J Hum Genet 37:524–532
Willard HF (1991) Evolution of alpha satellite. Curr Opin Genet Dev 1:509–514
Willard HF, Waye JS (1987) Hierarchical order in chromosome–specific human alpha satellite DNA. Trends Genet 3:192–198
Wolfe J, Darling SM, Erickson RP, Craig IW, Buckle VJ, Rigby PWJ, Willard HF, Goodfellow PN (1985) Isolation and characterisation of an alphoid centromeric repeat family from the human Y chromosome. J Mol Biol 182:477–485