2016

Clusters of Alpha Satellite on Human Chromosome 21 Are Dispersed Far onto the Short Arm and Lack Ancient Layers

William Ziccardi
Loyola University Chicago

Chongjian Zhao
Loyola University Chicago

Valery Shepelev
Russian Academy of Sciences

Lev Uralsky
Russian Academy of Sciences

Ivan Alexandrov
Russian Academy of Medical Sciences

Recommended Citation

Ziccardi, William; Zhao, Chongjian; Shepelev, Valery; Uralsky, Lev; Alexandrov, Ivan; Andreeva, Tatyana; Rogaev, Evgeny; Bun, Christopher; Miller, Emily; Putonti, Catherine; and Doering, Jeffrey. Clusters of Alpha Satellite on Human Chromosome 21 Are Dispersed Far onto the Short Arm and Lack Ancient Layers. Chromosome Research, 24, : 421-436, 2016. Retrieved from Loyola eCommons, Bioinformatics Faculty Publications, http://dx.doi.org/10.1007/s10577-016-9530-z

Authors
William Ziccardi, Chongjian Zhao, Valery Shepelev, Lev Uralsky, Ivan Alexandrov, Tatyana Andreeva, Evgeny Rogaev, Christopher Bun, Emily Miller, Catherine Putonti, and Jeffrey Doering

ABSTRACT

Human alpha satellite (AS) sequence domains that currently function as centromeres are typically flanked by layers of evolutionarily older AS that presumably represent the remnants of earlier primate centromeres. Studies on several human chromosomes reveal that these older AS arrays are arranged in an age gradient, with the oldest arrays furthest from the functional centromere and arrays progressively closer to the centromere being progressively younger. The organization of AS on human chromosome 21 (HC21) has not been well-characterized. We have used newly-available HC21 sequence data and an HC21p YAC map to determine the size, organization, and location of the AS arrays, and compared them to AS arrays found on other chromosomes. We find that the majority of the HC21 AS sequences are present on the p-arm of the chromosome and are organized into at least five distinct isolated clusters which are distributed over a larger distance from the functional centromere than that typically seen for AS on other chromosomes. Using both phylogenetic and L1 element age estimations, we found that all of the HC21 AS clusters outside the functional centromere are of a similar relatively recent evolutionary origin. HC21 contains none of the ancient AS layers associated with early primate evolution which is present on other chromosomes, possibly due to the fact that the p-arm of HC21 and the other acrocentric chromosomes underwent substantial reorganization about 20 million years ago.

List of Abbreviations

AS – alpha satellite

HCX – human chromosome X

HC21 – human chromosome 21

HC21p – human chromosome 21 short arm

HOR – higher order repeat

SD – segmental duplications

Clusters of Alpha Satellite on Human Chromosome 21 Are Dispersed Far onto the Short Arm and Lack Ancient Layers

William Ziccardi[1], Chongjian Zhao[1], Valery Shepelev[2,3,4], Lev Uralsky[2,4], Ivan Alexandrov[5], Tatyana Andreeva[3] , Evgeny Rogaev[3,4,6,7] Christopher Bun[8], Emily Miller[1], Catherine Putonti[1,8,9] and Jeffrey Doering[1] *

Affiliations:
[1] Department of Biology, Loyola University Chicago, 1032 W. Sheridan Road, Chicago, IL 60660 USA
[2] Institute of Molecular Genetics, Russian Academy of Sciences, Kurchatov sq. 2, Moscow 123182, Russia
[3] Department of Genomics and Human Genetics, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119991, Russia
[4] Center for Brain Neurobiology and Neurogenetics, Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia
[5] Research Center of Mental Health, Russian Academy of Medical Sciences, Zagorodnoe sh. 2, Moscow 113152, Russia
[6] Department of Psychiatry, Brudnick Neuropsychiatric Research Institute, University of Massachusetts Medical School, Worcester, Massachusetts 01604, USA
[7] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 119234, Russia
[8] Department of Computer Science, Loyola University Chicago. 820 N. Michigan Avenue, Chicago, IL 60611 USA
[9]Bioinformatics Program, Loyola University Chicago, 1032 W. Sheridan Road, Chicago, IL 60660 USA

***Corresponding author:** Jeffrey Doering; Tel: (773) 508 3627; FAX: (773) 508 3646;
E-mail: jdoerin@luc.edu

Running Title: Alpha satellite on human chromosome 21

Key words: alpha satellite, chromosome 21, chromosome evolution, centromere, acrocentric chromosome, chromosome mapping

INTRODUCTION

The centromeric regions of all human chromosomes contain AS, composed of tandemly arranged monomeric units approximately 171 bp long (Willard & Waye 1987). The actual functional centromere consists of monomers arranged in multimeric higher-order repeat (HOR) units which themselves repeat to form an array of homogenous HORs which can be megabases in length (Rudd & Willard 2004). The pericentromeric regions of chromosomes contain multiple layers of highly heterogeneous "monomeric" AS repeat arrays that lack any HOR, are composed of monomers or dimers of AS basic units, and are often found interspersed with other sequences (Rudd & Willard 2004, Shepelev et al. 2009). The HOR arrays are more recently evolved than the monomeric AS clusters which are more similar to the AS of lower primates and presumably represent the remnants of earlier primate centromeres (Alexandrov et. al. 2001, Schueler et al. 2001, Schueler et al. 2005, Rudd et al. 2006, Shepelev et al. 2009).

Functional HOR arrays may be classified into three "new" suprachromosomal families (SFs 1, 2 and 3), each residing on a number of chromosomes. The older monomeric AS is divided into the two large groups SF5 and SF4. SF5 is evolutionarily younger and immediately ancestral to the new families, and on most chromosomes it directly flanks the functional HOR arrays. The SF4 group contains all the older layers of monomeric AS (Alexandrov et al. 2001). Recently it has been subdivided into a number of SFs, most of which have not yet received formal names pending finalization of a new classification system. They are called dead AS layers and are color-coded (Table 1, Shepelev et al. 2009). These layers can be discriminated by a number of structural features and ultimately through full-scale phylogenetic analysis (Shepelev et al. 2009). In this paper, we refer to the old SF4 as the umbrella group named "SF4+", which includes the yellow layer and all the older layers of Shepelev et al. (2009). The new SFs 1-3, SF5 and SF4+ groups are all composed of their own classes of monomers (Alexandrov et al. 2001) recognizable by the PERCON program (Kazakov et al. 2003, Shepelev et al. 2015). Annotation of these groups in Human Genome assembly has recently been published (Shepelev et al. 2015).

Monomeric AS clusters are found both directly adjacent to centromeric HOR arrays and as separate, distinct clusters surrounded by non-AS sequences. Both types of monomeric clusters contain a lower frequency of transposed elements compared to the rest of the genome but do possess a higher frequency of such elements compared to HOR arrays (Mashkova et al. 1996, Schueler et al. 2001, Kazakov et al. 2003). This is consistent with monomeric AS clusters being older than HOR arrays but having a more recent history than the rest of the genome.

The sequence homogeneity of the new HOR clusters in the functional centromeres is thought to be maintained by recombination mechanisms like unequal crossover (Smith 1976) and gene conversion, while older clusters are not maintained by such events, and have accumulated mutations, transposon insertions and other rearrangements, eventually drifting into a highly divergent state (Alexandrov et al. 2001, Shepelev et al. 2009).

This model for AS evolution is supported by studies of the detailed organization of AS clusters on chromosomes X, 8 and 17 (Schueler et al. 2001, Schueler et al. 2005, Rudd et al. 2006, Shepelev et al. 2009). In all cases the HOR-containing functional centromere is surrounded by multiple monomeric AS clusters arrayed roughly symmetrically on both chromosome arms. The monomeric clusters are all more ancient evolutionarily than the centromeric AS and each monomeric cluster has a separate evolutionary history. The ages of the clusters were estimated using L1 insertions (Schueler et al., 2001, Kazakov et al. 2003, Shepelev et. al 2009), which revealed an age gradient. Clusters furthest from the functional centromere contain the oldest L1 variants, dating to early primate evolution, while clusters progressively closer to the centromere are progressively younger, with the functional centromere HOR dating to the great ape divergence or later. The more distant clusters show a greater degree of sequence heterogeneity than those found closer to the functional centromere. These data support the hypothesis that monomeric AS clusters represent arrays that once were homogeneous and served as the functional centromere but have since degenerated into heterogeneous clusters after the centromere function was superseded by a newly-generated HOR cluster. Subsequent rounds of replacement and drift would explain multiple layers of monomeric AS clusters on the same chromosome.

The organization of AS clusters on HC21 is much less well-characterized. The functional centromere region (α21-I) contains an 11-mer SF2 HOR structure of highly homogenous AS DNA (D21Z1), while numerous families of evolutionarily older monomeric AS are interspersed with a variety of other sequences over a long expanse of HC21p in the α21-II region (Carnahan et al. 1993, Trowell et al. 1993, Ikeno et al. 1994). Prior to the current work, little was known about the detailed organization of AS clusters in α21-II and their evolutionary relationships to each other and D21Z1.

In this work we used newly-available HC21 sequence data and an HC21p YAC map (Wang et al. 1999) to determine the size, organization, and location of the α21-II AS clusters and then compared them to AS clusters found on other chromosomes (Shepelev et al. 2009). Contrary to the situation on other chromosomes the α21-II AS clusters have a marked asymmetry in their distribution around the centromere. Several large AS clusters are located unusually far onto the short arm and interspersed with non-AS sequences. AS in these clusters at least partially consists of intra- and inter-chromosomal segmental duplications (SDs), and at least one of the SF4+ clusters has an HOR organization. Using both phylogenetic and L1 element age estimations, we found that all of the HC21 monomeric clusters are of a similar relatively recent evolutionary origin and none of the ancient monomeric layers associated with early primate evolution (Shepelev et al. 2009) are present on HC21. This organization of HC21p AS differs substantially from the more developed layered structure seen on chromosomes X, 8, and 17 (Shepelev et al. 2009).

MATERIALS & METHODS

AS Sequence Probes

The sequences of all HC21p AS probes used in this study can be found in GenBank (accession numbers in Suppl. Table 2). The pTRA clones were described by Choo et al. (1988) and Vissel and Choo (1991), and p11-4 and the pN clones were described by Ikeno et al. (1994). All clones were generously provided to us by these investigators. CEN2 and CEN3 clones were provided by M. Burmeister and were subcloned from HC21-specific phage lambda clones.

YAC Mapping of AS Sequences

All AS sequences were mapped to the HC21p YAC contig map (Wang et al. 1999) at stringencies high enough to minimize cross-hybridization between different AS sequences. YAC clones were digested with *Eco*RI to release the DNA inserts from the yeast vector. The digested DNAs were resolved by pulsed field gel electrophoresis, Southern blotted, and hybridized with a given ^{32}P-labelled AS sequence probe using standard methods (Doering et al. 1982). The pattern of YACs hybridizing to the same sequence permitted the AS sequences to be ordered. Since some of the YAC clones may be chimeras (Wang et al. 1999), only AS-containing fragments which were shown to be on HC21p by hybrid cell mapping were used to create the map. The 2E4, 3G8, and 2C9 YAC DNAs were also digested with restriction enzymes that do not cut in AS monomers, blotted and hybridized with specific probes to estimate the size of a given AS cluster.

Identifying BAC/Cosmids Containing α21-II Sequences

Bacterial artificial chromosomes (BACs) and cosmid clones containing AS sequences were identified by scanning the NCBI nucleotide collection of human sequences (taxon 9606) via Megablast searches using previously and newly-sequenced α21-II clone sequences. HC21 BACs found to have high matches (>90%) were then mapped to the appropriate α21-II cluster according to the previous YAC mapping (Figure 1).

Analysis of BAC/Cosmid Clones for Repetitive Sequences

HC21 BACs were scanned for AS and other repetitive sequences using the Repeat Masker web resource (http://www.repeatmasker.org/). Repeat Masker was also used to determine the precise location of AS sequences within clones from other human chromosomes (Shepelev et al. 2009). Dot Plot analyses (http://www.vivo.colostate.edu/molkit/dnadot) determined the presence or absence of any HOR structure within the BAC using a window size of 171 bp and mismatch limits ranging from 2 to 15%.

AS Suprachromosomal Family Classification

Classification of AS sequences into SFs was done using the PERCON program described in Kazakov et al. (2003) and Shepelev et al. (2015) and as shown in the hg38 assembly UCSC Genome Browser custom track (Shepelev et al. 2015). PERCON classifies AS monomers into 12 major types (Alexandrov et al. 2001) that define 5 AS SFs, of which SF1, 2 and 3 represent the live functional centromeres, and SFs 4+ and 5 represent the dead centromeric layers. The normal order of AS layers in a chromosome are: a new SF (1, 2 or 3) – SF5 – SF4+ (Alexandrov et al. 2001). In HC21, the functional centromere is SF2 (Table 1). Examples of PERCON monomeric maps are shown in Suppl. Figure 1. SF4+ is an umbrella group that includes a number of colored layers and respective monomeric classes which are not directly recognized by PERCON (Table 1). The tools we used to figure out the identity of these layers (rs analysis and 171/172 analysis) are described below.

Classification of SF4+ Dead Layers

In Shepelev et al. (2009), SF4+ was further differentiated into an age gradient of color-coded layers, of which the youngest is yellow and the next distal and older layer is yellow-striped. The latter two, together with the yet younger SF5 (blue layer) were termed old AS. The older olive-green, red and grey layers were termed ancient AS (Table 1).

We noted previously (Figure S1 in Shepelev et al. 2009) that ancient SF4+ AS could be distinguished from the old one by analyzing the statistics of rs (relative alignment score), a parameter which is calculated by the PERCON program and appears on its printout (Suppl. Figure 1). In this work, we used rs statistics gathered by PERCON to check for the presence of ancient monomers in the BAC/cosmid HC21 clones and chromosome assembly contigs.

To calibrate the detection method we used the known colored domains from the HCX pericentromeric region (Figure 5), the complete monomer map of which was published previously (Shepelev et al. 2009). Respective pieces of the HCX hg38 assembly were subjected to PERCON analysis and the rs statistics for SF4+ monomers were gathered (Suppl. Table 6). To avoid short monomer fragments which could spoil the analysis, we filtered PERCON produced monomers for the length of 90 bp or more. The old AS arrays processed in this way had 1-2% of rs<0.62 monomers, while even the youngest of the ancient layers (olive-green) had 55% of them. In yet older ancient layers this proportion increased to 90%, and in the complete HCX, 78% of all ancient monomers had rs<0.62. Similar results (Suppl. Table 6) were obtained with chromosomes 8 and 17 where the colored layers have also been mapped (Shepelev et al. 2009). We concluded that the ancient monomers could be easily discovered by this method and their real number could be up to twice as much as the number of rs<0.62 monomers.

We also noted previously (Supplementary text S1 in Shepelev et al. 2009) that in the yellow layer all monomers have a deletion in position 21 of the 172 bp AS monomer and therefore are 171 bp long, while the monomers of the yellow-striped layer have a ~1:3 ratio of 171 bp (pos. 21-) to 172 bp (pos. 21+) monomers. In order to use this feature for layer classification, we have supplied the PERCON program with a new module which detects the deletion in AS monomers in the vicinity of position 21 (positions 16-26 of a 172 bp monomer). The program examines the alignment of a given monomer to ALPHA-ALL consensus (see details in Shepelev et al. (2015)). If it does not include positions 16-26, the monomer remains unclassed ("u" in PERCON map), if it has a deletion in that region, the monomer is classed as "171 bp" and if it does not, the monomer is classed as "172 bp" (see the PERCON maps in Suppl. Figure 1). Given the absence of ancient AS in HC21 sequences, the SF4+ regions formed exclusively of 171 bp monomers were interpreted as yellow, and the ones containing many 172 bp monomers were interpreted as yellow-striped.

Phylogenetic Studies

Phylogenetic trees for the AS sequence-containing clones located on HC21 were constructed. The AS repeats identified by the MiIP program (Bun et al. 2012) were aligned using ClustalW2 in SeaView and a phylogenetic tree was derived using both the Neighbor Joining (NJ) and the phyML (Maximum Likelihood) method with 100 bootstrap replicates (Gouy et al. 2010). The African Green Monkey (AGM) sequence was selected as the root for each tree. The alignment was then used to derive consensus sequences ($\geq$ 60%) for each of the clones within the HC21 pericentromeric region. NJ and ML trees with 100 bootstrap replicates were also generated for this consensus tree using SeaView (Gouy et al. 2010**)**. Phylogenetic trees were visualized using PhyloWidget (Jordan & Piel 2008**)** and NJ plot (Perrière & Gouy 1996**)**.

Monomer mixing tests were done as previously described (Shepelev et al. 2009) to determine which SF4+ layers on HCX mix phylogenetically with the AS sequences on HC21. The HCX minimum evolution tree was constructed by MEGA5 using default parameters and contained 2319 monomers prepared as described in Shepelev et al. (2009). They represent all AS dead layers. The monomers from various HC21 AS-containing clones were added to this tree and their mixing with certain branches analyzed. For all HC21 clones and contigs listed in Suppl. Table 2 the results of the mixing tests and the results of PERCON analysis were the same.

Deep Sequencing Reads and WAV-17 test

DNA from mouse/human somatic cell hybrid WAV-17 was purchased from Coriell Cell Repositories (Catalog ID: GM08854) and used for whole-genome next-generation sequencing. This cell line contains, as its only human material, 1-5 copies per cell of HC21 which lacks major deletions or rearrangements (Lyle et al. 2007). Illumina pair-end sequencing libraries were constructed from 1 μg of WAV-17 DNA using the TruSeq DNA Sample Prep v2 Kit according to the manufacturer's protocol. The library was size selected to obtain an average DNA fragments size of 400-450 bp and sequenced on the Illumina HiSeq2000 platform as paired 101 bp reads. Genomic DNA of a female human individual was sequenced in the same manner and used as a control where needed.

We used three samples of ~50 million 101 bp reads each from which AS sequences were extracted by RepeatMasker. Each sequenced clone or contig subjected to the WAV-17 test (Suppl. Table 2) was aligned by BLAST (align two or more sequences option) to ~25000 AS reads extracted from the 50 million raw reads of WAV-17 DNA. For each sequence we counted the number of full-length (101 bp) 100% matches.

AS in WAV-17 HC21 Somatic Cell Hybrid

To determine whether there is ancient AS in the unknown portions of HC21p we applied rs analysis (described above) to WAV-17 AS deep sequencing read datasets. First, we re-evaluated this method for use with deep-sequencing reads of WAV-17 instead of whole AS monomers. The detection method was calibrated with truncated versions of the HCXp AS region (Figure 4, Suppl. Tables 6 and 7) that had only the old AS (ancient truncated), old plus olive-green (red and gray truncated) or the complete HCXp. These sequences were broken into 100 bp pieces to mimic the deep sequencing reads, subjected to PERCON analysis and the rs statistics for SF4+ monomers were gathered. Since PERCON starts the monomers at a fixed point (Kazakov et al. 2003, Shepelev et al. 2015), some 100 bp reads were divided into two monomers and some rather short monomeric fragments were generated. To avoid these short fragments which could spoil the analysis, we filtered PERCON-produced monomeric pieces for the length of 90-101 bp and then performed rs analysis.

We then proceeded to analyze the sequencing reads from the human control genomic sample and the WAV-17 cell line. AS was extracted from half a million human genomic reads and 4.5 million WAV-17 genomic reads by RepeatMasker and subjected to rs analysis.

RESULTS

YAC Mapping of AS Sequences

Eleven different AS sequences previously shown to be on chromosome 21 (names listed in Suppl. Table 2) were mapped to the HC21p YAC contig constructed by Wang et al. (1999). The results are shown in Figure 1. While some of the AS sequences are located proximal to the D21Z1 cluster, many lie at substantial distances from the centromere on the p-arm. Several of the AS probes detect sequences at two locations along the chromosome. In addition to the already-identified AS sequence adjacent to D21Z1 on the q-arm (Bozovsky et al. 2004), at least 5 additional clusters appear to be on the p-arm (Figure 1). The locations of YACs relative to each other in the contig were adjusted so as to reflect their content of AS sequences as well as the markers used to originally construct the map (Wang et al. 1999).

Cluster locations were confirmed and their sizes estimated by pulsed field gel mapping of specific YACs using restriction enzymes unlikely to cut AS sequences. The digests were hybridized with an AS probe known to hybridize to the specific α21-II region being sized. Mp2 could not be directly sized since the 250 kb band seen corresponds to undigested YAC 4E9 insert and not a fragment of Mp2. YAC 4E9 hybridized with probes for both Mp2 and Mp3, indicating these two clusters are closely linked and may form a single larger AS cluster.

Identifying BAC/Cosmid Clones Containing α21-II Sequences

All available α21-II plasmid clones were used to search the NCBI nucleotide database for BACs and cosmids that had greater than 90% sequence identity to any given α21-II clone. Three new HC21p BACs containing AS sequences were discovered: CT476838, FP236243, and CU638690. Fosmid FP565424, that overlaps the known HC21q contig, was also identified.

These clones were subjected to pair-wise BLAST comparisons with the entire collection of AS plasmid clones (Suppl. Table 2). We also analyzed three other clones previously shown to contain sequences from the α21-II region: cosmid AF105153 (Mashkova et al. 1996) and BAC AF254982, on HC21p, and BAC AP001464 on HC21q.

CT476838 and FP236243 have a significant overlap and assembling the sequences into a single contig results in a sequence 189097 nucleotides long with an overlap of 74461 nucleotides (Figure 2) that has sequence identity greater than 99%.

The HC21q clones FP565424 and AP001464 have ~1 kb overlap with 99% sequence identity, and they were merged into an Mq1 contig (Suppl. Figure 1) of 154959 bp, which contains 99% and 100% matches to CEN2-4 and CEN2-6, respectively (Suppl. Table 2). Mq1 contains 94% matches to D21Z1 on its one end and overlaps the main HC21q contig GL000151 on its other end. Sequence comparisons (Suppl. Figure 1) indicate that the D21Z1-like sequences in Mq1 are not the major centromeric cluster, which consists of 11-mer HOR units, but rather a variant region with 13-mer HOR units. The latter are composed of the same types of monomers, but differ both in sequence (94% identity) and structure (2 extra monomers at the end of the HOR unit, see Suppl. Figure 1).

Characterization of HC21 BACs/Cosmids

Using the information from both the YAC mapping (Figure 1) and the pair-wise sequence comparisons of the BAC/cosmid clones to the AS plasmid clone collection (Suppl. Table 2), the HC21 clones were mapped to the appropriate α21-II clusters. The CT476838/FP236243 contig was confidently placed in the Mp3 cluster based on the strong identity matches to both pTRA-2 and pN23 (Suppl. Table 2), which were located in the Mp3 region (Figure 1). CU638690 was mapped to the Mp5 cluster because of its strong match with pN31, although this placement was not made with absolute confidence as no other clones known to map to the Mp5 region showed a strong match to the BAC. AF254982 and AF105153 both map to the Mp1 cluster since

AF254982 contains a fragment of the TPTE pseudogene and AF105153 contains the ABM-C78 marker (Figure 2). These markers place the two BACs in the Mp1 cluster with AF105153 distal to AF254982 (Figures 1 and 2). There is no overlap between the two clones, suggesting that they are not part of a single uninterrupted contig (Figure 2).

All of these HC21p and q clones were analyzed by RepeatMasker to identify all repetitive elements included in them (Figure 2, Suppl. Table 3). In addition to AS sequences, the two major types of repetitive elements found within the clones are L1 insertions and satellite III (SatIII).

Dot plot analysis shows that only the Mp3 contig contains an HOR (Figure 3). This HOR cluster starts at a position approximately 52 kb into the contig sequence and has no lengthy period of transition between the HOR region and the monomeric cluster that precedes it. The HOR cluster extends 76.7 kb then ends abruptly, degenerating over a span of less than 3 kb into a monomeric AS organization that extends for another 75 kb to the end of the contig. The spacing of the diagonal lines in the dot plot indicates an HOR of approximately 3.9 kb, and the jagged appearance of these lines shows that the HOR structure of the cluster is degenerate with an imperfect match between HOR units and/or some variation in HOR size (Figure 3). Detailed analysis of the sequence shows that sequence identity between the HOR units is about 99%, but in many HORs some monomers are missing and some are duplicated (Suppl. Figure 1). The pTRA-2 AS sequence found in Mp3 (Suppl. Table 2) had previously been suggested to be part of a higher-order repeat structure of approximately 3.9 kb (Vissel & Choo 1991).

Much of the presently-available HC21p sequence contains SDs (Lyle et al. 2007), and we found that many of the HC21p AS sequences are also duplicated elsewhere in the genome. A short region of AF105153 is duplicated on chromosome 16, while small portions of CU638690 are also found on either chromosome 7 or 16. The HOR region of Mp3 is duplicated on a BAC of unknown location (AC144610), which is clearly distinct from HC21p. This locus is most likely on one or more of the other acrocentric chromosome short arms, which are known to contain sequences in the Mp3 HOR (Vissel & Choo 1991). There are also intrachromosomal duplications, as, for example, AF254982 (Mp1) which contains a 7 kb AS region that is 96% identical to CU638690 (Mp5). It should be noted that sequence similarities that involve satellite DNA are often not reflected in the SD track of the UCSC Genome Browser. However, wherever HC21 AS loci are flanked by available non-satellite sequence, the AS sequence is embedded in clusters of SDs.

Verifying Presence of AS Sequences on HC21

Since many AS sequences are present on other chromosomes as well as on HC21p, we verified the presence of Mp1-5 sequences on HC21 using deep sequencing of the WAV-17 somatic cell hybrid (Materials and Methods). We first verified the presence of AS plasmid probe sequences in the WAV-17 AS sequencing datasets. All probes except CEN3-1 had a number of full-length (101 bp) perfect matches in WAV-17 sequences (Suppl. Table 2). The CEN3-1 full-length matches observed in WAV-17 were all less than 95% identical. While both CEN3-1 and pTRA-7 do not have significant matches on the BAC/cosmid clones (Suppl. Table 2), pTRA-7 was confirmed to be on HC21 by the WAV-17 test. We conclude that CEN3-1 is absent from WAV-17 and is probably located on some other chromosome (e.g. chromosomes 5 and 19 to which it has 97% matches). However, a deletion in WAV-17 or allelic variation in HC21 are also possible. Hybridization of CEN3-1 to HC21 YACs (Figure 1, Suppl. Table 1) likely indicates that Mp4 contains AS sequences that are closely related but distinct from CEN3-1. Next we verified the presence in WAV-17 of AS sequences in the BAC/cosmid clones which define the Mq1, Mp1, Mp3 and Mp5 loci using the same deep sequencing method. All clones had a large number of perfect full-length matches (Suppl. Table 2) covering their AS regions without large visible gaps. This confirms the presence of all these AS clones, or nearly identical sequences on HC21.

L1 Insertions in HC21 Clones

L1 insertions are useful in estimating the age of AS clusters (Schueler et al. 2001, Kazakov et al. 2003, Shepelev et. al 2009), and so all L1 insertions in the previously described HC21 clones were characterized using the RepeatMasker program (Table 2, Suppl. Table 5). Overall, many more L1 insertions (72.4% of all L1s found) are found free from any association with AS clusters compared to those L1s embedded within (21%) or adjacent to (6.6%) AS sequence. Full length L1 inserts are also rare; only 5 full length L1 insertions (6.6% of all L1s identified) are found in all the BAC/cosmid clones studied. Four of the full length L1 inserts are embedded in or adjacent to AS sequences. L1 insertions were sorted into two groups by evolutionary age: modern and ancient (Smit et al. 1995). Modern L1s were defined as great ape-specific L1 insertions with very recent evolutionary origins, including L1PA3, L1PA2 and L1PHs (also known as L1PA1). Those L1 insertions with an origin older than L1PA3 were defined as ancient. L1 insertions not directly associated with any AS sequence are overwhelmingly ancient in origin (94.5% of L1 insertions in those regions) and include subfamilies that can be found in all mammalian species (Table 2). In contrast, modern L1 insertions are much more prevalent adjacent to (80%) and embedded within (93.8%) AS clusters (Table 2). Those modern L1 insertions adjacent to or embedded within AS sequences are primarily L1PA3, indicating a uniformly young age for most AS clusters on HC21p. The only AS-embedded ancient element L1PA4 was found in the beginning of the Mp3 contig (Suppl.

Figure 1, Suppl. Table 5), which suggests that this portion of AS might be older than the rest of AS sequences on HC21p.

Identifying AS Families on HC21

The AS sequences on HC21 were analyzed by the PERCON program and revisions of it (Kazakov et al. 2003, Shepelev et al. 2015, Materials and Methods) which identify the AS monomer types and SFs (Table1) present. D1 and D2 (SF2) monomers are found in D21Z1, R1 and R2 (SF5) and M1+ (SF4+) regions are in the Mq1 contig and only SF4+ monomers are in the Mp1, Mp3 and Mp5 contigs (Suppl. Table 2). The Mp2 and Mp4 regions have no contigs, so we analyzed the probes which mapped in these regions. CEN3-1, CEN2-4 and CEN2-6 all have R1 and R2 monomers and thus belong to SF5 (Table 1). However, CEN2-4 and CEN2-6 both have perfect matches in the Mq1 contig (Suppl. Table 2) and thus belong there. CEN3-1 was not confirmed on HC21 by WAV-17 reads and probably is located on some other chromosome. Therefore, hybridization with these probes in the Mp2 and Mp4 loci most likely just indicates the presence of some SF5 (blue) AS there.

SF4+ regions are further differentiated into an age gradient of color-coded layers, of which the youngest is yellow and the next distal and older layer is yellow-striped. The latter two, together with the yet younger SF5 (blue layer) are termed old AS. Yet older olive-green, red and grey layers are termed ancient AS (Table1). Using rs (relative similarity score) statistics gathered by PERCON (Materials and Methods), we next determined which SF4+ monomers are present in the Mp1, Mp3 and Mp5 loci, particularly looking for monomers with rs<0.62, which indicate ancient AS (Materials and Methods). A total of 1989 SF4+ monomers with length 90 bp or more were scored. Only in Mp1 clones did the number of rs<0.62 monomers rise above the 2% background, and only slightly (Suppl. Table 6). Only 4% of monomers in AF105153 and

3% of monomers in AF254982 have rs<0.62, which means that only about 5 such monomers above the expected background are present in these two clones (a total of 455 monomers in both clones). As the real number of ancient monomers could be about twice as much (Materials and Methods), an estimate of only 10 such monomers could be present among the total 1989 SF4+ monomers scored. Also, no rs<0.62 monomers were found in the short plasmid clones used in mapping HC21p (Suppl. Table 2).

Since significant numbers of ancient monomers were not found in HC21 clones, the SF4+ monomers identified there by PERCON could only belong to yellow or yellow-striped layers (Table 1). We used a previously-described AS monomer length (171bp versus 172 bp) polymorphism (Shepelev et al. 2009), to distinguish sequences in these two layers from each other (see Materials and Methods). Control typing of the known regions in HCX revealed the predominance of 172 bp monomers in the ancient layers and only one 172 bp monomer in the yellow regions (341 monomers total), while 70% and 90% of 172 bp monomers were found in the Xp and Xq yellow-striped regions, respectively. Thus, given the absence of ancient monomers on HC21, the presence of 172 bp monomers identifies the yellow-striped layer, and the real number of yellow-striped monomers can be 10-30% more than the number of 172 bp monomers (Materials and Methods). Examination of PERCON monomer maps of Mp1, Mp3, Mp5 and Mq1 allowed us to map yellow and yellow-striped regions in all 4 loci, (Figure 5, Suppl. Figure 1). Of particular interest is that the Mp3 HOR domain is located on the border between yellow-striped and yellow sequence in the contig. The HOR is on average 23 monomers long and has 19 monomers of 171 bp and 4 monomers of 172 bp (Suppl. Figure 1). As the true number of yellow-striped monomers can be up to one-third larger, we estimate that about 5-6 yellow-striped monomers are present in the HOR and the rest are yellow. This suggests that the HOR may have originated from amplification of a segment of the yellow/yellow-striped border.

The composition of the AS layers revealed in HC21 is consistent with our L1 dating results. The most common L1 in HC21 AS is L1PA3 (Table 2, Suppl. Table 5) which is characteristic of the yellow and blue layers (Shepelev et al. 2009). The only L1PA4 element present is in the yellow-striped array. The absence of any older L1 elements in the AS is consistent with the virtual absence of ancient AS layers on HC21.

Previous work (Shepelev et al. 2009) identified a number of different AS sequence domains in the long and short arm pericentromeric regions of chromosomes 8, 17 and X (Figure 5). These domains are arranged primarily in a symmetrical fashion on both sides of the functional centromere on any particular chromosome and are largely shared and arranged similarly on the non-homologous chromosomes. There also exists an age gradient on these chromosomes with evolutionarily older domains tending to be located further away from the centromere than newer domains. In contrast, HC21 has none of the ancient domains (gray, red and olive-green), and the domains that are present are not symmetrically arranged on both chromosome arms (Figure 5).

Phylogenetic Analysis of HC21 AS Monomers

To confirm our results on the sequence relationships of HC21 AS monomers we used formal phylogenetic analysis. Initial studies using Neighbor-Joining or Maximum-Likelihood trees indicated that while the AS sequences in all the monomeric clusters are, as expected, evolutionarily older than that in the HOR D21Z1, there are no large differences in the ages of the various AS monomer layers. This is consistent with the finding that there are no or almost no ancient AS layers on HC21 (Figure 5).

We also performed selective mixing tests, as described in Shepelev et al. (2009) to place the various HC21 AS monomers onto the phylogenetic tree of HCX monomers which was studied previously and contained all known AS layers (Figure 5). An example of the test for Mp3 monomers (Suppl. Figure 2) confirms the monomeric assignments made by PERCON analysis (Suppl. Table 2), with yellow and yellow-stripped Mp3 monomers grouping with their analogous SF4+ layers on HCX. The mixing tests did not result in any cases where significant numbers of HC21 monomers grouped with the ancient SF4+ layers, further confirming the younger age of HC21 AS.

HC21 AS in hg38 Human Genome Assembly

When the current work was already finished, the new human genome assembly (hg38; GCA_000001405.15) was released, which adds a number of AS sequences to HC21. Annotation of AS SFs in the hg38 assembly was recently published (Shepelev et al. 2015) and is available as a UCSC Genome Browser custom track. The only AS in the previous hg37 assembly was a fragment of the Mp1 locus (AF254982) on the p-arm side and a part of the Mq1 locus (AP001464) on the q-arm side. A number of HC21p clones have been added in the new assembly and the centromeric gap has been filled with so-called “reference models”, which are somewhat arbitrary representations of AS HOR domains. Reference models are not real DNA sequences like traditional GenBank contigs, but instead are collections of all WGS reads that match a certain HOR put into a contig by the stochastic approach of using a generative Markov process, which is not expected to generate the true long-range linear order across the entire array (Miga et al. 2014). They can however be very helpful in mapping the AS sequencing reads to the human genome assembly.

Due to the complex pattern of identities in centromeric sequences of the acrocentric chromosomes 13, 14, 21 and 22, the mapping protocol used in the new assembly was apparently unable to determine which reference model belonged to which chromosome and what were the precise locations of the AS sequences on the chromosomes. Thus, all the HOR domains, which are present on at least one of these chromosomes, were put together in a single block, and this block was placed in the former centromeric gap on each chromosome. The same block of 13 reference models arranged in the same order appears on all four chromosomes, but individual reference models have different names on every chromosome. We evaluated these reference models to see if they appear in WAV-17 and if they add anything new to our map.

The SF and colored layer composition according to PERCON, the results of the WAV-17 test and the results of BLAST with the HC21 clones described in the current work are presented in Suppl. Table 4. It appears that at least five of the reference models are not likely to belong to HC21, as they are not confirmed by WAV-17 reads. The largest of these is GJ212135 which represents the major SF2 HOR common to chromosomes 14 and 22. This HOR is well known to be absent from chromosomes 13 and 21 (Alexandrov et al. 2001). Two SF4+ (yellow-striped) models (GJ212125 and GJ212126) have only a few hits relative to their large size, so they most likely also belong to another chromosome (Suppl. Table 4). Of the 6 models that are confidently confirmed on HC21 by the WAV-17 test, the largest is GJ212154 which represents the major SF2 HOR shared by chromosomes 13 and 21 (D21Z1). Two (GJ212124 and GJ212131) represent sequences in the Mp3 contig, one (GJ212118) represents Mp1 sequences, and one contains an SF5 HOR of ~3.7 kb (GJ212128), a part of which is identical (99%) to the pTRA-7 probe. The remaining yellow model (GJ212132) is confidently confirmed but cannot be currently anchored to any known HC21 sequences. These analyses suggest first that Mp3 sequences are artificially duplicated in the assembly, being represented by both their actual genomic clones and the reference models, and the location of an Mp3 HOR reference model adjacent to the functional centromere is likely to be incorrect (Supp. Table 4).   Second, there is a previously-unidentified SF5 (blue) HOR (GJ212128) most likely located at Mp2/Mp3 and/or Mp1, as indicated by the pTRA-7 hybridization data in Figure 1. Finally there is likely to be at least one additional SF4+(yellow) HOR (GJ212132) at an unknown location.

The p-arm region of the assembly shows Mp5, part of Mp3 (FP236243, but not CT476838) and part of Mp1 (AF254982, but not AF105153) in an inverted orientation relative to our map. Such an inversion is not consistent with previous mapping results (Wang et al. 1999, Brun et al. 2003). Given the extraneous “reference models” in the new genome assembly these p-arm AS regions are not likely to be as far from the centromere as this assembly shows. One additional AS contig (AC079801), located at 5.3 Mb on the map, is distal to Mp5 and appears to be a segmental duplication of Mp5. This can be termed an Mp6 locus, parts of which along with Mp5 have strong sequence similarities to HC22 (see the Segmental Duplications track in UCSC Genome browser) and to Mp1 (not reflected in the track). Our WAV-17 sequencing data indicate that not all parts of the AC079801 contig are present in WAV-17, suggesting possible allelic variation at this locus or a deletion in WAV-17.

AS in WAV-17 HC21 Cell Hybrid

While we did not find any large pieces of AS older than yellow-striped on HC21, the map of HC21p is incomplete and may contain older AS sequences in the unmapped regions. To investigate this we looked for ancient AS in WAV-17 deep sequencing reads by evaluating the percentage of monomers with rs<0.62 among all SF4+ (M1+) monomers, as an indicator of ancient layers (Table 1, Materials and Methods). The fully-characterized HCX regions were broken into 100 bp pieces to mimic the deep sequencing reads and used as controls (Materials and Methods). HCXp and HCXq regions containing only old (Table 1) AS sequences revealed a background of reads with rs<0.62 of 1% , addition of the youngest ancient layer (olive-green) resulted in an increase of this fraction to 22% of total SF4+ monomers and addition of yet older ancient layers produced a further increase to 41% (Figure 4, Suppl. Table 7).

As the number of real ancient monomers in HCX was known from previous work (Shepelev et al. 2009), we calculated based on the data in Suppl. Table 7 that 38% of olive-green reads and 64% of total ancient reads in the complete HCX had rs<0.62. We conclude that reads of ancient AS sequences can be easily discovered by this method and their average real number is roughly twice the number of reads with rs<0.62.

We then analyzed AS sequences in reads from total human DNA and the WAV-17 cell line. In HC21, only 2% of SF4+ reads have rs<0.62 (Figure 4, Suppl. Table 7). Since this is only slightly above the 1% background, HC21 must have very low levels of ancient AS. In stark contrast, the total genomic sample of SF4+ monomers has 16% of reads with rs<0.62, meaning about 30% of real SF4+ monomers are ancient (Figure 4). Further work shows that whole pericentromeric regions of human chromosomes 8, 17 and X processed as 100 bp reads have 72%, 70% and 41% of SF4+ reads with rs<0.62, respectively (Suppl. Table 7). This means that the vast majority of SF4+ AS on chromosomes 8, 17 and X is ancient, which is consistent with the data in Shepelev et al. (2009). Thus, these three chromosomes have substantially more than the average genomic amount of ancient AS, which suggests that some other chromosomes may have very little of it or lack it entirely. HC21 is apparently an example of such a chromosome which carries only relatively young AS. Our preliminary analysis of whole hg38 assemblies of the human acrocentric chromosomes (Suppl. Table 8) indicates that they all have very little ancient AS and have only yellow and yellow-striped AS as their oldest SF4+. Hence their centromeres may have all been formed at about the same time around 23 million years ago (Table 1, Shepelev et al. 2009).

DISCUSSION

Evolutionarily old AS sequences were previously found on HC21p (Carnahan et al. 1993) which were thought to comprise a region (α21-II) consisting of uninterrupted monomeric AS sequences (Ikeno et al. 1994). Little information on the detailed organization or evolutionary relationships of these sequences existed. We have developed a map showing that AS on HC21p has a more complicated structure than previously proposed, consisting of at least five distinct AS clusters (Mp1-5) that are large in size (25 – 189 kb each) and extend over a distance of more than 5 Mb from D21Z1. These clusters are not continuous but interspersed with L1 insertions, satellite III arrays, and low copy number sequences (Figure 2). The monomeric AS clusters are substantially further away from D21Z1 than previous estimates had indicated (Trowell et al. 1993). While the α21-II region had been assumed to consist only of monomeric AS, the current work directly identifies a low copy number SF4+ HOR array in the Mp3 cluster and provides indirect evidence that few other SF5 and SF4+ HORs exist elsewhere on HC21. Such small scale secondary amplifications within SF5 and SF4+ dead layers were also reported in other chromosomes (Alexandrov et al. 2001, Rosandić et al. 2006, Hayden et al. 2013, Shepelev et al. 2015).

The ages of the various AS clusters on HC21 were estimated in this work using the L1 inserts found in those clusters, phylogenetic analyses of the AS monomers that comprise them, and detailed AS subfamily sequence analysis. The three methods agree that all α21-II sequences are evolutionarily young and most are of similar age, 16 – 23 million years. None of the ancient monomeric clusters associated with early primate evolution (Shepelev et al. 2009) are present on HC21. Deep sequencing of HC21 further confirms the absence of any ancient AS clusters on HC21, even in as-yet unmapped regions of the chromosome (Figure 4, Suppl. Table 7). Thus, the HC21 centromere contains a much shorter succession of chronologically-layered arrays of AS monomer clusters than is seen around the centromeres of other chromosomes (Shepelev et al. 2009). This suggests that the HC21 pericentromere may have an evolutionary history different from that of such regions found on chromosomes 8, 17 and X.

HC21 AS Cluster Organization Compared to Other Human Chromosomes

A map of the AS clusters found on HC21 was constructed (Figure 5) and compared to similar maps of the AS clusters on chromosomes X, 8, and 17 (Shepelev et al. 2009). The latest assembly (hg38) of HC21p is, for the most part, consistent with this map (Suppl. Table 4), but misses a number of clones placed on HC21p in this work. Even without a more detailed knowledge of the organization and sequence of the AS clusters of HC21, clear differences in the size and distribution of AS clusters on HC21 compared to those found on the other chromosomes are readily apparent.

HC21 lacks the symmetry in AS clusters seen on the other chromosomes. Chromosomes X, 8, and 17 all possess similar amounts of monomeric AS sequence on their short and long arms. HC21q, however, possesses only a small amount of monomeric AS sequence, the Mq1 cluster, while HC21p contains all the clusters that comprise the α21-II region (Figure 5). This lack of symmetry may be explained by the fact that the short arms of acrocentric chromosomes can frequently engage in non-homologous exchanges (Choo 1990), but such exchanges do not occur between the long arms of acrocentric chromosomes. In non-acrocentric chromosomes such exchanges may occur only between pericentromeric regions and facilitate the spread of SDs only in these regions of non-homologous chromosomes. In contrast, in acrocentric chromosomes, this process is not limited to pericentromeres, but spreads far into the short arms. This may be due to the presence of rDNA tandem repeats and the abundance of heterochromatic classical satellites in these locations. The acrocentric chromosome exchanges may also be responsible for the fact that the HC21p AS clusters are part of a system of intra- and inter-chromosomal SDs.

The furthest any of the AS arrays on chromosomes 8, X, and 17 are from their functional centromeres is 500 kb. The *closest* AS monomeric array on HC21p is more than 600 kb from the centromere. Again, non-homologous exchanges between the acrocentric short arms could explain this finding.

HC21p possesses more AS sequence (approximately 550 kb) collected in generally larger arrays (25 – 189 kb) across the entire short arm of the chromosome than is seen on the short or long arms of the other chromosomes analyzed (Figure 5). Also, the Mp3 HOR AS cluster is of large size (76 kb) and found at some distance from the functional centromere, a trait not shared by any other studied chromosome. These observations might be explained by the relatively recent evolutionary age of HC21p AS arrays, which has not provided time for their organization to deteriorate.

Many of the AS monomeric sequence families studied here on HC21 are also found on more than one of the other acrocentric chromosomes (Vissel & Choo 1991), where they share similar organizations (Trowell et al. 1993). Thus the HC21p map (Figure 5) may be similar to that for other acrocentric chromosomes.

Comparison of AS Phylogenies Between HC21 and Other Chromosomes

Monomeric AS clusters comparable in age to those on HC21p are also found on other chromosomes (Shepelev et al. 2009, Figure 5), although these chromosomes also contain a number of much older clusters not found on HC21p. This is demonstrated by both phylogenetic analysis (Figure 4, Suppl. Figure 2) and L1 insertion ages (Table 2).

The absence of monomeric AS clusters on HC21 older than about 20 million years (Table 1, Figure 5, Shepelev et al. 2009) could be explained by a more recent evolutionary history for the HC21 centromere compared to the centromeres of other chromosomes. While HC21q is thought to have been formed some time between the divergence of the New World and Old World monkeys from the human lineage (Stanyon et al. 2008), the p-arm has a much more recent history. The rDNA clusters were involved in major rearrangements resulting in the beginning of their presence on the acrocentric p-arms between the divergence of orangutan and gibbon (Miller 1977). Other repetitive sequence families start appearing on the acrocentric p- arms at about the same time (Cardone et al. 2004, Jarmuz et al. 2007), suggesting that these genomic regions were substantially reorganized some 20 million years ago. Any older AS clusters could have been lost as a result of this reorganization. A new centromere may have been created at the same time by repositioning or remodeling. Alternatively, HC21 may just always have a quicker pace of DNA loss in dead pericentromeric AS regions, a feature perhaps shared by other acrocentric chromosomes.

Our analyses of hg38 assemblies of the other human acrocentric chromosomes (Suppl. Table 8) indicate that, like HC21, they all lack ancient AS sequences in their current pericentromeric regions. This suggests that all the human acrocentric p-arms may have been similarly reorganized some 20 million years ago or share the same evolutionary trajectory in some other way. However, the current chromosomal assemblies of acrocentric chromosomes still have large gaps, so either completion of the assemblies or analysis of single chromosome somatic cell hybrids similar to what we did for HC21 are needed to draw any firm conclusions.

Acknowledgments

Parts of this work were supported by Loyola University Chicago. Other portions were supported by the Government of the Russian Federation (№ 14.B25.31.0033) and by the Russian Science Foundation (№ 14-50-00029). EIR was supported in part by R01 AG029360. IAA was supported by Research Center of Mental Health, Russian Academy of Medical Sciences. We thank Michael Shaffer for assistance with phylogenetic analyses.

The experiments in this article comply with the current laws of the countries in which they were performed. The authors declare that they have no conflicts of interest. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

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

Bandyopadhyay R, McQuillan C, Page SL, Choo KHA, Shaffer LG (2001) Identification and characterization of satellite III subfamilies to the acrocentric chromosomes. Chromosome Res 9: 223-233.

Bozovsky MR, Shukai, SA, Cummings MR, Doering JL (2004) Organization of the regions flanking the centromere of human chromosome 21. [abstract 1567]. Available from http://www.ashg.org/genetics/ashg04s/

Brun M-E, Ruault M, Ventura, M, Roizes G, De Sario A (2003) Juxtacentromeric region of human chromosome 21: a boundary between centromeric heterochromatin and euchromatic chromosome arms. Gene 312: 41-50.

Bun C, Ziccardi W, Doering JL, Putonti C (2012) MiIP: The Monomer Identification and Isolation Program. Evol Bioinform Online 8: 293–300.

Cardone MF, Ballarati L, Ventura M, Rocchi M, Marozzi A, Ginelli E, Meneveri R (2004) Evolution of beta satellite DNA sequences: Evidence for duplication-mediated repeat amplification and spreading. Mol Biol Evol 21: 1792-1799.

Carnahan SL, Palamidis-Bourtsos E, Musich PR, Doering, JL (1993) Characterization of an evolutionarily old human alphoid DNA. Gene 123: 219-225.

Choo KH, Vissel B, Brown R, Filby RG, Earle E (1988) Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian translocations. Nucl Acid Res 16: 1273-1284.

Choo KH (1990) Role of acrocentric cen-pter satellite DNA in Robertsonian translocation and chromosomal non-disjunction. Mol Biol Med 7: 437-449.

Doering JD, Jelachich ML, Hanlon, KM. (1982) Identification and genomic organization of human tRNALys genes. FEBS Lett 146: 1620-1624.

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27: 221-224.

Hayden KE, Strome, ED, Merrett, SL, Lee, HR, Rudd, MK, Willard, HF (2013) Sequences associated with centromere competency in the human genome. Mol Cell Biol. 33: 763-772.

Ikeno M, Masumoto H, Okazaki T (1994) Distribution of CENP-B boxes reflected in CREST centromeric antigenic sites on long range α-satellite DNA arrays of human chromosome 21. Hum Mol Genet 3: 1245-1257.

Jarmuz M, Glotzbach CD, Bailey KA, Bandyophahyay R, Shaffer LG (2007) The evolution of satellite III DNA subfamilies among primates. Am J Hum Genet 80: 495-501.

Jordan GE and Piel WH (2008) Phylowidget: web-based visualizations for the tree of life. Bioinformatics 24: 1641-1642.

Kazakov AE, Shepelev VA, Tumeneva IG, Alexandrov AA, Yurov YB, Alexandrov IA (2003) Interspersed repeats are found predominantly in the "old" alpha satellite families. Genomics. 82: 619-627.

Lyle R, Prandini P, Osoegawa K et al. (2007) Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. Genome Res 17: 1690-1696.

Mashkova TD, Tiumeneva IG, Zinov'eva OL, Romanova LIu, Jubs E, Alexandrov IA. (1996) Pericentromeric alpha-satellite DNA in human chromosome 21 bordering with euchromatin DNA. Mol Biol 30: 617-625.

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ (2014) Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res 24: 697-707.

Miller DA (1977) Evolution of primate chromosomes. Science 198: 1116-1124.

Perrière G and Gouy (1996) WWW-query: an on-line retrieval system for biological sequence banks. Biochimie 78: 364-369.

Romanova LY, Deriagin GV, Mashkova TD, Tumeneva IG, Mushegian AR, Kisselev LL, Alexandrov IA (1996) Evidence for selection in evolution of alpha satellite DNA: the central role of CENP-B/pJ alpha binding region. J Mol Biol 261: 334-40.

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

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

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

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF (2001) Genomic and genetic definition of a functional human centromere. Science 294: 109-115.

Schueler MG, Dunn JM, Bird CP, et al. (2005) Progressive proximal expansion of the primate X chromosome centromere. Proc Natl Acad Sci USA 102: 10563-10568.

Shepelev VA, Alexandrov AA, Yurov YB, Alexandrov IA. (2009) The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. PLoS Genet 5: e1000641.

Shepelev VA, Uralsky LI, Alexandrov AA, Yurov YB, Rogaev EI, Alexandrov IA (2015) Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. Genomics Data 5: 139-146.

Smit AF, Tóth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246: 401-417.

Smith G (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191: 528-535.

Stanyon, R, Rocchi, M, Capozzi, O, et al. (2008) Primate chromosome evolution: Ancestral karyotypes, marker order and neocentromeres. Chromosome Res 16: 17-39.

Trowell HE, Nagy A, Vissel B, Choo KH (1993) Long-range analyses of the centromeric regions of human chromosomes 13, 14 and 21: identification of a narrow domain containing two key centromeric DNA elements. Hum Mol Genet 2: 1639-1649.

Vissel B and Choo KH (1991) Four distinct alpha satellite subfamilies shared by human chromosomes 13, 14 and 21. Nucl Acid Res 19: 271-277.

Wang SY, Cruts M, Del-Favero J, et al. (1999) A high-resolution physical map of human chromosome 21p using yeast artificial chromosomes. Genome Research. 9: 1059–1073.

Willard HF and Waye JS (1987) Hierarchical order in chromosome-specific human alpha satellite DNA. Trends Genet 3: 192-198.

Zhao C. (1999) Physical mapping of alphoid sequences on human chromosome 21. Masters Thesis. Loyola University Chicago.

LEGENDS

Figure 1. *YAC Map of the α21-II Region of HC21p* - YACs are shown as horizontal lines drawn to scale. Selected markers are indicated by thin vertical lines, and hybrid cell line breakpoints are indicated by arrows. The positions of 11 different AS sequence-containing plasmid clones are indicated by vertical dashed lines, and their placement in α21-II clusters is noted by red brackets. Circles are shown where a minor deletion in a YAC may be present.

Figure 2. *HC21p BACs Annotated* – All of the HC21p BAC/cosmids analyzed in this work shown to scale with major repetitive sequences noted. BAC/cosmids are shown as solid black lines, and repetitive sequences are shown as solid boxes with color determined by sequence type. The centromere proximal end of the Mp1 contig is to the right in this diagram. Mp1 is at least 140 kb long (Suppl. Table 1), which leaves a total of 21.3 kb of missing sequence from the cluster after subtracting the sizes of the AS clusters within AF105153 and AF254982. The AS cluster in Mp3 is more than 161 kb long, and this must be considered a minimum estimate since the Mp3 contig ends with AS sequence followed by an L1 insertion, making it possible that there are additional AS sequences directly adjacent to the contig.

Figure 3. *Dot Plot of Mp3* - Result of a dot plot analysis of the Mp3 contig (see Figure 3) created as described in Materials and Methods using a window size of 171 nucleotides and a mismatch limit of 8 nucleotides (5%). Boundaries of repetitive clusters are noted by vertical lines. L1 insertions are indicated by a vertical line topped by an *. A polymorphic HOR was detected in a portion of the AS sequence in the cluster, starting at approximately 52 kb into the sequence and extending 78 kb. The size of the HOR unit is approximately 3.9 kb, but many individual copies of the HOR contain deletions or duplications.

Figure 4. *Rs Analysis of WAV17*- The proportion of monomers with rs<0.62, as identified by the PERCON program (blue bars), was used as an indicator of the presence of ancient AS monomers in deep sequencing reads of human genomic DNA and in the HC21 single chromosome somatic cell hybrid WAV-17. Various truncated versions of the HCX human genomic assembly (hg38), where the ancient monomers have been previously mapped and their real proportion (black bars) is known (Shepelev et al. 2009), are shown as controls (Figure 5, Suppl. Table 6). The bars show the percentage of a given category in total SF4+ AS monomers scored in a sample. HCXp and HCXq regions containing only old AS sequences revealed that monomers with rs<0.62 are virtually absent in these domains (1% background). Such monomers are 22% in the youngest ancient layer (olive-green) and are about 40% in yet older ancient layers

of the p- and q-arms of HCX. The overall ratio of monomers with rs<0.62 to real ancient monomers, established for the whole HCX (0.64), was used to calculate predicted proportions (red bars) of ancient monomers in HC21 (WAV-17) and human genomic DNA. Controls show that predicted values can be both higher and lower than real ones depending on the composition of the ancient layers.

Figure 5. *Alpha satellite layers in human chromosomes 8, 17, X, and 21* - An updated version of Figure 2 from Shepelev et al. (2009), modified to include chromosome 21. As in the original figure, each AS array is depicted as a colored domain whose color indicates the identity of the monomeric types and relative ages of the AS monomers that comprise that array (Table 1). New AS HOR domains which act as centromeres are not drawn to scale and are shown as diagonally-crossed white and light blue boxes. On HC21p, the location of Mp1 is determined from a complete BAC contig while the locations of the other Mp clusters are estimated from the YAC mapping data in Figure 1. The polymorphic HOR cluster in the Mp3 region of HC21p is shown as both yellow and yellow-stripped since both monomer variants are part of the HOR unit. The maximum size estimate for each HC21p AS array (Suppl. Table 1) is used in the map.

Table 1. *Characteristics of AS layers* – Adapted from Shepelev et al. (2009). The color-coded dead layers identified within the former SF4 group are now referred to as subdivisions of the umbrella group named SF4+, pending the development of a final classification scheme. "rs" indicates the relative similarity score of a significant proportion of monomers in the respective layer that assists in AS sequence classification (see Materials and Methods). "Monomer length" refers to the presence (171 bp) or absence (172 bp) of a characteristic deletion at position 21 of the AS consensus monomer.

Table 2. *L1 analysis by genomic location and age* - L1s in the HC21 BAC/cosmids are grouped by their location relative to AS sequences (embedded within an AS cluster, directly adjacent to an AS cluster, or free from association with any AS sequence). All L1s were typed by evolutionary age with ancient L1s defined as insertions older in origin than L1PA3. Modern L1s were defined as L1PA3 and more recently evolved L1 insertions (Smit et. al 1995). The table shows the total numbers of L1s found in each location, the percentage of L1s found in that region, the numbers of modern and ancient L1 insertions found in each location, and the relative percentages of modern and ancient L1s in each region. L1s are most often found free from AS sequences, and those insertions are most often ancient. L1s are more rarely found embedded in or adjacent to AS sequences, and in those cases they are most often modern in evolutionary origin.

Line figure

Click here to download Line figure Fig 1 YAC Map.jpg

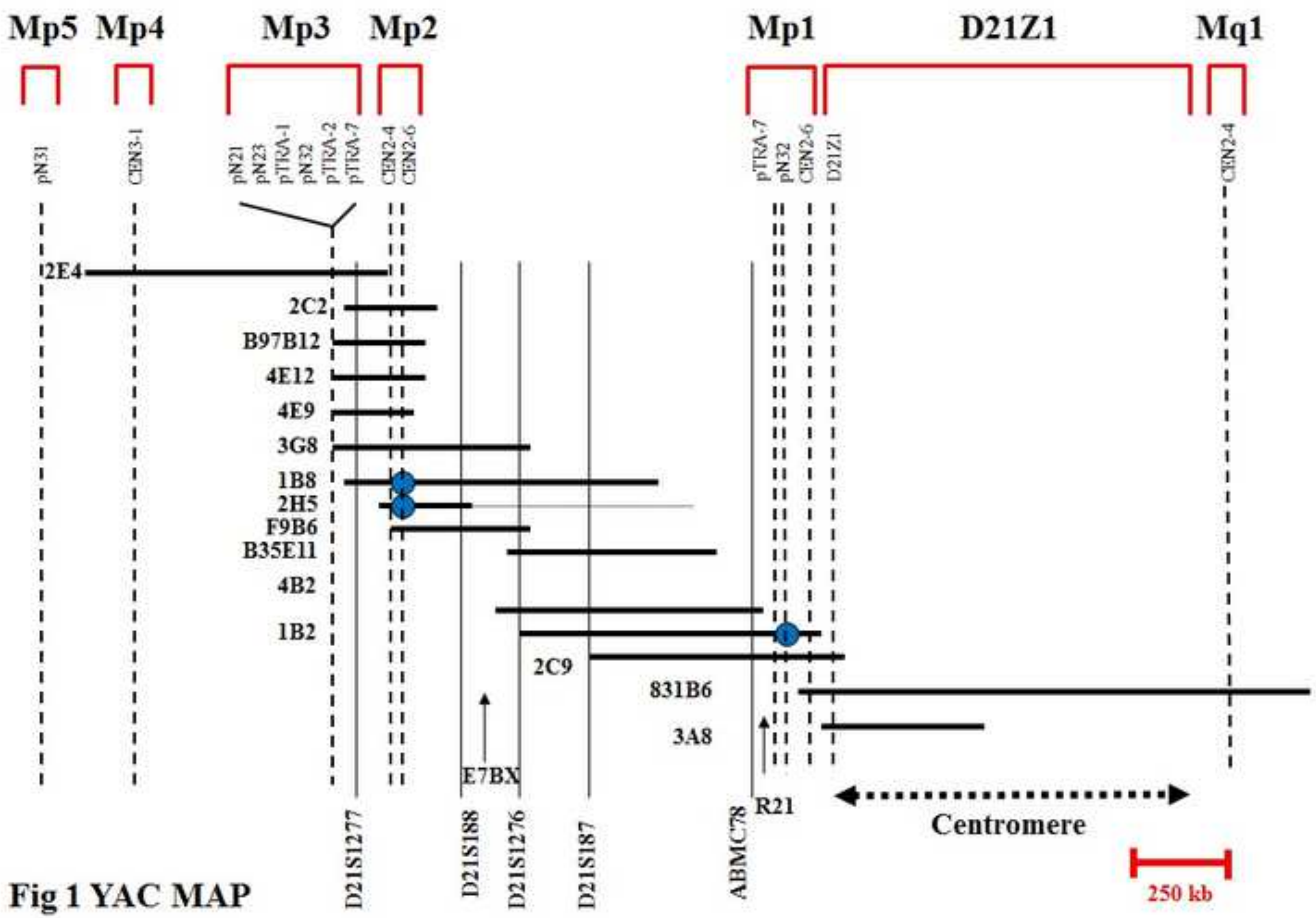


Fig 1 YAC MAP

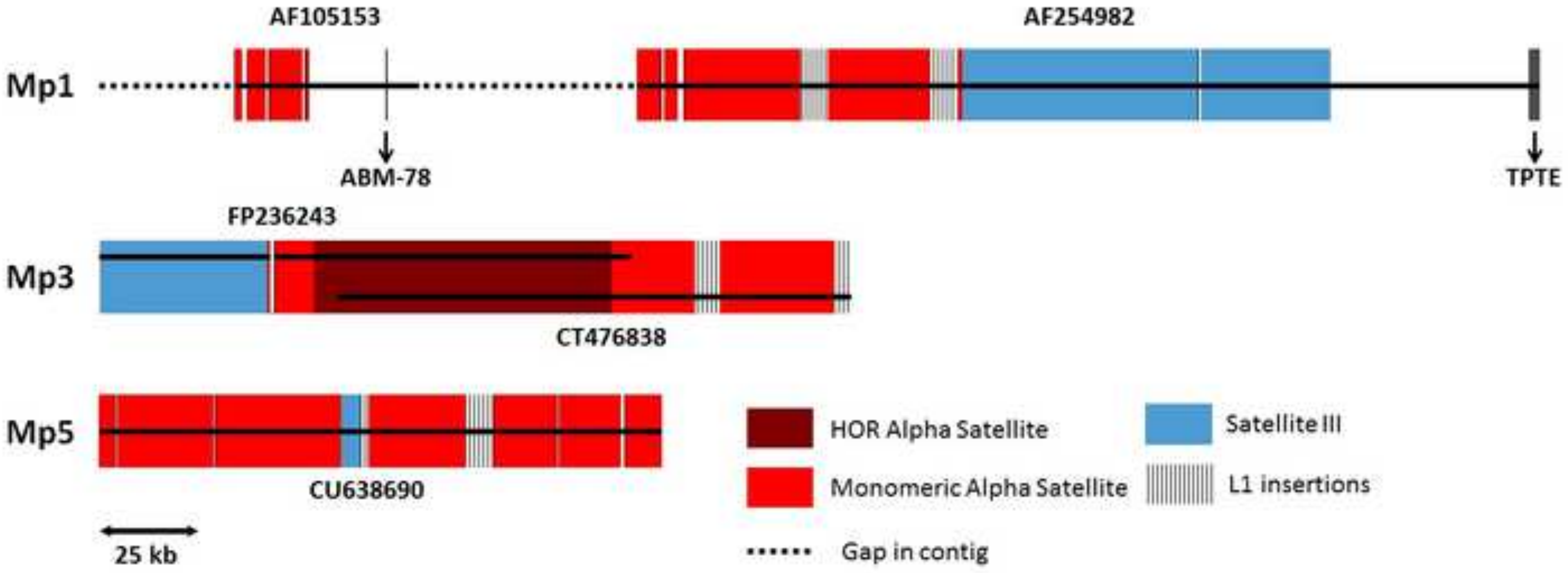


Fig 2 HC21p BACs Characterized

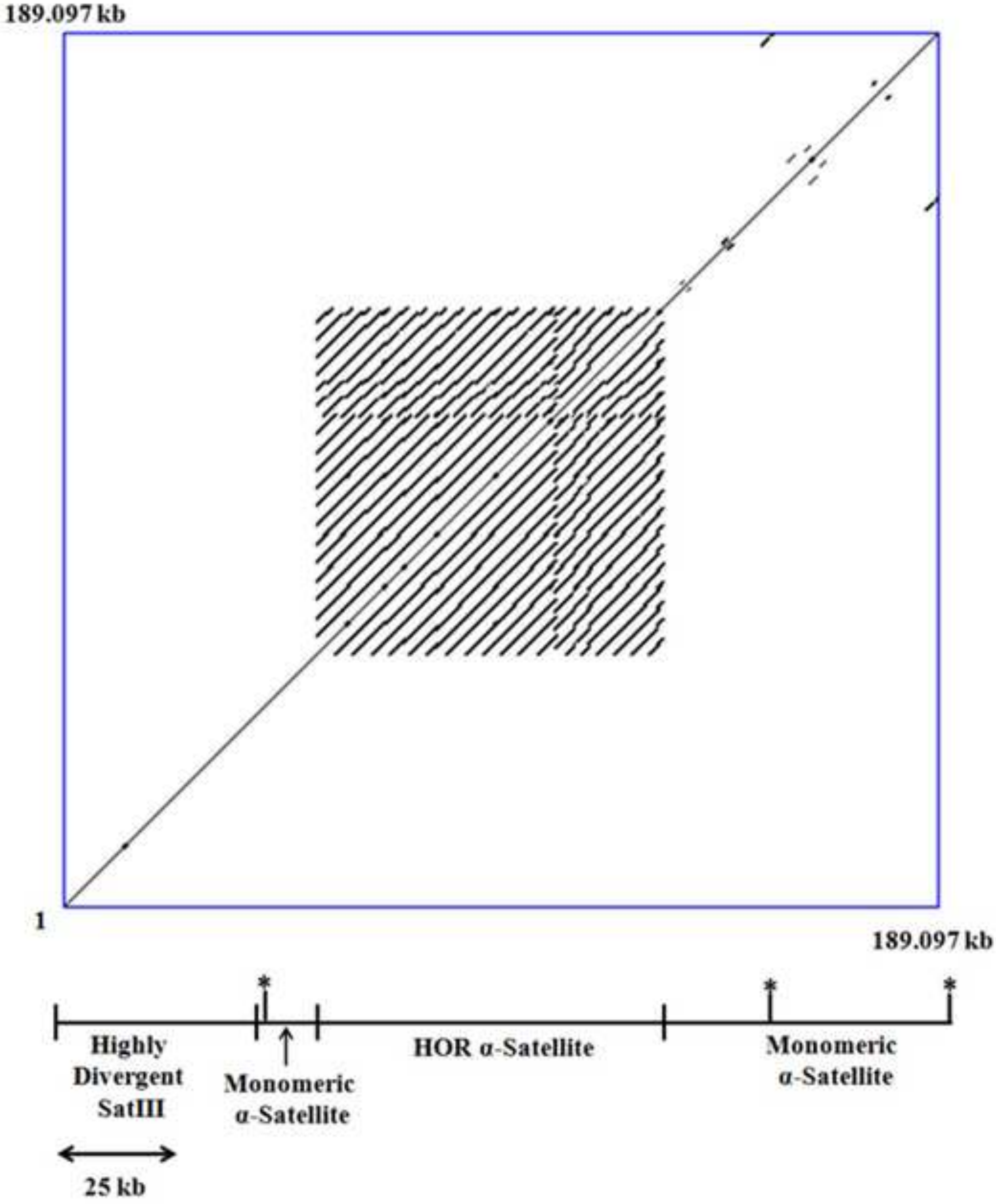


Fig 3 Mp3 Dot Plot

Click here to download Colour figure Fig 4 rs Analysis of WAV17.jpg

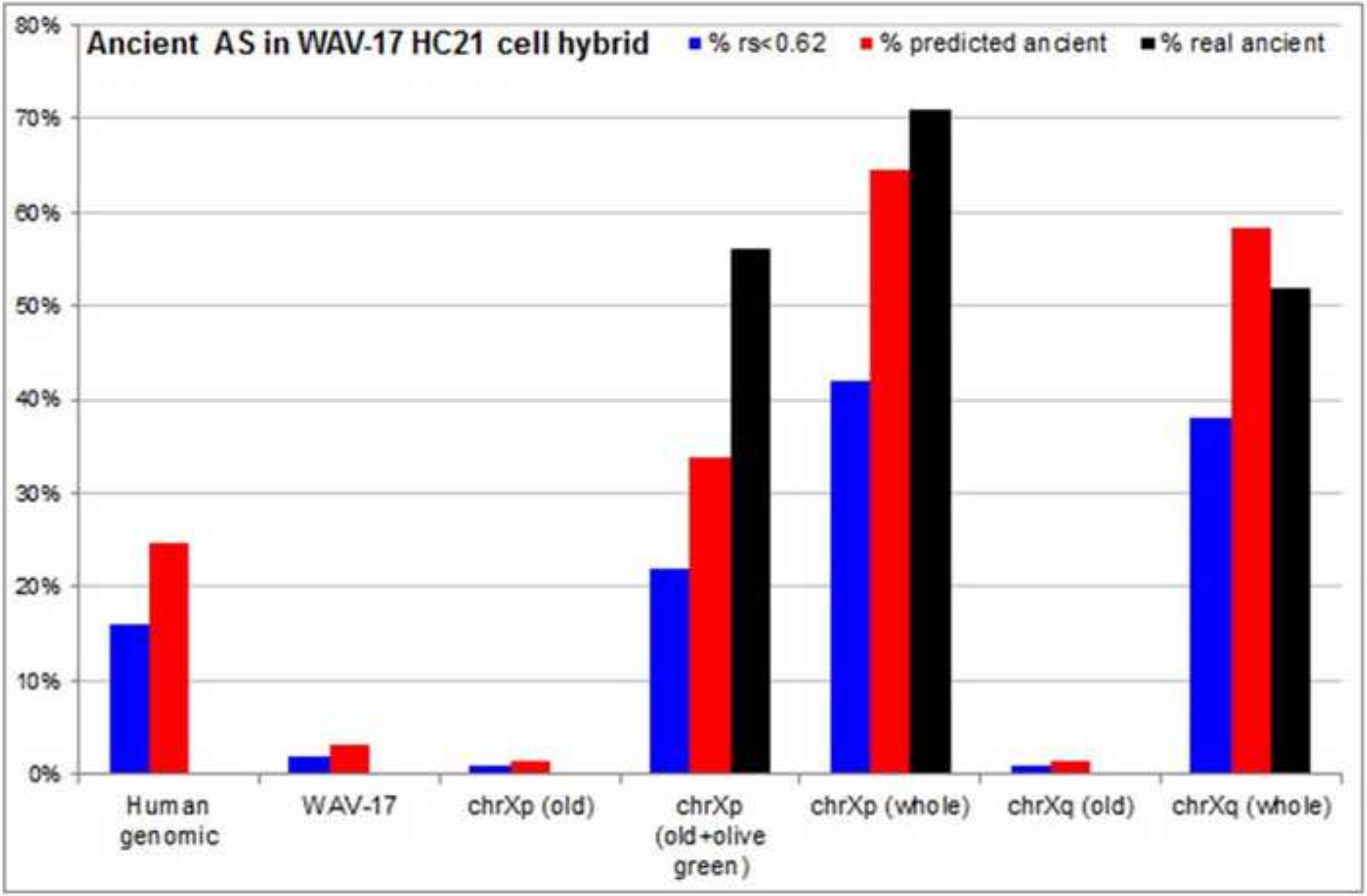


Fig 4 rs Analysis of WAV17

Colour figure

Click here to download Colour figure Fig 5 Alpha Satellite Layers in Human Chromosomes.jpg

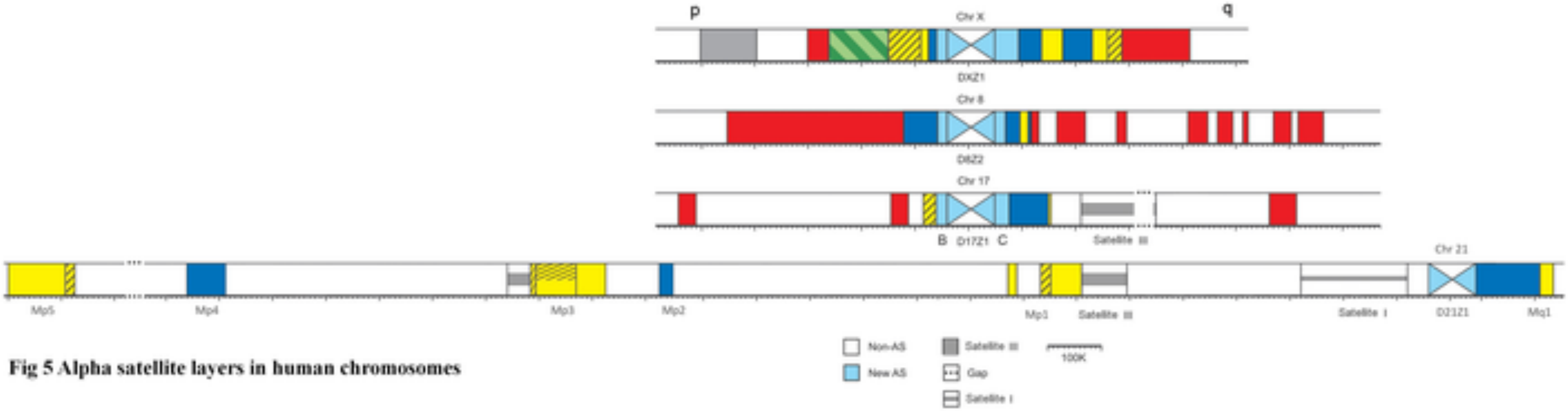


Fig 5 Alpha satellite layers in human chromosomes

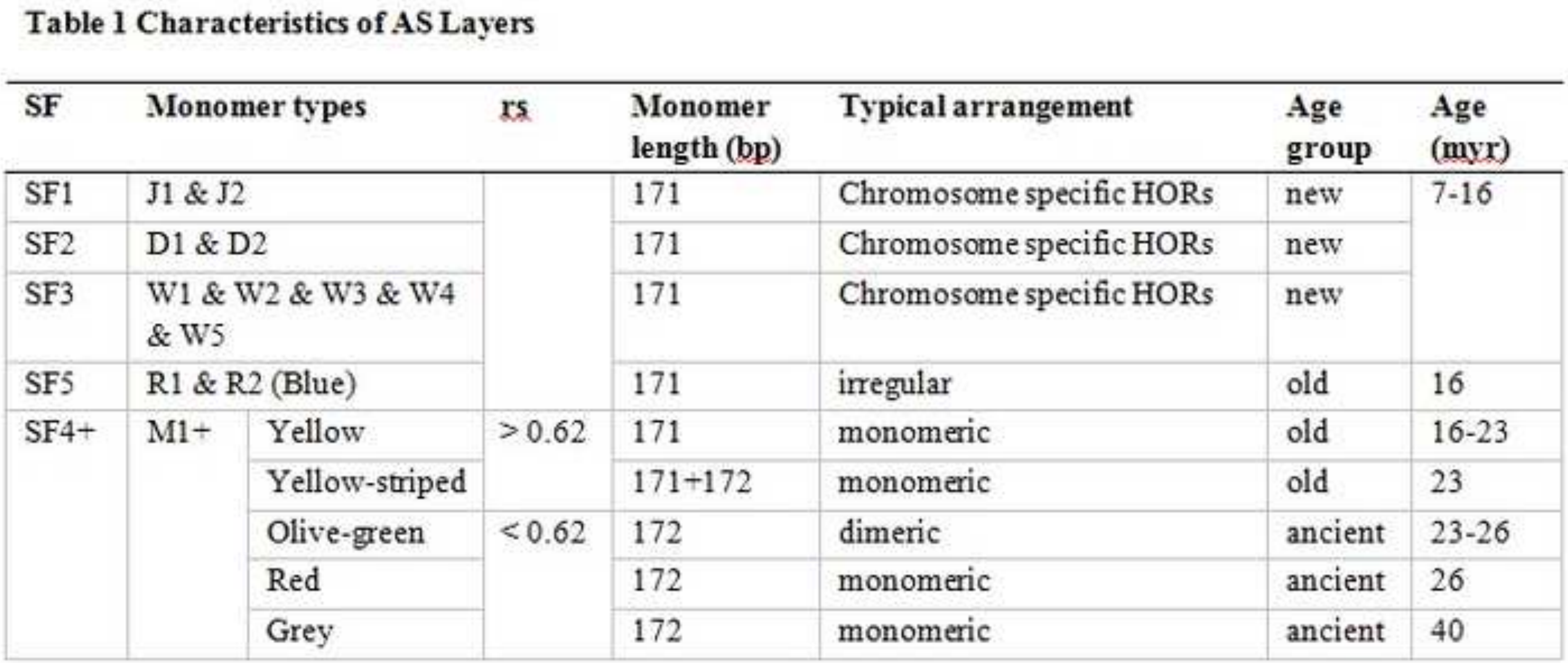
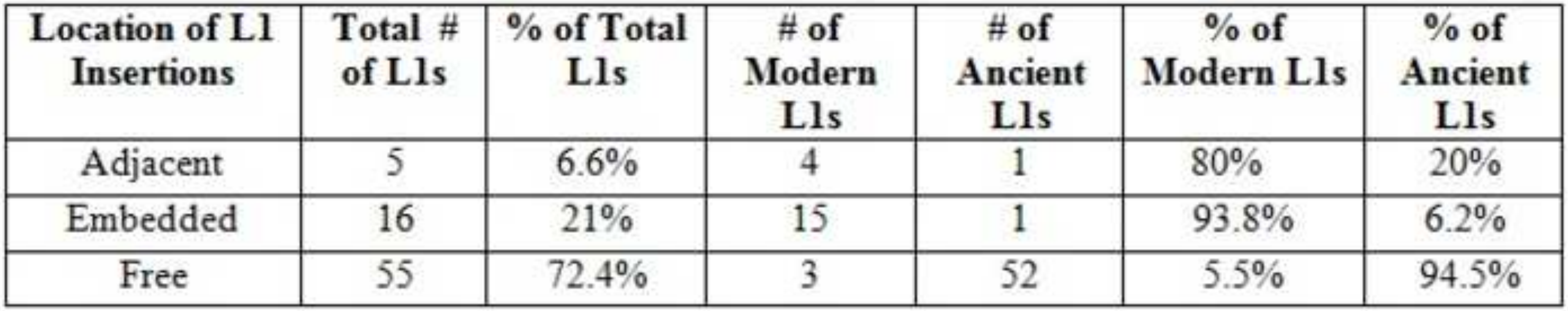


Table 1 Characteristics of AS Layers

SF	Monomer types		rs	Monomer length (bp)	Typical arrangement	Age group	Age (myr)
SF1	J1 & J2			171	Chromosome specific HORs	new	7-16
SF2	D1 & D2			171	Chromosome specific HORs	new	
SF3	W1 & W2 & W3 & W4 & W5			171	Chromosome specific HORs	new	
SF5	R1 & R2 (Blue)			171	irregular	old	16
SF4+	M1+	Yellow	> 0.62	171	monomeric	old	16-23
		Yellow-striped		171+172	monomeric	old	23
		Olive-green	< 0.62	172	dimeric	ancient	23-26
		Red		172	monomeric	ancient	26
		Grey		172	monomeric	ancient	40

Table 2 L1 analysis by genomic location and age



Location of L1 Insertions	Total # of L1s	% of Total L1s	# of Modern L1s	# of Ancient L1s	% of Modern L1s	% of Ancient L1s
Adjacent	5	6.6%	4	1	80%	20%
Embedded	16	21%	15	1	93.8%	6.2%
Free	55	72.4%	3	52	5.5%	94.5%

Supplementary material

Supplementary material

Click here to access/download

Supplementary material

Suppl. Table 5 L1 Insertions in HC21p BAC and Cosmid Clones.xlsx