Perspective

Short arms of human acrocentric chromosomes and the completion of the human genome sequence

Stylianos E. Antonarakis[1,2,3]

[1]Department of Genetic Medicine and Development, University of Geneva Medical Faculty, 1211 Geneva, Switzerland; [2]Foundation Campus Biotech, 1202 Geneva, Switzerland; [3]Medigenome, Swiss Institute of Genomic Medicine, 1207 Geneva, Switzerland

The complete, ungapped sequence of the short arms of human acrocentric chromosomes (SAACs) is still unknown almost 20 years after the near completion of the Human Genome Project. Yet these short arms of Chromosomes 13, 14, 15, 21, and 22 contain the ribosomal DNA (rDNA) genes, which are of paramount importance for human biology. The sequences of SAACs show an extensive variation in the copy number of the various repetitive elements, the full extent of which is currently unknown. In addition, the full spectrum of repeated sequences, their organization, and the low copy number functional elements are also unknown. The Telomere-to-Telomere (T2T) Project using mainly long-read sequence technology has recently completed the assembly of the genome from a hydatidiform mole, CHM13, and has thus established a baseline reference for further studies on the organization, variation, functional annotation, and impact in human disorders of all the previously unknown genomic segments, including the SAACs. The publication of the initial results of the T2T Project will update and improve the reference genome for a better understanding of the evolution and function of the human genome.

Five human chromosomes have been named as acrocentrics by the 1960 Denver nomenclature report (Lejeune et al. 1960); these are Chromosomes 13, 14, 15, 21, and 22. In these chromosomes, the centromere appeared to be at one end of the chromosomal structure, and from the Greek word *άκρο* (summit or extreme tip), they were named acrocentric. It was recognized, however, that these acrocentric chromosomes do have a short arm (p-arm according to the cytogenetics nomenclature), which peculiarly was of variable size: from almost invisible in some of the acrocentric chromosomes of certain individuals to quite a sizable chromosomal material in other individuals (for more historical notes, see Levan et al. 1964). Cytogeneticists have recognized that the short arms of the five acrocentric chromosomes contain similar sequences and that even a complete deletion of one of these short arms such as in the translocation t(21p;21p) does not have any phenotypic consequence, presumably because the remaining short arms are functionally sufficient. In this review, I refer to the short arms of acrocentric chromosomes as SAACs. In the early description of the SAACs, cytogeneticists recognized three distinct regions: the proximal short arm adjacent to the centromere, the satellite stalk, and the more distal-telomeric region, named satellite. These three regions roughly correspond to the three chromosomal bands of the SAACs p11, p12, and p13, respectively. During the 1960s and 1970s, cytogeneticists drew the acrocentric chromosomes as shown in Figure 1.

Nucleolar organizing regions and rDNA genes

The term nucleolar organizing region (NOR) was first coined by Barbara McClintock in 1934 in plant chromosomes (McClintock 1934). The NORs are chromosomal regions (DNA sequences) around which the nucleoli form; these chromosomal regions map on the SAACs in humans and contain tandem repeats (TRs) of ribosomal RNA genes (rDNA or RNR genes). It has been estimated that humans have approximately 300 copies of RNR genes in the diploid genome distributed on the SAACs, namely, Chromosomes 13, 14, 15, 21, and 22 (Henderson et al. 1972; Schmickel 1973; Stults et al. 2008; Floutsakou et al. 2013).

During metaphase, the upstream binding transcription factor (UBTF, also known as UBF), an HMG-box protein, binds to the RNR gene–containing region of each acrocentric chromosome and provides a marker for "active" NORs. These NORs are the content of the satellite stalks or chromosomal band p12. The binding of UBTF marks the secondary constriction and allows the staining with silver nitrate (AgNORs). The AgNOR staining is a method to visualize the NORs with AgNO3, which was first introduced by Howell et al. (1975) and Denton et al. (1976). The AgNOR staining has been used extensively in cytogenetic studies and diagnosis of chromosomal abnormalities (Fig. 2).

The unit of rDNA in the SAACs is an "operon" of ~13 kb, which contains the gene for 45S transcribed by the RNA polymerase I (Pol I). The mapping of the 45S rDNA to the five acrocentric chromosomes was performed in the early 1970s (Henderson et al. 1972). The names of RNR1, RNR2, RNR3, RNR4, and RNR5 have been given to them for the tandem clusters on chromosomes 13p, 14p, 15p, 21p, and 22p, respectively (OMIM 180450, 180451, 180452, 180453, and 180454).

The 45S rRNA transcript is processed into 18S, 5.8S, and 28S rRNA components. The 18S and 5.8S are separated by the internal transcribed spacer 1 (ITS1) sequence; the 5.8S and 28S, by the ITS2 sequence. Each unit has a 5′ and a 3′ external transcribed spacer (ETS). The tandem units of the 45S rDNA are connected by a 32-kb intergenic spacer (IGS) sequence (Sylvester et al. 1986; Gonzalez and Sylvester 1995). The 18S rRNA is part of the small 40S subunit of ribosomes, whereas the 28S, 5.8S, and 5S rRNAs are incorporated in the large 60S subunit of the eukaryotic

Corresponding author: stylianos.antonarakis@unige.ch
Article published online before print. Article and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.275350.121.

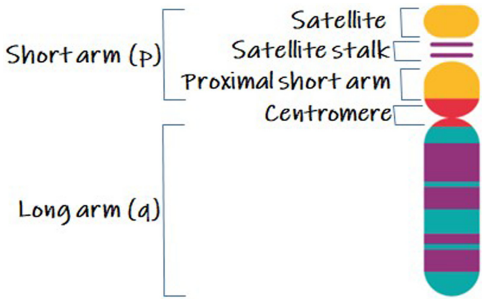

Figure 1. Schematic representation of an acrocentric chromosome and nomenclature of the short arm regions. The satellite stalk contains the rDNA.

ribosomes. Note that the 5S rRNA is encoded by rDNA RNA5S1 on Chromosome 1 in humans (OMIM 180420), which is transcribed by RNA polymerase III (Pol III). Figure 3 schematically shows the rDNA genes, the rRNAs, and their incorporation in the ribosome. Remarkably 70% of the cellular RNA in a eukaryotic cell is rRNA (Warner 1999). Usually about half of the rDNA genes are actively transcribed (Grummt 2003); this transcription is dynamic and fluctuates during the difference stages of the cell cycle (Iyer-Bierhoff and Grummt 2019).

The boundaries of the rDNA cluster have been termed the proximal junction region (PJ) of 207 kb and the distal junction region (DJ) of 380 kb (Floutsakou et al. 2013). The PJ sequences were previously labeled as the Chr 21 short arm pericentromeric region (Lyle et al. 2007). Interestingly, both DJ and PJ can themselves be duplicated. Both PJ and DJ are, as expected, found in all acrocentric human chromosomes. PJ and DJ contain blocks of satellite repeats known as centromeric repeats or CER satellites (Jurka et al. 2005). Other repeats in these junction regions include HSAT5, SATR1, GSATII, ACRO1, ACRO138, and BSA beta. The current understanding of the structure of PJ and DJ, including an inverted duplication within DJ, and initial appreciation of the polymorphic variability can be found in work by Floutsakou et al. (2013) and van Sluis et al. (2019). Gene prediction programs have identified eight putative genes in DJ and four in PJ. Experimental validation of these putative genes has not been performed. Histone marks, however, provide evidence for transcription units within DJ (Floutsakou et al. 2013).

There is extensive variation in the sequences of rDNA copies. Many single-nucleotide variants (SNVs) have been documented since the early 1990s (Leffers and Andersen 1993). In addition, structural variants (SVs) have also been observed since the late 1990s (Gonzalez and Sylvester 2001). "Giant" short arms have been observed in certain individuals. Figure 4 (from Friedrich et al. 1996) shows a short arm of Chromosome 15 that appears as large as the long arm. Pulsed-field gel electrophoresis and preparations of very high molecular-weight genomic DNA were used to determine the length of the rDNA clusters in blood cells in normal volunteers. The cluster length showed extensive variability within and between individuals, ranging from 50 kb to >6 Mb. In addition, there was complete heterozygosity of the length, and each individual showed a unique rDNA electrophoretic pattern (Stults et al. 2008). Using samples from multigeneration families, this study showed a meiotic rearrangement of the length of each cluster with a frequency of >10% per cluster per meiosis (Stults et al. 2008).

An illustrative recent example of a study of the variation of the rDNA sequences used Chromosome 21 p-arm BAC clones and short- and long-read sequences. Analysis of 13 clones revealed many SNVs, short indels, and larger SVs, including partial deletions. An updated reference sequence was proposed that contains TRs, long repeats (LRs), and simple sequence repeats (SSRs) in the IGS (Kim et al. 2018).

The early literature has concluded that there was homogenization among the different copies of the rDNA genes within a chromosome and among chromosomes. There was evidence for inter-chromosomal exchanges (among chromosomes) by unequal crossover (Krystal et al. 1981). Most of these exchanges have been proposed to be the result of the satellite associations that have been observed since the early 1960s in unbanded chromosomes and later, in the mid 1970s, in cultured human leukocytes using silver staining (Fig. 5; Denton et al. 1976). With the use of superresolution microscopy, it was shown recently that rDNA can form linkages between chromosomes. These linkages are coated by UBTF, indicating that they occur between transcriptionally active loci (Potapova et al. 2019).

An increase in the number of specific variant-containing copies of rDNA from one generation to the next has added to the argument of inter-chromosomal exchanges (Schmickel et al. 1985). A remarkable example of inter-chromosomal exchanges is provided by the study of a family in which an unusually large satellite stalk (rDNA sequences) was found on a free Chromosome 22 in a grandfather, a translocated Chromosome 22 in the daughter, and the free Chromosome 21 in the granddaughter (jumping satellites) (Gimelli et al. 1976). Notably, the inter-chromosomal exchanges were concluded to be not only among homologous chromosomes but also, remarkably, among nonhomologous chromosomes. This last conclusion is compatible with the cytogenetic observation of satellite associations of the SAACs (Gonzalez and Sylvester 2001). The rDNA contains extensive regions of short sequence repeat segments, which are known as hotspots for recombination.

On the other hand, there was evidence of intra-chromosomal exchanges (within a chromosome) that could contribute to the homogenization of the rDNA sequences (Seperack et al. 1988).

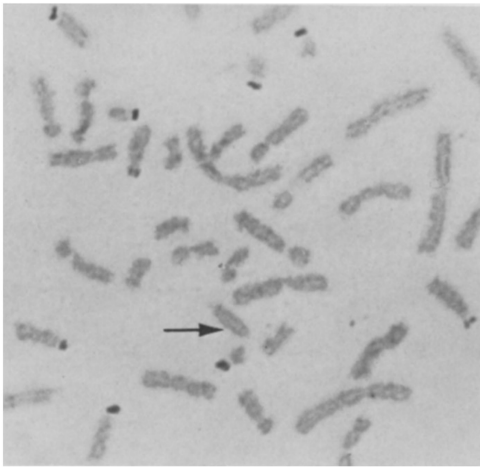

Figure 2. Metaphase chromosome spread with nine differentially stained satellite stalks, which appear as dark areas above the centromere of the acrocentric chromosomes. Arrow points to a D group chromosome, which lacks a satellite staining region. Figure reprinted with permission from Howell et al. (1975).

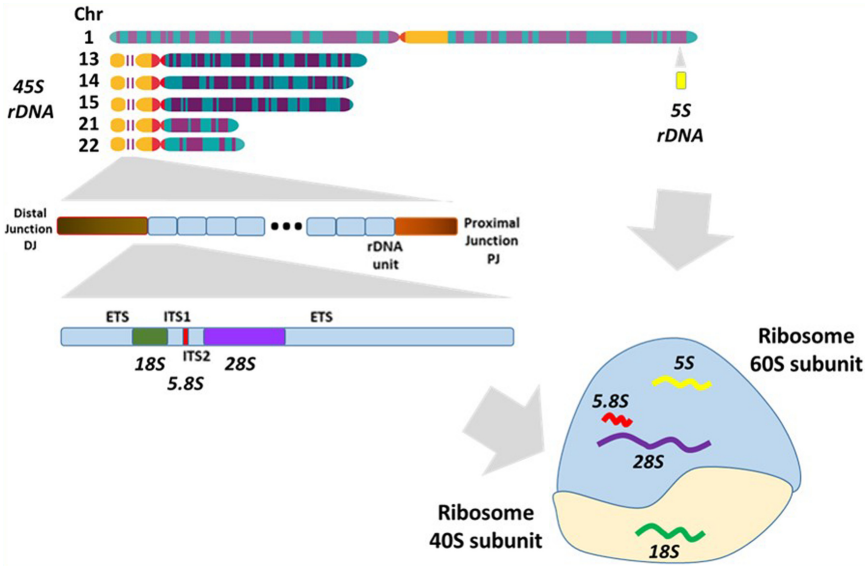

Figure 3. Schematic representation of chromosomes encoding rDNA genes, along with the produced rRNAs, and the incorporation of these RNAs into the ribosome. For detailed description, see text.

Figure 6 provides a schematic representation of inter- and intra-chromosomal exchanges in the SAACs.

A recent study using high-coverage short-read genome sequence (which apparently provides more reliable results than low-coverage sequencing) has estimated the copy number of the 45S genes of the diploid genome in different individuals. The coverage was approximately 35× for Chromosome 1, which was used as a metric. The samples tested were from 86 individuals with European ancestry and 77 with African ancestry, respectively. The copy number of 18S ranged from 115 to 550 copies with a mean of 293 copies and standard deviation of 70. Similar copy number estimates have been obtained for the 5.8S (minimum 113, maximum 562, mean 295, SD 83) and 28S (minimum 83, maximum 426, mean 226, SD 64). In this data set, there was a strong correlation of the values for these three genes ($\rho = 0.97$). Thus, on average there are approximately 30 copies of the rDNA genes in each acrocentric chromosome; however, it remains to be seen if there are substantially unequal numbers of rDNA copies in each acrocentric chromosome. This estimation agrees with earlier estimates from the 1970s (Schmickel 1973). The mean number of each of these genes was higher in the samples of African ancestry than that of European; for example, the mean for the 18S gene was 322 and 282 in Africans and Europeans, respectively (Hall et al. 2021).

Repeated sequences in SAACs other than rDNAs

Several different arrays of TRs have been found that map in the SAACs. Most of them are called satellite repeats, and they provide a major challenge in the assembly of sequence reads. The major classes of these repeated sequences in SAACs that are included in the GRCh38 (hg38) genome assembly are (1) the alpha satellite with a repeat unit of 171 nucleotides (nt; these sequences account for ~2.58% of the human genome) (Waye and Willard 1987); (2) the beta satellite (Waye and Willard 1989) with a repeat unit of 68 nt, accounting for 0.02% of the genome; (3) the gamma satellite with a repeat unit of 220 nt, composing 0.13% of the genome; (4) the HSAT1 with a 42-nt repeat unit, accounting for 0.12% of the genome; (5) the HSAT2 and HSAT3 of a repeat unit of 5 nt, which account for 1.42% of the genome (Jones et al. 1973; Altemose et al. 2014); (6) the ACRO1 of a repeat unit of 147 nt, representing 0.01% of the genome; and (7) the CER of 96-nt unit, accounting for 0.008% of the diploid genome. (Levy et al. 2007). All of these sequences together with the addition of the rDNA gene repeats have been estimated to compose at least 3.54% of the diploid genome or equivalent to 56 megabases, a genome equivalent to a small-size chromosome such as Chromosome 19.

The alpha satellite sequences comprise most of the centromeric regions and spread over the SAACs. The technological advances of long-read sequences and the improvement of the computational methods have provided the opportunity to better understand the composition and architecture of the satellite repeats (Jain et al. 2018). An important and expected observation is the extensive variability of the structure, composition, and length (copy number) of the various satellite repeat arrays in different individuals. For example, the alpha satellites of a pair of X Chromosomes vary in length from 1 Mb to 3.5 Mb; in addition, the composition of the subfamilies of alpha satellite sequences also varies (Miga et al. 2014, 2020; Miga 2015). Figure 7 provides a schematic representation of the extensive variability of the satellite repeats.

In a recent study of more than 800 individuals, the alpha satellite had an estimate of 3.1% median value of the diploid genome, with a range between 1% and 5% (Miga et al. 2014).

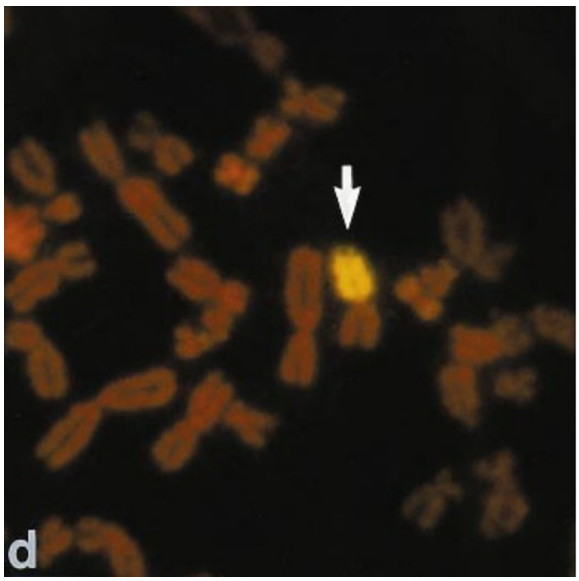

Figure 4. Fluorescence in situ hybridization (FISH) with rDNA. Extreme variant of the short arm of Chromosome 15. Reprinted with permission from Friedrich et al. (1996).

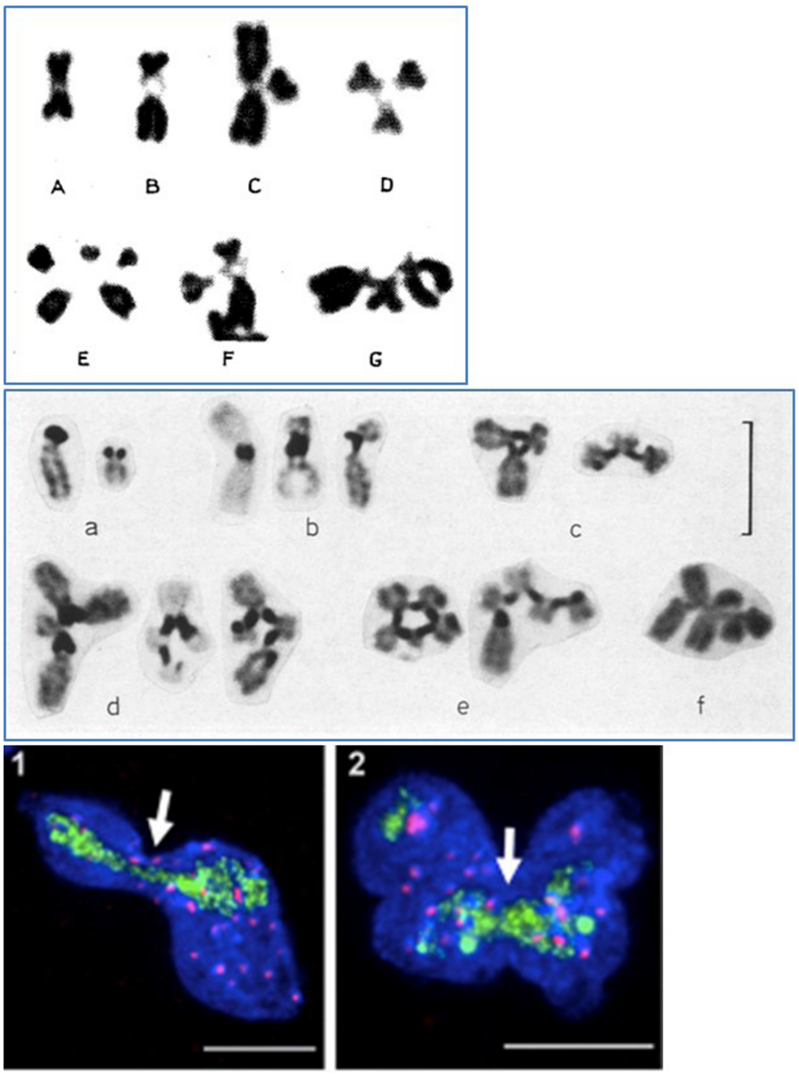

Figure 5. Satellite associations of the short arms of human acrocentric chromosomes. (*Top*) Various types of satellite association from different cells. (*Middle*) Human nucleolar organizer chromosomes: satellite associations. (*Bottom*) Localization of rDNA (green) and centromeres in cells that divided in the presence of topoisomerase inhibitor ICRF-193, showing inter-chromosomal rDNA linkage. Figures reproduced with permission from Ferguson-Smith et al. (1961), Denton et al. (1976), and Potapova et al. (2019), respectively.

Genes other than rDNAs on SAACs

A protein-coding gene named transmembrane phosphatase with tensin homology (*TPTE*) has been identified on Chromosome 21p. The predicted polypeptide of 551 amino acids encoded by *TPTE* has significant homology with tensin and auxilin domains, cyclin-G associated kinase (GAK) and the tumor suppressor PTEN. The gene contains 24 exons and spans 87 kb; the mRNA is ~2.5 kb. On Chromosome 21 the gene maps between the D21Z1 and D21Z4 repeats, and the orientation of the transcript is from the centromere (21cen) to telomere (21pter). Monochromosomal cell hybrids showed the presence of homologous sequences on Chromosomes 13, 15, 22, and Y. The estimated number of copies in the haploid human genome was seven in males and six in females. The gene is highly expressed in testis (Chen et al. 1999; Guipponi et al. 2000). The mouse homologous gene maps to mouse Chromosome 8, which shows synteny to human Chromosome 13q14.2-q21 between *NEK3* and *SUGT1*. The syntenic region on the human chromosome contains a partial, highly divergent copy of *TPTE*, now considered a pseudogene named *TPTE2P2*, that is likely to represent the ancestral copy from which all the other copies of *TPTE* arose through duplication events (Guipponi et al. 2001). Several alternative spliced isoforms have been identified. In the original description, the Chr 21 and Chr 13 transcripts were described as producing a functional transcript, whereas that of Chr 22 was transcribed under the control of an LTR and was predicted not to code for a peptide with transmembrane domains (Tapparel et al. 2003). In the current GRCh38 assembly of the UCSC Genome Browser, the Chr 21 gene sequence is listed as *TPTE* and that of Chr 22 as *TPTEP1*, and it is shown as mapping in the centromeric region. *TPTE2* on Chr 13 is also shown as mapped to the 13q next to the centromere. Several pseudogenes of *TPTE2*, namely, *TPTE2P1*, *TPTE2P2*, *TPTE2P3*, *TPTE2P5*, and *TPTE2P6*, all map in the middle of Chr 13q, whereas *TPTE2P4* maps on Chr Y. The update of the centromeric and short arm regions of the acrocentric chromosomes using long-read sequencing and de novo assembly will clarify the exact mapping positions and determine the potential positional variability of these genes. One of the objectives of the T2T Project is to complete the sequence and the analysis of the acrocentric chromosomes in the near future.

There are several additional potential gene sequences on the SAACs. Interestingly, in the Genome Browser, these sequences are shown only on Chr 21p, probably because there are more studies on BAC sequences of Chr 21p than any other chromosome. Most of these sequences are likely to also map on other acrocentric p-arms too, and a more accurate mapping will await the results of the T2T Project (Miga et al. 2020; Logsdon et al. 2021; https://sites.google.com/ucsc.edu/t2tworkinggroup). These include members of the BAGE family of genes (Boël et al. 1995), tektin pseudogenes, long intergenic non-protein-coding RNA 1667 (*LINC01667*), microRNA 3156-3 (*MIR3156-3*), small nucleolar RNA, H/ACA box 70 (*SNORA70*); the list is not exhaustive. The expression of some of these genes has been tested in the GTEx project (https://gtexportal.org/home/): The *TPTE* gene is exclusively expressed in testis with an average level of 136 transcripts per million (TPM); the *BAGE2* gene is also expressed in the testis with an average level of 2.9 TPM. Similarly, the *LINC01667* gene is expressed in testis with an average level of 10.4 TPM, whereas the *TEKT4P2* pseudogene is ubiquitously expressed with levels ranging from 1.5 to 5.8 TPM in different tissues.

The older sequencing efforts of 1.1 Mb of Chr 21p using BACs and short reads has identified five potential gene models in a BAC

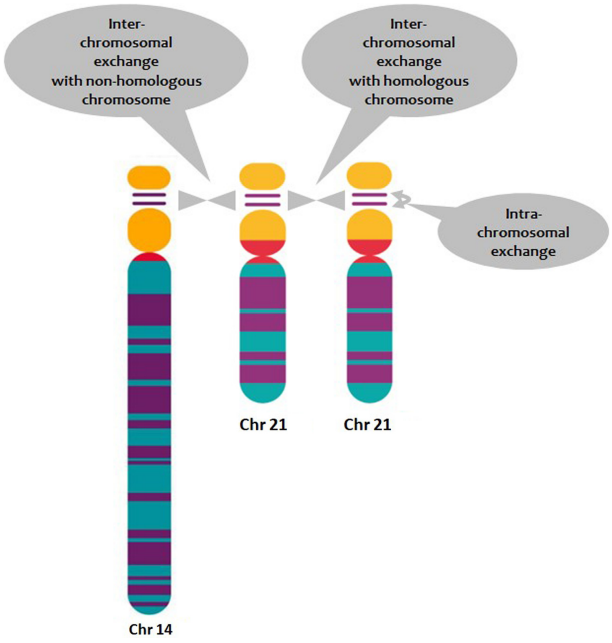

Figure 6. Schematic representation of acrocentric chromosomes that depict the inter-chromosomal exchanges in the rDNA sequences among (1) nonhomologous and (2) homologous chromosomes, and (3) intra-chromosomal exchanges among rDNA sequences within a single chromosome. These exchanges are likely to be the cause of the extensive variation in rDNA copy numbers and the homogenization of these sequences among the alleles of all five acrocentric chromosomes.

proximal to the rDNA repeats. These gene models GM9, GM11, GM10, GM12, and GM28 were shown transcription signals by RT-PCR in a panel of 24 tissues (Lyle et al. 2007).

Clinical significance and disorders related to sequences within the SAACs

Robertsonian translocations

Robertsonian translocations are special kinds of chromosomal translocations in which two acrocentric chromosomes are joined with breakpoints and junction sequences in their short arms. These were first described by W.R.B. Robertson in 1916 in insect speciation (Robertson 1916). Robertsonian translocations could occur among nonhomologous chromosomes (e.g., Chromosomes 14 and 21) or among homologous chromosomes (between, e.g., Chromosomes 21). Robertsonian translocations are among the most common chromosomal rearrangements in humans. A series of cytogenetic studies of more than 110,000 karyotypes of both spontaneous abortions and liveborn individuals showed a frequency of Robertsonian translocations of 0.9–1.2 per 1000 individuals (Hamerton et al. 1975; Jacobs 1981; Nielsen and Wohlert 1991). The most common is the Robertsonian translocation between Chromosomes 13 and 14, which accounts for ~75% of all Robertsonian translocations. The translocations t13;21, t14;21, t15;21, and t21;22 in the parents provide an increased risk for trisomy in the offspring. Furthermore, a t21;21 carrier could only produce children with trisomy 21. Chromosome 13 is involved in 81% of Robertsonian translocations, whereas Chromosomes 14, 15, 21, and 22 are involved in 89%, 9%, 14%, and 6% of such translocations, respectively (Therman et al. 1989). These translocations could occur in oogonial/spermatogonial mitosis, in meiotic prophase I, in the zygote, or in the early postzygotic mitotic divisions. Robertsonian translocations likely occur because of the similarity/identity of sequences in the SAACs and inappropriate recombination events between two nonhomologous chromosomes during the satellite associations. The fusion, nonhomologous Robertsonian translocation chromosome usually has two centromeres (dicentric) in ~90% of the cases (Blouin et al. 1994). In many instances of homologous Robertsonian translocations, dicentric chromosomes have also been found (Shaffer et al. 1991). Of note is that the homologous Robertsonian translocation could be either isochromosomes or the fusion of two different homologs (Blouin et al. 1994). An early study has concluded that the breakpoints in Robertsonian translocations occur preferentially in repetitive DNA, which is located between the satellite III and the rDNA (Gravholt et al. 1992). In a similar later study, most of the breakpoints in Robertsonian translocations 13;14 and 14;21 were between repetitive sequences TRI-6 (subfamily of satellite I) and rDNA on Chromosome 13p, between TRS-47 and TRS-63 (subfamilies of satellite III) on Chromosome 14, and between TRI-6 and rDNA on Chromosome 21 (Page et al. 1996).

The true nature of Robertsonian translocations, the exact recombination mechanism, and the phenotypic correlations could

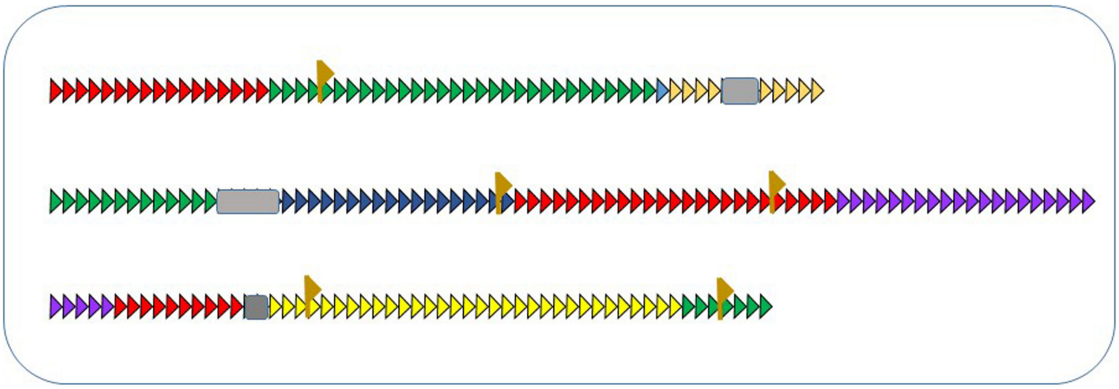

Figure 7. Schematic representation of the variation of satellite repeats in the pericentromeric and short arm regions of human acrocentric chromosomes. Three different representative polymorphic alleles are shown. The unit of the repeat is shown as an arrowhead; the color of each repeat represents the different groups of sequence variation: Gray boxes represent nonsatellite sequences; brown flags depict inserted transposable elements. The figure is inspired by Miga (2015).

be elucidated using the genomic structure from the results of the T2T Project.

Variation of rDNAs

The variation in the total number of the rDNA copies is possibly associated with phenotypic variation and/or disease states, particularly in individuals with an extremely low or high number of rDNA copies (Kampen et al. 2020). In addition, the SNVs of the resulting rDNA copies may result in heterogeneous ribosomes (Parks et al. 2018) with variation in their function. Genomic instability of the rDNA has been reported in Bloom syndrome and ataxia-telangiectasia (Parks et al. 2018), both conditions with increased cancer risk. Somatic losses of rDNA copies have been described in several tumor types (Xu et al. 2017; Udugama et al. 2018; Wang and Lemos 2019). Thus, the nucleotide and copy number variation of rDNA may become an important contributor to phenotypic variation.

Methylation of rDNA, as well as rRNA sequences

Several studies have shown that a considerable fraction of the copies of the rDNAs are methylated. The methylation level of rDNA is strongly associated with age, linking the biological age to nucleolar biology (Wang and Lemos 2019). In another study, methylation of the rDNA transcription unit including the upstream control element (UCE), core promoter, 18S rDNA, and 28S rDNA in human sperm also significantly increased with donor's age (Potabattula et al. 2020).

The rRNA is also heavily modified, and among these modifications, the most prominent is the methylation of adenosine at position 6. The biological significance of this and other modifications is under investigation in human disorders, including cancer (Barbieri and Kouzarides 2020).

Expression of satellite repeats and nearby genes

Satellite sequences could be transcribed, and the functional consequences of this transcription are inadequately studied and poorly understood. Alpha-satellite expression, for example, occurs through RNA polymerase II–dependent transcription. Single-molecule fluorescence in-situ hybridization (smFISH) detects alpha-satellite RNA transcripts in intact human cells. The levels of alpha-satellite RNA vary across cell lines and over the cell cycle (Bury et al. 2020). The topology of this transcription and the spatial relationship of the nucleolus and centromeres need further study.

Ectopic insertion of satellite repeats

The insertion of 18 monomeric (~68-bp) beta-satellite repeat units was inserted in the *TMPRSS3* gene caused one form of autosomal recessive congenital deafness (DFNB10) (Scott et al. 2001). Thus, sequences from the SAACs could "jump" and reinsert in a protein-coding gene outside of these SAACs. The mobile nature of repetitive sequences on SAACs is well documented (Farrell et al. 1993). Circular extrachromosomal molecules present in many eukaryotic cells, small polydisperse circular DNAs (spcDNA), may contain beta-satellites (Assum et al. 1993) or other repeats on the SAACs (Gaubatz 1990); these are likely to be produced by unequal homologous recombination between or within repetitive sequences. The insertion into *TMPRSS3* in the DFNB10 family may have arisen by recombination of spcDNA containing beta-satellites with a region of minimal homology spanning exon 11 of *TMPRSS3*. The complete sequence of the SAACs will provide insights into the events of ectopic insertion of certain sequences.

The completion and analyses of the SAAC sequences through the T2T Project; promises and challenges

The use of long-read sequences of tens or hundreds of kilobases (for review, see Logsdon et al. 2020) now provides the opportunity for the first time to investigate the structure of the SAAC. Currently, the v1.0 T2T assembly that includes the completed acrocentric chromosomes has been deposited in GenBank (https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.2). The recent publication using long-read and strand-specific sequencing technologies to study SVs in a cohort of 32 genomes did not resolve the sequences of the SAACs (Ebert et al. 2021). Computational methods for de

Figure 8. Schematic representation of the two alternative hypotheses regarding some DNA sequences in the SAACs. The alternative on the *left* depicts the situation in which there are no chromosome-specific sequences in the short arms; the alternative on the *right* includes chromosome-specific sequences shown as bars of different colors in the short arms of each acrocentric chromosome.

novo assembly have been developed to capture and visualize the complexity and the variability of the sequence assemblies. The genomes of hydatidiform moles have been used to assemble the complete haplotype of DNA sequences (Steinberg et al. 2014). Such moles originate from a single sperm that has undergone postmeiotic chromosomal duplication; these genomes are, therefore, uniformly homozygous for one set of alleles. The hydatidiform mole CHM13, with stable chromosomal content in culture, is used for the telomere-to-telomere sequence of Chromosomes X and 8 (Miga et al. 2020; Logsdon et al. 2021). The genomes of the moles simplify the establishment of the single and continued haplotype without the complication of the presence of the second homologous chromosome.

The promises and expectations of the T2T collaborative project regarding the SAACs could be briefly summarized below:

1. Establish the linear structure and nucleotide composition of these five SAAC regions (Chromosomes 13p, 14p, 15p, 21p, 22p).
2. Identify candidate transcribed sequences in the SAACs (both coding and noncoding) and elucidate their copy number and mapping in one or more genomic locations.
3. Identify other functional genomic elements of low-copy number and initially define their potential functional significance.
4. Identify, within the SAACs, sequences specific to each acrocentric chromosome. It is not clear if such sequences exist, and the T2T Project could provide the knowledge infrastructure for further population-based investigations. Figure 8 provides the two extreme alternative hypotheses regarding the existence of chromosome specific sequences within the SAACs.
5. Identify novel classes of repetitive elements localized primarily in the SAACs. The structure of these elements may inform potential function or may provide the reagents for further studies. In addition, a revision and update of the nomenclature of the different satellite sequences will greatly facilitate the communications in the genomic communities.
6. Provide a more accurate total length of the whole genome and each chromosome. An indication of the common variation of the genome length could be also provided after the T2T sequence of the genomes of different individuals from a wide variety of geoethnic ancestry.

The potential challenges include the following:

1. It is possible that there is no such thing as chromosome-specific short arm sequences, because of the extensive exchanges of sequences at the cellular and population level. The study of the variation of sequences within the SAACs will require extensive population studies using long-read methodologies that should become financially affordable, computational methods available to the scientific community, and sharing of data in appropriate databases.
2. Development of methods to establish continuous haplotypes in the context of the diploid genome are needed so that the extent of the sequence variability could be better understood, as well as the dynamics of recombinational exchanges and other mechanisms generating structural sequence variation.
3. The functional analysis of the low-copy sequences, and the high-copy repeats will be important in order to understand the potential involvement of the SAACs in human cellular and organismal biology.
4. That the involvement of SAACs in the phenotypic diversity may include Mendelian and complex traits is a medical objective that cannot be underestimated. The current knowledge of SAAC sequences in disease stats is abysmally low, and it is expected that some phenotypic traits could be related or caused by pathogenic variability in some of the SAAC sequences.

Competing interest statement

The author declares no competing interests.

Acknowledgments

I thank Drs. Adam Phillippy of the National Human Genome Research Institute of the National Institutes of Health and Evan Eichler of the University of Washington, Seattle, for the expert and constructive comments on this manuscript. I also thank the ChildCare Foundation for partial financial support, as well as the three expert reviewers for the opportunity to improve the original manuscript.

References

Altemose N, Miga KH, Maggioni M, Willard HF. 2014. Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Comput Biol* **10:** e1003628. doi:10.1371/journal.pcbi.1003628

Assum G, Fink T, Steinbeisser T, Fisel KJ. 1993. Analysis of human extrachromosomal DNA elements originating from different β-satellite subfamilies. *Hum Genet* **91:** 489–495. doi:10.1007/BF00217778

Barbieri I, Kouzarides T. 2020. Role of RNA modifications in cancer. *Nat Rev Cancer* **20:** 303–322. doi:10.1038/s41568-020-0253-2

Blouin JL, Binkert F, Antonarakis SE. 1994. Biparental inheritance of chromosome 21 polymorphic markers indicates that some Robertsonian translocations t(21;21) occur postzygotically. *Am J Med Genet* **49:** 363–368. doi:10.1002/ajmg.1320490333

Boël P, Wildmann C, Sensi ML, Brasseur R, Renauld JC, Coulie P, Boon T, van der Bruggen P. 1995. BAGE: a new gene encoding an antigen recognized on human melanomas by cytolytic T lymphocytes. *Immunity* **2:** 167–175. doi:10.1016/S1074-7613(95)80053-0

Bury L, Moodie B, Ly J, McKay LS, Miga KH, Cheeseman IM. 2020. α-Satellite RNA transcripts are repressed by centromere–nucleolus associations. *eLife* **9:** e59770. doi:10.7554/eLife.59770

Chen H, Rossier C, Morris MA, Scott HS, Gos A, Bairoch A, Antonarakis SE. 1999. A testis-specific gene, TPTE, encodes a putative transmembrane tyrosine phosphatase and maps to the pericentromeric region of human chromosomes 21 and 13, and to chromosomes 15, 22, and Y. *Hum Genet* **105:** 399–409. doi:10.1007/s004390051122

Denton TE, Howell WM, Barrett JV. 1976. Human nucleolar organizer chromosomes: satellite associations. *Chromosoma* **55:** 81–84. doi:10.1007/BF00288330

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372:** eabf7117. doi:10.1126/science.abf7117

Farrell SA, Winsor EJ, Markovic VD. 1993. Moving satellites and unstable chromosome translocations: clinical and cytogenetic implications. *Am J Med Genet* **46:** 715–720. doi:10.1002/ajmg.1320460624

Ferguson-Smith MA, Handmaker SD, Hopkins ABJ. 1961. Observations on the satellited human chromosomes. *Lancet* **277:** 638–640. doi:10.1016/S0140-6736(61)91655-5

Floutsakou I, Agrawal S, Nguyen TT, Seoighe C, Ganley AR, McStay B. 2013. The shared genomic architecture of human nucleolar organizer regions. *Genome Res* **23:** 2003–2012. doi:10.1101/gr.157941.113

Friedrich U, Caprani M, Niebuhr E, Therkelsen AJ, Jørgensen AL. 1996. Extreme variant of the short arm of chromosome 15. *Hum Genet* **97:** 710–713. doi:10.1007/BF02346177

Gaubatz JW. 1990. Extrachromosomal circular DNAs and genomic sequence plasticity in eukaryotic cells. *Mutat Res* **237:** 271–292. doi:10.1016/0921-8734(90)90009-G

Gimelli G, Porro E, Santi F, Scappaticci S, Zuffardi O. 1976. "Jumping" satellites in three generations: a warning for paternity tests and prenatal diagnosis. *Hum Genet* **34:** 315–318. doi:10.1007/BF00295297

Gonzalez IL, Sylvester JE. 1995. Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics* **27:** 320–328. doi:10.1006/geno.1995.1049

Gonzalez IL, Sylvester JE. 2001. Human rDNA: evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. *Genomics* **73:** 255–263. doi:10.1006/geno.2001.6540

Gravholt CH, Friedrich U, Caprani M, Jørgensen AL. 1992. Breakpoints in Robertsonian translocations are localized to satellite III DNA by fluorescence in situ hybridization. *Genomics* **14:** 924–930. doi:10.1016/S0888-7543(05)80113-2

Grummt I. 2003. Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. *Genes Dev* **17:** 1691–1702. doi:10.1101/gad.1098503R

Guipponi M, Yaspo ML, Riesselman L, Chen H, De Sario A, Roizès G, Antonarakis SE. 2000. Genomic structure of a copy of the human TPTE gene which encompasses 87 kb on the short arm of chromosome 21. *Hum Genet* **107:** 127–131. doi:10.1007/s004390000343

Guipponi M, Tapparel C, Jousson O, Scamuffa N, Mas C, Rossier C, Hutter P, Meda P, Lyle R, Reymond A, et al. 2001. The murine orthologue of the Golgi-localized TPTE protein provides clues to the evolutionary history of the human TPTE gene family. *Hum Genet* **109:** 569–575. doi:10.1007/s004390100607

Hall AN, Turner TN, Queitsch C. 2021. Thousands of high-quality sequencing samples fail to show meaningful correlation between 5S and 45S ribosomal DNA arrays in humans. *Sci Rep* **11:** 449. doi:10.1038/s41598-020-80049-y

Hamerton JL, Canning N, Ray M, Smith S. 1975. A cytogenetic survey of 14,069 newborn infants. I. Incidence of chromosome abnormalities. *Clin Genet* **8:** 223–243. doi:10.1111/j.1399-0004.1975.tb01498.x

Henderson AS, Warburton D, Atwood KC. 1972. Location of ribosomal DNA in the human chromosome complement. *Proc Natl Acad Sci* **69:** 3394–3398. doi:10.1073/pnas.69.11.3394

Howell WM, Denton TE, Diamond JR. 1975. Differential staining of the satellite regions of human acrocentric chromosomes. *Experientia* **31:** 260–262. doi:10.1007/BF01990741

Iyer-Bierhoff A, Grummt I. 2019. Stop-and-go: dynamics of nucleolar transcription during the cell cycle. *Epigenet Insights* **12:** 2516865719849090. doi:10.1177/2516865719849090

Jacobs PA. 1981. Mutation rates of structural chromosome rearrangements in man. *Am J Hum Genet* **33:** 44–54.

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018. Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* **36:** 321–323. doi:10.1038/nbt.4109

Jones KW, Prosser J, Corneo G, Ginelli E. 1973. The chromosomal location of human satellite DNA III. *Chromosoma* **42:** 445–451. doi:10.1007/BF00399411

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462–467. doi:10.1159/000084979

Kampen KR, Sulima SO, Vereecke S, De Keersmaecker K. 2020. Hallmarks of ribosomopathies. *Nucleic Acids Res* **48:** 1013–1028. doi:10.1093/nar/gkz637

Kim JH, Dilthey AT, Nagaraja R, Lee HS, Koren S, Dudekula D, Wood WH Iii, Piao Y, Ogurtsov AY, Utani K, et al. 2018. Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. *Nucleic Acids Res* **46:** 6712–6725. doi:10.1093/nar/gky442

Krystal M, D'Eustachio P, Ruddle FH, Arnheim N. 1981. Human nucleolus organizers on nonhomologous chromosomes can share the same ribosomal gene variants. *Proc Natl Acad Sci* **78:** 5744–5748. doi:10.1073/pnas.78.9.5744

Leffers H, Andersen AH. 1993. The sequence of 28S ribosomal RNA varies within and between human cell lines. *Nucleic Acids Res* **21:** 1449–1455. doi:10.1093/nar/21.6.1449

Lejeune J, Levan A, Böök JA, Chu EHY, Ford CE, Fraccaro M, Harnden DG, Hsu TC, Hungerford DA, Jacobs PA, et al. 1960. A proposed standard system of nomenclature of human mitotic chromosomes. *The Lancet* **275:** 1063–1065. doi:10.1016/S0140-6736(60)90948-X

Levan A, Fredga K, Sandberg AA. 1964. Nomenclature for centromeric position on chromosomes. *Hereditas* **52:** 201–220. doi:10.1111/j.1601-5223.1964.tb01953.x

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5:** e254. doi:10.1371/journal.pbio.0050254

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21:** 597–614. doi:10.1038/s41576-020-0236-x

Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, et al. 2021. The structure, function, and evolution of a complete human chromosome 8. *Nature* **593:** 101–107. doi:10.1038/s41586-021-03420-7

Lyle R, Prandini P, Osoegawa K, ten Hallers B, Humphray S, Zhu B, Eyras E, Castelo R, Bird CP, Gagos S, 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. doi:10.1101/gr.6675307

McClintock B. 1934. The relation of a particular chromosomal element to the development of the nucleoli in Zea mays. *Zeitschrift für Zellforschung und Mikroskopische Anatomie* **21:** 294–326. doi:10.1007/BF00374060

Miga KH. 2015. Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Res* **23:** 421–426. doi:10.1007/s10577-015-9488-2

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. doi:10.1101/gr.159624.113

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585:** 79–84. doi:10.1038/s41586-020-2547-7

Nielsen J, Wohlert M. 1991. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. *Hum Genet* **87:** 81–83. doi:10.1007/BF01213097

Page SL, Shin JC, Han JY, Choo KH, Shaffer LG. 1996. Breakpoint diversity illustrates distinct mechanisms for Robertsonian translocation formation. *Hum Mol Genet* **5:** 1279–1288. doi:10.1093/hmg/5.9.1279

Parks MM, Kurylo CM, Dass RA, Bojmar L, Lyden D, Vincent CT, Blanchard SC. 2018. Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Sci Adv* **4:** eaao0665. doi:10.1126/sciadv.aao0665

Potabattula R, Zacchini F, Ptak GE, Dittrich M, Müller T, El Hajj N, Hahn T, Drummer C, Behr R, Lucas-Hahn A, et al. 2020. Increasing methylation of sperm rDNA and other repetitive elements in the aging male mammalian germline. *Aging Cell* **19:** e13181. doi:10.1111/acel.13181

Potapova TA, Unruh JR, Yu Z, Rancati G, Li H, Stampfer MR, Gerton JL. 2019. Superresolution microscopy reveals linkages between ribosomal DNA on heterologous chromosomes. *J Cell Biol* **218:** 2492–2513. doi:10.1083/jcb.201810166

Robertson WRB. 1916. Chromosome studies. I. Taxonomic relationships shown in the chromosomes of tettigidae and acrididae: V-shaped chromosomes and their significance in Acrididae, Locustidae, and Gryllidae: chromosomes and variation. *J Morphol* **27:** 179–331. doi:10.1002/jmor.1050270202

Schmickel RD. 1973. Quantitation of human ribosomal DNA: hybridization of human DNA with ribosomal RNA for quantitation and fractionation. *Pediatr Res* **7:** 5–12. doi:10.1203/00006450-197301000-00002

Schmickel RD, Gonzalez IL, Erickson JM. 1985. Nucleolus organizing genes on chromosome 21: recombination and nondisjunction. *Ann N Y Acad Sci* **450:** 121–131. doi:10.1111/j.1749-6632.1985.tb21488.x

Scott HS, Kudoh J, Wattenhofer M, Shibuya K, Berry A, Chrast R, Guipponi M, Wang J, Kawasaki K, Asakawa S, et al. 2001. Insertion of β-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness. *Nat Genet* **27:** 59–63. doi:10.1038/83768

Seperack P, Slatkin M, Arnheim N. 1988. Linkage disequilibrium in human ribosomal genes: implications for multigene family evolution. *Genetics* **119:** 943–949. doi:10.1093/genetics/119.4.943

Shaffer LG, Jackson-Cook CK, Meyer JM, Brown JA, Spence JE. 1991. A molecular genetic approach to the identification of isochromosomes of chromosome 21. *Hum Genet* **86:** 375–382. doi:10.1007/BF00201838

Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, Shiryev SA, Morgulis A, Surti U, Warren WC, et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res* **24:** 2066–2076. doi:10.1101/gr.180893.114

Stults DM, Killen MW, Pierce HH, Pierce AJ. 2008. Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res* **18:** 13–18. doi:10.1101/gr.6858507

Sylvester JE, Whiteman DA, Podolsky R, Pozsgay JM, Respess J, Schmickel RD. 1986. The human ribosomal RNA genes: structure and organization of the complete repeating unit. *Hum Genet* **73:** 193–198. doi:10.1007/BF00401226

Tapparel C, Reymond A, Girardet C, Guillou L, Lyle R, Lamon C, Hutter P, Antonarakis SE. 2003. The *TPTE* gene family: cellular expression, subcellular localization and alternative splicing. *Gene* **323:** 189–199. doi:10.1016/j.gene.2003.09.038

Therman E, Susman B, Denniston C. 1989. The nonrandom participation of human acrocentric chromosomes in Robertsonian translocations. *Ann Hum Genet* **53:** 49–65. doi:10.1111/j.1469-1809.1989.tb01121.x

Udugama M, Sanij E, Voon HPJ, Son J, Hii L, Henson JD, Chan FL, Chang FTM, Liu Y, Pearson RB, et al. 2018. Ribosomal DNA copy loss and repeat

instability in ATRX-mutated cancers. *Proc Natl Acad Sci* **115:** 4737–4742. doi:10.1073/pnas.1720391115

van Sluis M, Gailín MO, McCarter JGW, Mangan H, Grob A, McStay B. 2019. Human NORs, comprising rDNA arrays and functionally conserved distal elements, are located within dynamic chromosomal regions. *Genes Dev* **33:** 1688–1701. doi:10.1101/gad.331892.119

Wang M, Lemos B. 2019. Ribosomal DNA harbors an evolutionarily conserved clock of biological aging. *Genome Res* **29:** 325–333. doi:10.1101/gr.241745.118

Warner JR. 1999. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* **24:** 437–440. doi:10.1016/S0968-0004(99)01460-7

Waye JS, Willard HF. 1987. Nucleotide sequence heterogeneity of α satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res* **15:** 7549–7569. doi:10.1093/nar/15.18.7549

Waye JS, Willard HF. 1989. Human β satellite DNA: genomic organization and sequence definition of a class of highly repetitive tandem DNA. *Proc Natl Acad Sci* **86:** 6250–6254. doi:10.1073/pnas.86.16.6250

Xu B, Li H, Perry JM, Singh VP, Unruh J, Yu Z, Zakari M, McDowell W, Li L, Gerton JL. 2017. Ribosomal DNA copy number loss and sequence variation in cancer. *PLoS Genet* **13:** e1006771. doi:10.1371/journal.pgen.1006771

Short arms of human acrocentric chromosomes and the completion of the human genome sequence

Stylianos E. Antonarakis

Genome Res. 2022 32: 599-607 originally published online March 31, 2022
Access the most recent version at doi:10.1101/gr.275350.121