Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly

V. A. Shepelev
Russian Academy of Sciences

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/oapubs

Part of the Genomics Commons

Repository Citation
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. Open Access Publications by UMMS Authors. https://doi.org/10.1016/j.gdata.2015.05.035. Retrieved from https://escholarship.umassmed.edu/oapubs/2618

Creative Commons License
This work is licensed under a Creative Commons Attribution 4.0 License.
This material is brought to you by eScholarship@UMassChan. It has been accepted for inclusion in Open Access Publications by UMMS Authors by an authorized administrator of eScholarship@UMassChan. For more information, please contact Lisa.Palmer@umassmed.edu.
Data in Brief

Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly

V.A. Shepelev a,b,c, L.I. Uralskya,c, A.A. Alexandrov d, Y.B. Yurovd, E.I. Rogaev b,c,f,⁎, I.A. Alexandrov d,⁎⁎

⁎ Correspondence to: Brudnick Neuropsychiatric Research Institute, University of Massachusetts Medical School, 303 Belmont Street, Worcester MA 01604, USA. Tel.: +1 508 856 4383.
⁎⁎ Corresponding author. Tel./fax: +7 495 420 5996.
E-mail addresses: Evgeny.Rogaev@umassmed.edu (E.I. Rogaev), ivanals@hotmail.com (I.A. Alexandrov).

**Correspondence to: Brudnick Neuropsychiatric Research Institute, University of Massachusetts Medical School, 303 Belmont Street, Worcester MA 01604, USA. Tel.: +1 508 856 8495, +1 508 856 4383.

1. Direct link to deposited data

Deposited data are available here: \textit{http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&hgt.customText=https://dl.dropboxusercontent.com/u/22994534/AS-tracks/human-GRC-hg38-M1SFs.bed.bz2.}

2. Materials and methods

2.1. A general layout of AS sequences in \textit{hg38} assembly

Centromeric regions of human chromosomes in \textit{hg38} assembly \cite{1} (GCA_000001405.15) can be divided in two main parts. One is a functional homogeneous core of each centromere which consists of thousands of copies of ~98\% identical higher-order repeats (HORs) composed of 2–20 divergent copies of an ~170 bp AS monomer \cite{2,3}.
As a rule, HORs are different on different chromosomes, but 3 non-homologous pairs of chromosomes share almost identical or very similar HORs (the so-called “paired domains”) [3, 13, 21, 14/22 and 5/19]. Each core is flanked by layers after layers of sequence formed by divergent monomeric or dimeric arrays devoid of homogeneous HORs [3–6]. These layers are composed of slightly different types of monomers and represent the “dead” remnants of the centromeres of our pre-great ape ancestors, which had no chromosome-specific HORs, but rather monomeric or dimeric AS identical in all chromosomes with the possible exception of the Y [3, 6, 7]. The farther from the “live” homogeneous core, the older and more divergent the dead layers are [6] and more signs of “post-mortal” damage such as deletions, inversions and insertions of mobile elements they display [3, 6]. The structure of the flanking pericentromeric regions is more or less symmetrical and each specific layer is often present on both sides of homogeneous core, which performs the centromeric function and forms a kinetochore [6]. The dead divergent layers cannot function as a centromere, but form pericentromeric heterochromatin [8].

With the exception of the Y chromosome, functional HOR arrays can be classified into three “new” suprachromosomal families (SFs 1, 2 and 3), each residing on a number of chromosomes. The older non-HOR AS is divided into the two large groups SF5 and SF4. SF5 is evolutionarily younger and immediately ancestral to the new families. On most chromosomes it directly flanks the functional HOR arrays [3]. SF4 group contains all the older layers of non-HOR AS. 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 [6]. Here we refer to the old SF4 as the SF4+ umbrella group, which includes the yellow layer (SF4 proper) and all the older layers defined in [6]. The new SF 1–3, SF5 and SF4+ groups are all composed of their own classes of monomers ([3] and Table 1) recognizable by the PERCON program [7]. In this work, we do not annotate the colored layers within SF4+ (monomeric group M1+), as their classification has yet to be completed.

In previous assemblies of the human genome, most of the HOR AS was absent and the core was occupied by a centromeric gap. In the latest hg38 assembly, the 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 recreate the true long-range linear order across the entire array [1, 9]. They can however be very helpful in mapping the AS deep sequencing or WGS reads to the human genome assembly.

Due to the complex pattern of intra- and inter-chromosomal identities in the pericentromeric regions of the acrocentric chromosomes 13, 14, 21 and 22, the mapping protocol used for 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 into the former centromeric gap on each chromosome. The same block of 13 reference models arranged in about the same order appears on all four chromosomes, but individual reference models have different names on every chromosome. Note that this block includes two live centromeres (paired domains 13/21 and 14/22), of which only one is actually present on any particular chromosome. Also, the identical sets of 3 AS reference models (of which only one is alive) appear on chromosomes 5 and 19 (paired domain 5/19), and the live model from this set also appears on chromosome 1 where the HOR is very similar to 5/19 paired domain and apparently cannot be distinguished by reference model assembly process (see Tables 2 and S1).

### 2.2. AS classification used by PERCON in the context of the human genome

AS was classed into five suprachromosomal families (SFs 1–3, SF4+ and SF5) according to monomeric classes in the sequence (Table 1), as described earlier [7]. Of those, SFs 1–3 are the new families of homogeneous HORs residing in functional centromeres in all autosomes and the X. In many chromosomes, on the periphery of the live HOR domain, much smaller domains formed by different new family HORs may also be present [3, 10]. These could be the remnants of formerly functional centromeres, which have been recently replaced by other new family HOR domains and have been heavily deleted since their death. Such damaged dead centromeric domains are expected to appear on both sides of a live centromere and to be somewhat less homogeneous and less regular. On the other hand, they could be just occasional amplifications of a piece of AS, the HORs which have never had centromeric function which we termed pseudocentromeres. If such pseudocentromeric HOR has amplified a piece of AS residing in a segment duplication (SD) or a piece of some atypical border sequence, it may appear as an AS domain with unexpected location or composition. Also, if a piece of a damaged old centromere is amplified, it may once again appear as

### Table 1

| Functional state, location | Old classification<sup>a</sup> | New classification<sup>[6]</sup> | Ancestral arrangement<sup>b</sup> | Age group |
|---------------------------|--------------------------------|---------------------------------|---------------------------------|-----------|
| Live SFs                  | SF1                            | SF1 colored layer               | Dimeric                         | New       |
|                           | J1(A)                          | J1(A)                           |                                 |           |
|                           | J2(B)                          | J2(B)                           |                                 |           |
| Core centromere           | SF2                            | SF2                             | Dimeric                         | New       |
|                           | D1(B)                          | D1(B)                           |                                 |           |
|                           | D2(A)                          | D2(A)                           |                                 |           |
|                           | SF3                            | SF3                             | Pentameric                       | New       |
|                           | W1(B)                          | W1(B)                           |                                 |           |
|                           | W2(B)                          | W2(B)                           |                                 |           |
|                           | W3(B)                          | W3(B)                           |                                 |           |
|                           | W4(A)                          | W4(A)                           |                                 |           |
| Dead SFs/layers           | SF5                            | Blue (SF5)                      | Irregular                        | Old       |
|                           | W5(A)                          | R1(B)                           |                                 |           |
|                           | R2(A)                          | R2(A)                           |                                 |           |
|                           | SF4+                           | Yellow (SF4)                    |                                 | Old       |
|                           | M1(A)                          |                                  |                                 |           |
|                           | Yellow-striped (SF6)           | V1(A)                           |                                 | Old       |
|                           | Olive-green (SF7)              | H1(A)                           |                                 | Dimeric   |
| Peripheral centromere     | SF4+                           | Red (SF7)                       |                                 | Ancient   |
|                           | M1+ (A)                        | H3(A)                           |                                 |           |
|                           | Gray (SF7)                     | H4(A)                           |                                 | Ancient   |

The table summarizes the data reviewed in [3] and reported in [6].

<sup>a</sup> Used in this paper.

<sup>b</sup> In live domains ancestral arrangement can only be observed or deduced from monomer order within a HOR unit.
Table 2
The list of unique AS reference models in hg38 assembly.

| # | Chrom | Name        | Size (bp) | SF | State | HOR symbol*[3] |
|---|-------|-------------|-----------|----|-------|----------------|
| 1 | chr1  | GJ211836.1  | 198,076   | 3  |       |                |
| 2 | chr1  | GJ211837.1  | 278,512   | 3  |       |                |
| 3 | chr1  | GJ211855.1  | 63,597    | 3  |       |                |
| 4 | chr1  | GJ211857.1  | 83,495    | 3  |       |                |
| 5 | chr1, 5, 19b | GJ212202.1 | 2,282,185 | 1  | Live | D1Z7/D5Z2/D19Z3 |
| 6 | chr2  | GJ211860.1  | 1,902,412 | 2  | Live | D2Z1          |
| 7 | chr3  | GJ211866.1  | 461,128   | 1.5|       |                |
| 8 | chr3  | GJ211867.1  | 13,936    | 1.5|       |                |
| 9 | chr3  | GJ211871.1  | 2,102,155 | 1  | Live | D3Z1          |
| 10 | chr4  | GJ211881.1  | 2,031,890 | Live | D4Z1 |
| 11 | chr5  | GJ211882.1   | 1,843,521 | Live | D5Z2 |
| 12 | chr5  | GJ211883.1   | 1,902,412 | Live | D5Z2 |
| 13 | chr5  | GJ211884.1   | 264,463   | 3  |       |                |
| 14 | chr5  | GJ211886.1   | 46,345    | 5  |       |                |
| 15 | chr5  | GJ211887.1   | 142,630   | 1  |       |                |
| 16 | chr5, 19b | GJ211902.1 | 53,672    | 5  |       |                |
| 17 | chr5, 19b | GJ211906.2 | 338,504   | 5  |       |                |
| 18 | chr6  | GJ211907.1   | 1,276,046 | 1  | Live | D6Z1          |
| 19 | chr7  | GJ211908.1   | 2,658,581 | 1  | Live | D7Z1          |
| 20 | chr7  | GJ211914.1   | 150,232   | 5  |       |                |
| 21 | chr8  | GJ211909.1   | 88,022    | 4+ |       |                |
| 22 | chr9  | GJ211929.1   | 1,282,922 | Live | D9Z4 |
| 23 | chr10 | GJ211930.1   | 85,412    | 5  |       |                |
| 24 | chr10 | GJ211932.1   | 1,561,440 | Live | D10Z1 |
| 25 | chr10 | GJ211933.1   | 46,180    | 1  |       |                |
| 26 | chr10 | GJ211936.1   | 147,701   | 1  |       |                |
| 27 | chr11 | GJ211938.1   | 11,969    | 5  |       |                |
| 28 | chr11 | GJ211943.1   | 3,251,982 | Live | D11Z1 |
| 29 | chr11 | GJ211948.1   | 82,575    | 3  |       |                |
| 30 | chr12 | GJ211949.1   | 47,204    | 1  |       |                |
| 31 | chr12 | GJ211954.1   | 2,349,957 | Live | D12Z3 |
| 32 | chr13, 14, 21, 22a | GJ21552.1 | 22,537    | 4+ |       |                |
| 33 | chr13, 14, 21, 22a | GJ21961.2 | 88,022    | 4+ |       |                |
| 34 | chr13, 14, 21, 22a | GJ21962.2 | 54,133    | 4+ |       |                |
| 35 | chr13, 14, 21, 22a | GJ21963.2 | 63,535    | 4+ |       |                |
| 36 | chr13, 14, 21, 22a | GJ21965.2 | 20,670    | 5  |       |                |
| 37 | chr13, 14, 21, 22a | GJ21967.2 | 6690      | 4+ |       |                |
| 38 | chr13, 14, 21, 22a | GJ21968.2 | 3245      | 4+ |       |                |
| 39 | chr13, 14, 21, 22a | GJ21969.2 | 22,561    | 4+ |       |                |
| 40 | chr13, 14, 21, 22a | GJ21972.2 | 1,134,211 | 2  | Live | D14Z2/D12Z2 |
| 41 | chr13, 14, 21, 22a | GJ21986.2 | 1198      | 4+ |       |                |
| 42 | chr13, 14, 21, 22a | GJ21991.2 | 632,586   | Live | D13Z1/D12Z1 |
| 43 | chr13, 14, 21, 22a | GJ22005.1 | 340       | 1  |       |                |
| 44 | chr13, 14, 21, 22a | GJ22061.1 | 340       | 1  |       |                |
| 45 | chr15 | GJ20361.1   | 415,278   | 4+ |       |                 |
| 46 | chr15 | GJ20421.1   | 855,957   | 4+ |       |                 |
| 47 | chr15 | GJ20451.1   | 1,370,146 | Live | D15Z3 |
| 48 | chr16 | GJ20461.1   | 23,302    | 2  |       |                |
| 49 | chr16 | GJ2051.1    | 1,928,003 | 1  | Live | D16Z2          |
| 50 | chr17 | GJ2053.1    | 817,381   | 3  |       |                |
| 51 | chr17 | GJ2054.1    | 3,371,415 | Live | D17Z1 |
| 52 | chr17 | GJ2056.1    | 49,431    | 3  |       |                |
| 53 | chr18 | GJ2060.1    | 319,478   | 2  |       |                |
| 54 | chr18 | GJ2062.1    | 7,693,584 | Live | D18Z1 |
| 55 | chr18 | GJ2066.1    | 93,042    | 2  |       |                |
| 56 | chr18 | GJ2067.1    | 39,636    | 2  |       |                |
| 57 | chr18 | GJ2069.1    | 76,958    | 2  |       |                |
| 58 | chr18 | GJ2071.1    | 24,109    | 2  |       |                |
| 59 | chr20 | GJ2091.1    | 150,723   | 2  |       |                |
| 60 | chr20 | GJ2093.1    | 1,880,394 | Live | D20Z2 |
| 61 | chr20 | GJ2095.1    | 47,956    | 2.5|       |                |
| 62 | chr20 | GJ2101.5    | 80,766    | 4+ |       |                |
| 63 | chr20 | GJ2107.1    | 78,875    | 4+ |       |                |
| 64 | chr20 | GJ2117.1    | 120,944   | 5  |       |                |
| 65 | chrX  | GJ2192.1    | 3,806,963 | 3  | Live | DXX2          |
| 66 | chrY  | GJ2193.1    | 227,095   | 4+ | Live | DYY2          |

* a Identity of reference models marked as “live” with the known live HORs of respective chromosomes was verified by blasting the sequences in our HOR list in [3] to the first 10,000 bp of respective reference model. In all cases multiple hits of 93% or higher were obtained.

b Only one representative member of a group of identical reference models is listed. For complete list, see Supplementary Table S1.

c Corrected versions of these reference models were obtained from K. Miga and used for analysis.
known monomeric classes [14]. Position 1 in the monomer by tradition was arbitrarily assigned to the first nucleotide of the BamHI site in chromosome X-specific HOR DXZ1 [18]. The N positions in the A/B box of ALPHA-ALL were set to the A configuration as shown in Fig. S1. The repeated alignment was performed by a modified Smith–Waterman–Gotoh algorithm [19,20] and was stopped when relative alignment score (rs) of the monomers obtained became 0.29 or less. rs was calculated as an alignment score divided by reward for a match multiplied by the length of alignment. The alignment score was the number of matches multiplied by reward for a match minus the number of gaps multiplied by a gap opening penalty minus the number of nucleotides in gaps multiplied by a penalty for gap extension minus a number of mismatches multiplied by a penalty for mismatch. At the very ends of the monomer the alignment was not always precise and small gaps of up to 5 bp often separated the adjacent monomers. This did not affect monomer classification. Next, every monomer was classified into one of the 12 known standard monomer classes ([J1,J2,D1,D2,W1–W5,R1,R2 and M1] [3]) or defined as unclassified (Um) or random (Xm) by a simple Bayesian classification procedure that utilized consensus matrices of the 12 classes of monomers together with the random matrix (shown in Fig. S1). The program estimated the probability of a hypothesis that a given monomer belonged to one of the known monomeric classes and, if it met the threshold (typically 0.9), assigned the classification. Otherwise, the monomer was deemed “unclassified” (e.g. chimeric monomers where half belongs to one class and half to another class). Altogether, 14 groups were identified by PERCON. Long sequences were processed in consecutive 5 kb windows, which overlapped by about 200 bp.

Independently, the region of the A/B box (positions 35–51) was classed in every monomer in the sequence (A box, B box, X for random, U for unclassified and Q if, in the truncated monomer, the box region was not present). Classification was performed by Bayesian classifier in the same way as for the whole monomers. The matrices for the A and B boxes were generated by summation of all the consensus matrices of individual monomeric classes that belonged to type A and type B (shown in Fig. S1). Note that the A/B classification did not assess the functionality of the B-box in CENP-B binding—it just determined to which ancestral type the monomer belonged. In some rare cases, the box classification and the monomer classification may contradict each other (e.g. R2 monomer which has a B-box). At least some studied instances of such monomers are hybrids with the box region coming from one class and the rest of the monomer from the other (data not shown).

PERCON is available for download at: https://github.com/alsrat/PERCON.

2.4. UCSC Browser Track

The track was created by PERCON program developed by V.A. Shepelev and I.A. Alexandrov [7]. AS monomers were identified by PERCON similarity search, extracted and distributed into the classes characteristic of the 5 SFs by a Bayesian classifier. Program output contained detailed information on AS monomers, including monomeric class, result of independent typing of the A/B box, genomic coordinates and strand orientation, which were used for the annotation track. For incomplete monomers with length less than 140 bp, the monomer class was shown in lowercase letters; for longer monomers in uppercase letters. These data generated for hg38 human genome assembly were transformed into a Browser Extensible Display (BED) format suitable for viewing as a custom annotation track in the UCSC Human Genome Browser (http://genome.ucsc.edu/). To convert PERCON output to BED format we wrote an AWK script (available at https://github.com/enigene/prcn2BED) which also color-codes the monomers according to the monomeric type. After using the prcn2BED, the resulting BED file was processed by a second AWK script (available at https://github.com/enigene/remisct) to remove duplicate segments resulting from overlap of the 5 kb windows. Of two overlapping monomers, the longer one remained intact and the shorter one was trimmed. The resulting annotation track is available at http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&hgt=customText=https://dl.dropboxusercontent.com/u/22994534/AS-tracks/human-GRC-hg38-M1SFs.bed.bz2. The track is self-explanatory, as in the full-view mode each monomer is marked with respect to its monomeric class and the A/B type according to Table 1. By using the Table Browser, the track data can be analyzed in text format and filtered or transformed to generate various statistics. For instance, different classes of monomers can be counted per individual chromosome or chromosomal region. Also, the overlaps with other tracks can be created and retrieved as a new track, DNA sequence, or in text format.

2.5. Overall statistics of AS in hg38 assembly

The overall statistics of AS was collected from the UCSC Browser Track using Table Browser and analyzed to control how the track data corresponded to what was known from other methods and sources. The overall detection of AS by PERCON (70.1 Mbp, 2.30%) did not differ significantly from that of RepeatMasker (http://repeatmasker.org/), as used in UCSC Browser RepeatMasker Track (70.8 Mbp, 2.32%). RepeatMasker records that had at least 98% overlap with the PERCON track constituted 69.5 Mbp or 98.2% of total RepeatMasker detection.
RepeMasker records that had no overlap with PERCON AS SF track constituted 3628 bp or 0.01% of total detection. The latter records were all small fragments shorter than one monomer. The size of DNA occupied by monomers of each SF determined in hg38 assembly is shown in Fig. 1 where it is compared to the data obtained in the analysis of WGS database. One million HuRef WGS reads (PRJNA19621) obtained by random DNA fragmentation were processed by PERCON in the same way as described previously for analysis of the BAC ends [7]. There was no dramatic difference in SF profiles between the sets, except the proportion of unclassed monomers in WGS (6%) was predictably higher than in the assembly (2.5%). Both a large number of truncated monomers at the ends of WGS reads and the low quality of sequence at the ends of Sanger reads contributed to this difference (Fig. S2). To correct for these factors we evaluated the effect of trimming the bad ends using the LUCY program with default parameters [22] and of filtering the dataset for monomers 140 bp or longer (such monomers are marked with upper case letters in PERCON output). The results are shown in Fig. S2, where it can be seen that each step reduced both the number of unclassed monomers and the total AS detection. To combine good detection with more accurate SF measurements we used the SF proportions obtained in the double-filtered sample to divide the AS amount obtained in the unfiltered sample. After such correction (see legend to Fig. S2 for more details) the proportion of unclassed monomers dropped to 2% and the WGS data appeared to be in fairly good concordance with the assembly. The only significant difference was a larger SF3 in the assembly. The reasons for this discrepancy we did not investigate. A more detailed discussion of the possible sources of unclassed monomers was provided in [7]. The proportions of SFs in human genome shown in Fig. 1 are, as much as we know, the first accurate estimate obtained in a non-biased sample. The results differ significantly from the ones in [7] which was our previous attempt to SF-class a large sample of AS fragments. The difference is presumably due to restriction enzyme bias in the older sample.

The above statistics suggest that monomeric maps for most human AS sequences can be obtained without running PERCON, but simply by finding the same exact sequence or a very similar copy of it in the hg38 assembly.

2.6. Annotation of AS HOR reference models

For quick reference, we also provide tables of SF-annotated AS reference models as they appear in the assemblies of individual human chromosomes. Although annotation of individual HORs is beyond the scope of this work, we indicate the known live HORs, which are the largest reference models (except for the Y chromosome). Altogether 109 reference models are used in hg38 assembly (Table S1). After correction for identical models in different chromosomes, 66 unique models remain (Table 2). Of these, 18 unique models represent 22 live centromeres of autosomes, as chromosomes 13/21, 14/22 and 1/5/19 share the same live reference models. Two additional models represent live centromeres of sex chromosomes. All of them are classed in SFs as previously reported for respective live centromeres and can be recognized by high identity to the known centromeric sequences from the list shown in [3] (see Tables 2 and S1). Of the remaining 46 models, 24 contain variant live HORs, dead HORs or pseudocentromeres of the new families (SF1–3), and 10 contain SF5 HORs which were not known to form live centromeres and were traditionally perceived as pseudocentromeres. Previously, such HORs were reported on chromosomes 4, 5, 7, 19 and acrocentrics [15,16,21]. However, recent evidence shows that live centromeres of orangutan are likely formed by SF5 HORs [23], so some SF5 HOR domains may in fact be dead centromeres. Finally, 12 other unique pericentromeric reference models were classed as SF4+. Except for the live SF4 + centromere of chromosome Y, SF4 + HORs were not widely reported in man. The only example known to us was the pTRA-2 AS clone [21] which was classed as SF4 [3] and shown to form a cluster of 75 HORs in the short arm of chromosome 21 [24] (see the Discussion section below). In particular, SF4 + HORs are present in chromosomes 15, 20 and the 13/14/21/22 group.

Due to some problem in the reference model assembly process, 5 unique reference models (marked in Tables 2 and S1) were assembled incorrectly, with a reverse order of monomers in a HOR (K. Miga, personal communication). The corrected versions of these reference models were obtained from K. Miga and used throughout our analysis. Note that non-HOR AS is not supposed to appear in reference models.

Evaluation of HORs in the sequences listed in Tables 3–4 was performed using the dot-matrix construction program from http://www.vivo.colostate.edu/molkit/dnadot/, Gepard program [25] or the REVN program written by V.A. Shepelev [26].

3. Discussion

Inspection of the AS assembly track shows that PERCON adequately and comprehensively classes AS monomers and that they are organized predominantly into the major known modes characteristic of the known SFs. No long arrays of unclassed monomers are observed. SF1 and SF2 sequences are uniformly arranged in arrays with J1J2 and D1D2 dimeric periodicities, the remnants of W1W2W3W4W5 pentameric order can be discerned in SF3 sequences, and SF5 clusters demonstrate irregular alternation of R1 and R2 monomers. Although HORs are not annotated in the track, these repeats can be easily seen in SF3 and SF5 due to reiteration of complex patterns of W1-5 or R1 and R2 monomer classes. In more dimeric sequences, HORs can often be seen due to some irregularity which occurs once or twice in a HOR, like the occasional Um, R1 or R2 monomer often found in SF2 HORs, or some other breach of dimeric periodicity like D1D1 dimer in D18Z2. In most cases, these structural peculiarities faithfully reflect the features of individual HORs reported in literature, but often the reiteration of HORs in reference models appears to be imperfect or even dramatically disturbed. Whether this reflects the true genomic arrangement or some problem in the algorithm of reference model formation remains to be investigated.

We noted a few previously unreported or poorly-reported minor modes of AS organization, as follows: (i) relatively long clusters
composed entirely or almost entirely of R2 with very few or no R1; (ii) mixed occurrence of J1 and J2 monomers alternating with R1 and R2 over relatively long regions, or the same kind of pattern with D1/D2 or R1/R2 alternation; (iii) SF4+ HORs, which appear to be no less common than long known SF5 HORs. Below we will briefly comment on these unusual modes.

Pure R2 clusters were predicted by our scenario of SF5 formation and of the origin of the new families [3,6,7,14], but were actually found only recently due to their relative rarity [23]. We proposed that, in the common ancestor of orangutan and man, the centromeres were formed by pure R2 arrays and this was the last generation of type A centromeres in the human lineage. These centromeres, like all previous AS centromeres, had no CENP-B binding sites. At some point, R1 (type B) monomer, which had a CENP-B binding site, but was otherwise very similar to R2, formed due to several point mutations. The presence of CENP-B endowed this monomer with an ability to spread by irregular transposition-like process only within live R2 centromeres resulting in increased homogeneity and the presence in chimpanzee and gorilla genomes. We found inter-HOR identity of 98.5% and no high identity hits in great ape WGS datasets (data not shown). Thus, the HOR is unlikely to be a relic of an era of SF5 centromeres, but is probably a more recent pseudocentromere which appeared via amplification of a piece of a dead pure R2 array.

Mixed SF1/SF5 and SF2/SF5 domains documented in Table 4 occupy relatively large space in the assembly because they are amplified in chromosomes 3 and 20, while non-HOR mixed regions are a tiny fraction. Potentially, three explanations of such mixes are feasible. They could be: (i) former pure SF1 or SF2 domains that were heavily deleted with formation of a large number of hybrid monomers (e.g. half D1, half D2) which are usually classed as Um or either R1 or R2, depending on configuration of the A/B box in the monomer (data not shown). Solitary (one per HOR) apparent SF5, but actually hybrid D1/D2 monomers are often present in SF2 HORs (data not shown); (ii) mixes of genuine new SF and SF5 monomers, which perhaps formed on the border of SF5 and live centromeres by recombination across the border, and; (iii) clusters formed by early versions of D and J monomers, which differed from their respective HORs possess some of the mutations characteristic of J1 and J2 or D1 and D2, respectively. High homogeneity and absence in African apes indicate that both mixed SF5 HORs are likely pseudocentromeres. However, as both have divergent ancestral internal periodicities (see Supplementary note 1), these pseudocentromeres may have been formed by recent amplification of more ancient, dead HOR centromere material. Options (i) and (ii) may apply to other mixed domains which have no clear HORs. They warrant careful detailed investigation.

Table 4
SF1/SF5 and SF2/SF5 mixed AS regions in hg38 assembly.

| SF     | Location       | Contig          | Sizea (bp) | SF1% | SF2% | SF5% | HORs on dot-matrix |
|--------|----------------|-----------------|------------|------|------|------|-------------------|
| SF2/SF5 2q21.2 | chr:2:132,237-392-132,247,263 | AC0975312.3 | 9871 | 0 | 0 | 100 | No HOR |
| SF1/SF5 3p11.1 | chr:3:90,482,385-90,722,299 | ABBA01004652.1, AEKP01209350.1, ABBA01004653.1, AEKP01209355.1, ABBA01004654.1, ABBA01004655.1 | 22,441 | 0 | 0 | 100 | No HOR |
| SF1/SF5 3q11.1 | chr:3:90,722,554-91,233,510 | GJ211866.1 | 460,956 | 48 | 0 | 52 | HOR 1.7 kb |
| SF1/SF5 3q11.1 | chr:3:91,247,785-91,247,457 | GJ211867.1 | 13,765 | 57 | 0 | 33 | No HOR other than AB dimer, identity ~93% HOR 1.7 kb |
| SF1/SF5 3q11.1 | chr:3:92,716,246-92,725,964 | ABBA01000927.1, ABBA01000928.1, ABBA01000929.1, ABBA01000930.1, ABBA01000931.1 | 17,430 | 50 | 0 | 30 | No HOR |
| SF1/SF5 6q11.1 | chr:6:20,320,028-60,241,613 | AC244258.2 | 11,401 | 28 | 0 | 70 | No HOR |
| SF1/SF5 6q11.1 | chr:6:61,371,445-61,427,364 | AEKP01189806.1, AEKP01189805.1, AEKP01189804.1, AEKP01189803.1, AEKP01189802.1, FP325349.3 | 55,519 | 30 | 0 | 70 | No HOR |
| SF2/SF5 7q11.1 | chr:7:61,096,433-61,103,082 | AC142121.2, AC017075.8 | 6649 | 0 | 38 | 62 | No HOR |
| SF1/SF5 8p11.1 | chr:8:43,940,231-43,965,733 | AC127507.4, AC144576.3 | 22,886 | 15 | 0 | 70 | No HOR |
| SF1/SF5 8q11.1 | chr:8:45,946,092-45,971,262 | AC118650.5 | 22,549 | 16 | 0 | 70 | No HOR |
| SF2/SF5 9p11.2 | chr:9:40,556,928-40,565,104 | AMYH02020868.1, FP325318.4 | 7524 | 0 | 40 | 60 | No HOR |
| SF2/SF5 9p11.2 | chr:9:40,862,745-40,873,147 | AL353626.5 | 10,402 | 0 | 28 | 70 | No HOR |
| SF1/SF5 10p11.1 | chr:10:39,548,571-39,555,979 | 7408 | 35 | 0 | 52 | No HOR |
| SF2/SF5 10p11.1 | chr:10:39,548,571-39,555,979 | AC145653.4, AUXG0100432.1 | 28,658 | 60 | 0 | 30 | No HOR |
| SF2/SF5 12p11.1 | chr:12:34,686,342-34,715,037 | AC143531.4, AC143531.5 | 30,263 | 0 | 35 | 65 | No HOR |
| SF2/SF5 16p11.2 | chr:16:34,219,066-34,252,724 | AC136932.4, ABBA01017803.1 | 47,783 | 0 | 27 | 54 | HOR 1.9 kb |
| SF2/SF5 20q11.1 | chr:20:28,509,094-28,556,877 | GJ212095.1 | 6987 | 0 | 41 | 58 | No HOR |

Notes:
- a Size has been corrected to exclude L1-repeats and gaps.
- b Same HOR as in GJ211866.
So far, only two different non-orthologous live Y chromosome SF4+ HORs were known (DY23 in man and BACs AC195625, AC156580, AC163730, AC163738 in chimpanzee) [27,28]. Also, evidence for live SF4+ HORs in gibbons has been reported [6,29,30]. Additionally, the D21Z5 SF4+ HOR, also known as pTRA-2, (X55367, corresponds to GJ212124) was convincingly demonstrated on the sequence level [21, 24]. The latter appears to be a segmental duplication also present on other acrocentric chromosomes. In the human assembly, in addition to DY23, we class 12 reference models as SF4+. Eight of them are part of the group assigned to chromosomes 13/14/21/22, but not all of them necessarily reside on all of these chromosomes. For instance, another member of this group, the SF2 D12Z1/D21Z1 live HOR forms the live centromeres of chromosomes 13 and 21, and is not massively present in chromosomes 14 and 22 [3]. The same applies to chromosomes 14/22 live HOR which is absent in chromosomes 13 and 21. Also, not all of them are confirmed by regular contigs (see Table S2). Another four SF4+ models reside on chromosomes 15 and 20 (2 on each). The GJ212105 HOR is also located in regular contigs ABBAO1015870 and A1837517 in the same area on 20q. The latter contig has 54 kb HOR domain with HOR length 1872 bp. The high HOR similarity of 99.9%-100% is very unusual, which implies recent amplification. The large size of chromosome 15 reference models (GJ212036 and GJ212042) raises an interesting possibility that they are dead SF4+ HOR centromeres. These centromeres were reported only in gibbons [30], but not in great apes, with the notable exceptions of human and chimpanzee Y chromosomes. In these cases, however, the SF4+ centromeres function in chromosomes that do not have more recent generations of AS, such as SF5 or the new families, which presumably would compete for centromeric function more efficiently than SF4+ [6]. Arguably, in chromosome 15, which has SF2 functional centromere and some SF5 as well, the SF4+ HOR domains may only be the dead centromeres that date back to the time before the great apes, when SF5 and the new families did not exist. However, high homogeneity (99%) of these HORs, both in the reference models and in regular contigs (listed in Table S3) and their absence in genomes of chimpanzee and gorilla (data not shown) argue that they are recent mega-scale pseudocentromeres.

Our theoretical scenario of AS centromere evolution [6] states that only a functional centromere can stably maintain the mega-amplified state and homogeneity characteristic of live centromeres due to the involvement of hypothetical kinetochore-associated recombination machine (KARM). There are at least two possibilities to reconcile the existence of mega-scale pseudocentromeres in the short arm of chromosome 15 with this view. First, it could be just a rare occasion of the active centromeres of human chromosomes X and Y satellite genomes. Genes Res. 24 (2014 Apr) 697–709, http://dx.doi.org/10.1016/j.gene.2014.02.026 (Epub 2014 Feb 5. PubMed Central PMCID: PMC3795068).

KE. Hayden, E.D. Strome, S.L. Merrett, H.R. Lee, M.K. Rudd, H.F. Willard, Sequences associated with centromere competency in the human genome. Mol. Cell. Biol. 33 (4) (2013 Feb) 763–772, http://dx.doi.org/10.1128/MCB.01108–12 (Epub 2012 Dec 10. PubMed Central PMCID: 32320566; PubMed Central PMCID: 3571344).

A. Aebersold, D. Maier, S. Sillence, T. Driscoll, P. Hannon, J. Dean, D. Houlton, J. Cassiman, Functional epialleles at an endogenous human centromere. Proc. Natl. Acad. Sci. U. S. A. 109 (34) (2012 Aug 21) 13704–13709, http://dx.doi.org/10.1073/pnas.1203126109 (2012 Jul 30. PubMed Central PMCID: 22847449; PubMed Central PMCID: PMC3420787).

K.H. Miga, Y. Newton, M. Jain, A. Altermeze, H.F. Willard, W.J. Kent, Centromere reference models for human chromosomes X and Y satellite genomes. Genome Res. 14 (2004 Apr) 697–709, http://dx.doi.org/10.1101/gr.159624.113 (Epub 2014 Feb 5. PubMed Central PMCID:PMC3795068).

M.K. Rudd, G.A. Wray, H.F. Willard, The evolutionary dynamics of alpha satellite. Genome Res. 16 (1) (2006 Jan) 88–96 (Epub 2005 Dec 12. PubMed Central PMCID:PMC1596132).

M.K. Rudd, G.A. Wray, H.F. Willard, The evolutionary dynamics of alpha satellite. Genome Res. 16 (1) (2006 Jan) 88–96 (Epub 2005 Dec 12. PubMed Central PMCID:PMC1596132).

M. Maloney, P. McRae, I. Jones, M. Ewens, R. Tavare, J. Durrant, Analysis of the human Y chromosome. Proc. Natl. Acad. Sci. U. S. A. 109 (34) (2012 Aug 21) 13704–13709, http://dx.doi.org/10.1073/pnas.1203126109 (2012 Jul 30. PubMed Central PMCID: 22847449; PubMed Central PMCID: PMC3420787).

L.Y. Romanyova, G.V. Derrigan, T.D. Mashkova, G. Yurkov, A.R. Mushegian, L.L. Kisselev, L.A. Alexandrov, Reference models for selection in evolution of alpha satellite DNA: the central role of CENP-B at alpha binding domain. J. Mol. Biol. 261 (3) (1996 Aug 23) 334–340 (PubMed PMID: 8780776).

M.K. Rudd, H.F. Willard, Analysis of the centromeric regions of the human genome assembly. Trends Genet. 20 (11) (2004 Nov) 529–533 (PubMed PMID:15475110).

M. Rosandici, C. Viar, I. Basar, M. Gluncic, N. Pavlin, I. Plas, CENP-B box and alpha satellite sequence distribution in human alpha satellite higher-order repeats (HOR). Chromosom. Res. 14 (7) (2006) 735–753 (Epub 2006 Nov 22. PubMed PMID:17115529).

L.A. Alexandrov, L.I. Maslovskaya, L.I. Kisselev, L.A. Romanyova, Y.B. Yurkov, A.R. Mushegian, L.L. Kisselev, L.A. Alexandrov, Definition of a new alpha satellite superfamily characterized by monomer organization. Nucleic Acids Res. 21 (1993 May) 2209–2215 (PubMed PMID:8502563; PubMed Central PMCID: PMC309486).

A.J. Wray, H.F. Willard, Evolution of alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. Nucleic Acids Res. 13 (18) (1985 Apr 25) 2731–2741 (PubMed PMID:2987865; PubMed Central PMCID: PMC341190).

T.F. Smith, M.S. Waterman, Identification of common molecular subsequences. J. Mol. Biol. 147 (1) (1981 Mar 25) 195–197 (PubMed PMID:7265238).

O. Gotoh, An improved algorithm for matching biological sequences. J. Mol. Biol. 162 (3) (1982 Dec 15) 705–708 (PubMed PMID:1716870).

H. E. Trowell, A. Nagy, A. Long, Analysis 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 (10) (1993 Oct) 1639–1649 (PubMed PMID:8288917).

H. Chou, M.H. Holmes, DNA sequence quality trimming and vector removal. Bioinformatics 17 (12) (2001) 1093–1104 (PubMed PMID:11751217).

Shepelev VA, Gusev FE, Reshetov DA, Alexandrov AA, Yurov YB, Rogaei EA, Alexandrov IA. The evolution of alpha satellite Suprachromosomal Family 5 from
orangutan to man: a story of invasion, death and rebirth. Molecular Phylogenetics.
Contributions to the 3rd Moscow International Conference "Molecular Phylogenetics" (Moscow, Russia, July 31–August 4, 2012) compiled by A. Troitsky, L. Rusin and V. Aleoshin. Moscow, TORUS Press, p2012–172, [p136–137].

[24] W. Ziccardi, Organization and Composition of the α21-II Region of Human Chromosome 21Masters Thesis 2012 (Available from: http://ecommons.luc.edu/luc_theses/730).

[25] J. Krumsieck, R. Arnold, T. Rattei, Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23 (8) (2007 Apr 15) 1026–1028 (Epub 2007 Feb 19. PubMed PMID: 17309856).

[26] V.A. Shepelev, N.V. Yanishevsky, Multidimensional dot-matrices. Comput. Appl. Biosci. 10 (6) (1994 Dec) 605–611 (Supplement 2).

[27] J.F. Hughes, H. Skaletsky, T. Pyntikova, T.A. Graves, S.K. van Daalen, P.J. Minx, R.S. Fulton, S.D. McGrath, D.P. Locke, C. Friedman, B.J. Trask, E.R. Mardis, W.C. Warren, S. Repping, S. Rozen, R.K. Wilson, D.C. Page, Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature 463 (7280) (2010 Jan 28) 536–539, http://dx.doi.org/10.1038/nature08700 (Epub 2010 Jan 13. PubMed PMID: 20072128; PubMed Central PMCID: PMC3653425).

[28] C. Tyler-Smith, W.R. Brown, Structure of the major block of alphoid satellite DNA on the human Y chromosome. J. Mol. Biol. 195 (3) (1987 Jun 5) 457–470 (PubMed PMID: 2821279).

[29] C. Alkan, M. Ventura, N. Archidiacono, M. Rocchi, S.C. Sahinalp, E.E. Eichler, Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. PLoS Comput. Biol. 3 (9) (2007 Sep) 1807–1818 PubMed PMID: 17907796; PubMed Central PMCID: PMC1994983.

[30] S. Terada, Y. Hirai, H. Hirai, A. Koga, Higher-order repeat structure in alpha satellite DNA is an attribute of hominoids rather than hominids. J. Hum. Genet. 58 (11) (2013 Nov) 752–754, http://dx.doi.org/10.1038/jhg.2013.87 (Epub 2013 Aug 15. PubMed PMID: 23945983).

[31] T.H. Eickbush, D.G. Eickbush, Finely orchestrated movements: evolution of the ribosomal RNA genes. Genetics 175 (2) (2007 Feb) 477–485, http://dx.doi.org/10.1534/genetics.107.071399 (Review, PubMed PMID: 17322354; PubMed Central PMCID: PMC1806002).