

- 10 Gu, X. (2003) Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet.* 19, 354–356
- 11 Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–65
- 12 Chen, F.C. and Li, W.H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68, 444–456
- 13 Ebersberger, I. *et al.* (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* 70, 1490–1497
- 14 Bailey, J.A. *et al.* (2002) Recent segmental duplications in the human genome. *Science* 297, 1003–1007
- 15 Yunis, J.J. and Prakash, O. (1982) The origin of man: a pictorial legacy. *Science* 215, 1525–1530
- 16 Navarro, A. and Barton, N.H. (2003) Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes. *Science* 300, 321–324
- 17 Navarro, A. *et al.* (2003) Response to comment on ‘Chromosomal speciation and molecular divergence – Accelerated evolution in rearranged chromosomes’. *Science* 302, 988
- 18 Lu, J. *et al.* (2003) Comment on ‘Chromosomal speciation and molecular divergence – Accelerated evolution of genes in rearranged chromosomes’. *Science* 302, 988
- 19 Vieira, J. *et al.* (2001) Evidence for selection at the fused1 locus of *Drosophila americana*. *Genetics* 158, 279–290
- 20 Rieseberg, L.H. *et al.* (2000) Hybridization, introgression and linkage evolution. *Plant Mol. Biol.* 42, 205–224
- 21 Noor, M.A.F. *et al.* (2001) Chromosomal inversions and the persistence of species. *Proc. Natl. Acad. Sci. U. S. A.* 98, 12084–12088
- 22 Zhang, J. *et al.* (2004) Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome Res.* 14, 845–851
- 23 Williams, E.J.B. and Hurst, L.D. (2000) The proteins of linked genes evolve at similar rates. *Nature* 407, 900–902
- 24 Navarro, A. and Barton, N.H. (2003) Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution Int. J. Org. Evolution* 57, 447–459
- 25 Hey, J. (2003) Speciation and inversions: chimps and humans. *BioEssays* 25, 825–828
- 26 Spitz, F. *et al.* (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* 113, 405–417
- 27 Phippard, D. *et al.* (2000) The *sex-linked fidget* mutation abolishes *Brn4/Pou3f4* gene expression in the embryonic inner ear. *Hum. Mol. Genet.* 9, 79–86
- 28 Tanimoto, K. *et al.* (1999) Effects of altered gene order or orientation of the locus control region on human-globin gene expression in mice. *Nature* 398, 344–348
- 29 Puig, M. *et al.* (2004) Silencing of a gene adjacent to the breakpoint of a widespread *Drosophila* inversion by a transposon-induced antisense RNA. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9013–9018

0168-9525/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.  
doi:10.1016/j.tig.2004.08.009

# Analysis of the centromeric regions of the human genome assembly

M. Katharine Rudd and Huntington F. Willard

Institute for Genome Sciences and Policy, Duke University, Durham, NC 27710, USA

**The sequence of the human genome is not yet complete, and major gaps remain at the centromere region of each chromosome, which is comprised of repetitive  $\alpha$  satellite DNA. In this article, we describe the sequences in the vicinity of the centromere that are included in the current genome assembly, analyze the ~7 Mb of  $\alpha$  satellite that have been assembled thus far and anticipate the nature of the sequences that remain to be accounted for.**

The centromere of most complex eukaryotic chromosomes is a specialized locus comprising repetitive DNA that is responsible for chromosome segregation during mitosis and meiosis [1,2]. Normal human centromeres consist of megabases of  $\alpha$  satellite DNA, a repeat family containing ~171-bp monomers [3]. These monomers can be arranged either in a highly homogeneous, multimeric organization or in a more heterogeneous monomeric form that lacks this higher-order periodicity [4–6]. Despite their obvious functional significance, centromeric regions and their constituent  $\alpha$  satellite sequences were largely omitted by the Human Genome Project because of their repetitive nature and the expected paucity of genes [7]; the reported

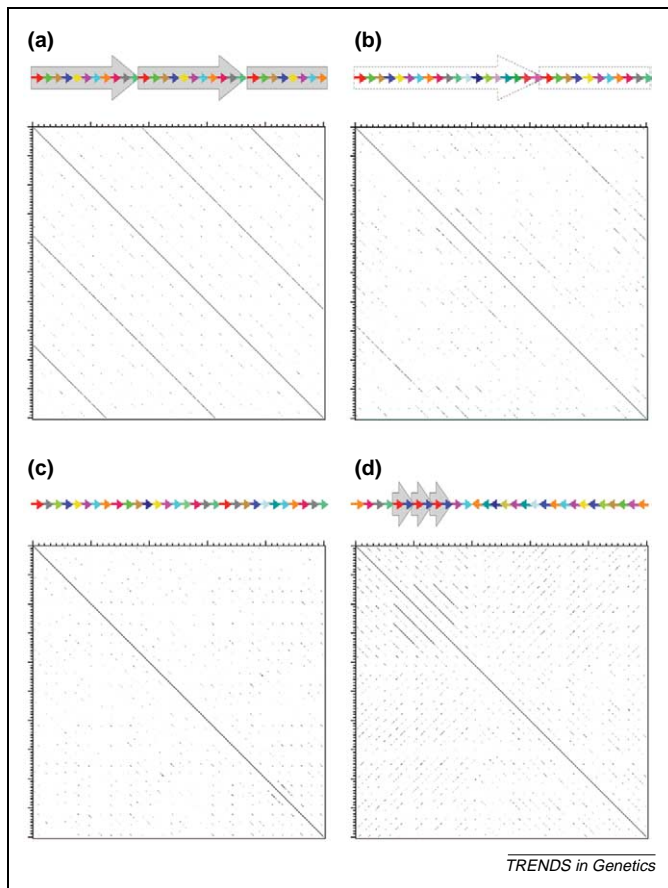
assemblies [8,9] of each chromosome arm thus end an uncertain distance from the functional centromere [10]. Although such regions are often considered to be difficult to sequence, in fact it is the assembly, not the sequencing itself, which presents a challenge because of the high degree of sequence homogeneity among many hundreds or thousands of copies of a given repeated sequence.

## Alpha satellite organization and function

Alpha satellite DNA has been identified at every human centromere [5,6]; however, among reported chromosome assemblies, the amount and type of  $\alpha$  satellite varies. There are two major types of  $\alpha$  satellite DNA: higher-order and monomeric [4,5] (Figure 1). Higher-order  $\alpha$  satellite DNA consists of ~171-bp monomers organized in arrays of multimeric repeat units that are highly homogeneous (typically 97–100% identical); by contrast, monomeric  $\alpha$  satellite DNA lacks any higher-order periodicity, and its monomers are only on average ~70% identical [11]. In addition to their different sequence organization, monomeric and higher-order  $\alpha$  satellite DNA also differ in their functionality. Higher-order  $\alpha$  satellite DNA was shown to be associated with centromere function on the basis of genomic [10,12], biochemical [13,14] and artificial

Corresponding author: Huntington F. Willard (hunt.willard@duke.edu).

Available online 11 September 2004



**Figure 1.** Four types of  $\alpha$  satellite DNA are apparent in the current assembly of the human genome. DOTTER plots of 5 kb of  $\alpha$  satellite compared with itself are shown for each type of  $\alpha$  satellite. **(a)** Highly homogeneous higher-order  $\alpha$  satellite consists of multimeric repeat units that are 97%–100% identical. Higher-order repeat units (shown by colored arrow heads) are organized in tandem arrays that typically have a uniform repeat unit size and can span several megabases [5,17]. In Build 34 of the current genome assembly, nine chromosome arms have reached higher-order  $\alpha$  satellite of this type, a total of ~200 kb of sequence. Our analysis of 73 higher-order repeat units shows that within chromosome arm contigs, higher-order repeat unit identity ranges from  $97.5\% \pm 0.5\%$  (8q; n=9 higher-order repeats) to  $99.3\% \pm 0.4\%$  (19p, n=6 higher-order repeats), with an average of  $98.4\% \pm 0.5\%$  identical. Between different chromosome arrays, however, higher-order repeats are divergent; this has been well documented previously [3,5]. **(b)** Other higher-order  $\alpha$  satellite shows clear evidence of multimeric structure; however, these multimeric units are less regular and more divergent in sequence and are ~82%–100% identical, with an average identity of  $93.7\% \pm 2.6\%$ . Seven chromosome assemblies contain this kind of higher-order  $\alpha$  satellite, comprising ~100 kb of the current genome assembly. **(c)** Monomeric  $\alpha$  satellite lacks any evidence of higher-order periodicity [11,36]. Its monomers are ~50%–100% identical to one another, with an average pairwise percent identity of  $71.6\% \pm 8.3\%$ . **(d)** Short zones of multimeric, highly homogeneous  $\alpha$  satellite (<1–10 kb) have been found in the middle of larger expanses of monomeric  $\alpha$  satellite. Although not part of a larger higher-order array, tandem repeat units within these zones are highly homogeneous (98%–100% identical). Such zones were predicted to arise via local homogenization events, thus giving rise to transition states in the early stages of sequence family homogenization [24,37].

chromosome assays [10,15,16]. By contrast, there is no evidence for the direct involvement of monomeric  $\alpha$  satellite DNA in centromere function [16].

Higher-order  $\alpha$  satellite DNA is the predominant type in the genome, present in megabase quantities at each centromere [3,5,17]. Where it has been studied, monomeric  $\alpha$  satellite DNA lies at the edges of higher-order arrays and is less abundant [10,11,18]. This expectation notwithstanding, the vast majority of  $\alpha$  satellite DNA in the current assembly is of the monomeric type (discussed

in the following section), reflecting the currently incomplete nature of centromeric contigs.

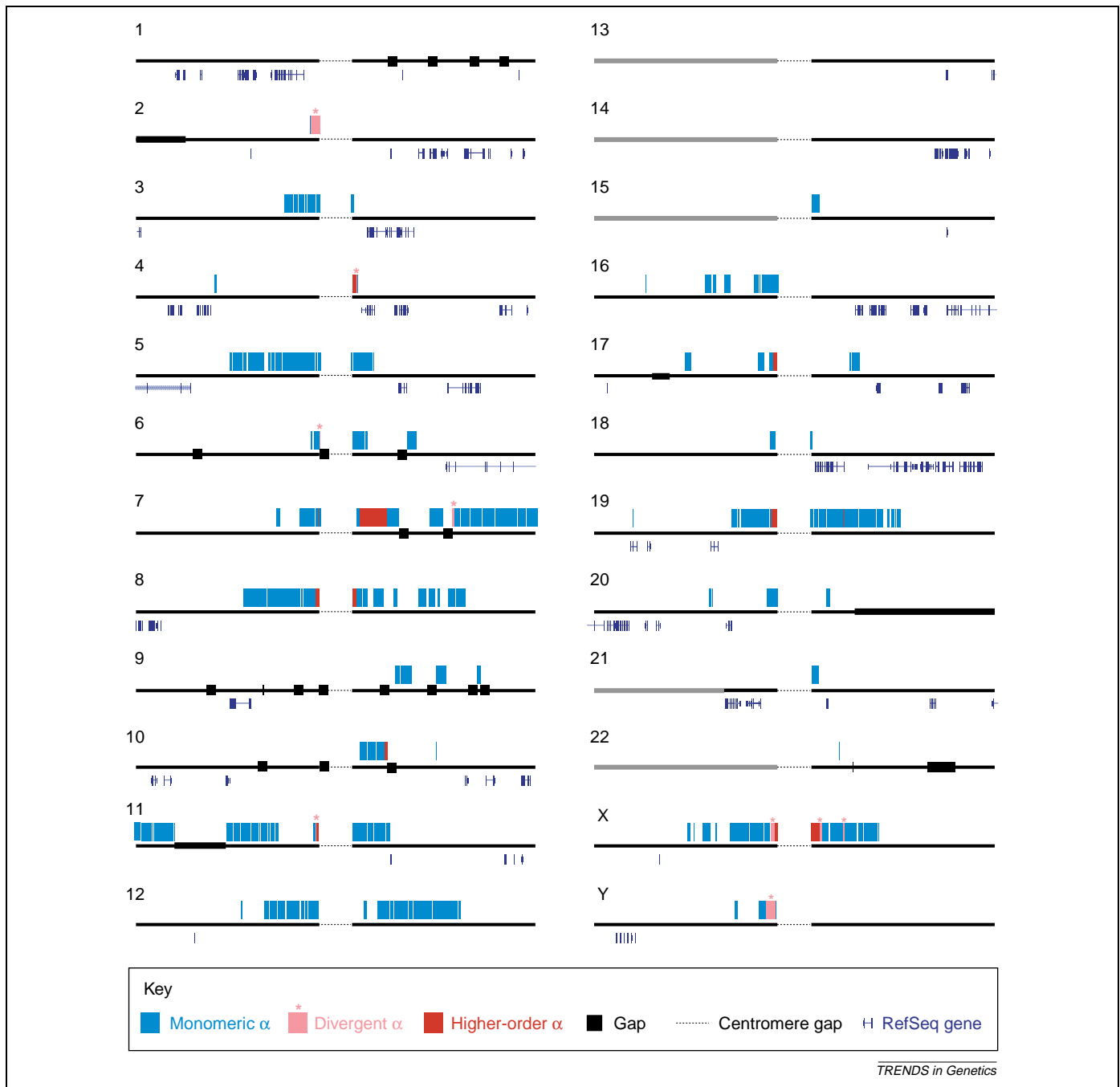
### Alpha satellite in the genome assembly

Despite the difficulty of assembling  $\alpha$  satellite DNA and the lack of specific attention to the centromere regions for most chromosomes, several chromosome assemblies do include  $\alpha$  satellite DNA in their contigs. The July 2003 (Build 34; <http://genome.ucsc.edu/>) assembly of the human genome contains 6.6 Mb of a satellite (an estimated ten times under-representation [19,20]), of which 5.7 Mb lie within the most centromere proximal megabase of the reported p and q arm contigs, adjacent to the centromere gaps. As expected, there is a sharp drop in  $\alpha$  satellite content outside of the first megabase (supplementary Table 1 online). To annotate the major  $\alpha$  satellite regions of the reported genome assembly, we focused on the most proximal megabase of each chromosome arm. The validation of the sequence assembly of centromeric regions remains an important goal for future work; nonetheless, general features of the reported contigs have been confirmed in several instances, by long-range pulsed field gel mapping [10,21] and by fluorescence *in situ* hybridization (M.K. Rudd, unpublished).

Alpha satellite content adjacent to the centromere gap varies widely among chromosomes (Figure 2). Of the 43 chromosome arm assemblies examined (the five acrocentric chromosomes contain heterochromatic short arms and are not represented in the current genome assembly), nine assemblies have not reached any  $\alpha$  satellite, suggesting that these contigs are a substantial distance from the centromere. Only six chromosomes have >100 kb of  $\alpha$  satellite assembled on both p and q arm contigs; the longest reported assembly on any chromosome is only 836 kb (supplementary Table 1 online), substantially less than the amount known to be located at each centromere on the basis of earlier molecular, cytogenetic and genomic studies [5,6,17]. It is likely that this variation in coverage reflects the assembly progress on particular chromosomes rather than interchromosome differences in  $\alpha$  satellite organization.

To characterize the types of  $\alpha$  satellite in the current assembly, we used a combination of BLAST and DOTTER alignment tools (supplementary data online and Figure 1). Using this analysis, >92% of  $\alpha$  satellite DNA in the current assembly is of the monomeric type, and in only 11 chromosomes (chromosomes 2, 4, 6, 7, 8, 10, 11, 17, 19, X and Y) have assemblies reached into higher-order centromere proximal  $\alpha$  satellite (Figure 2). Notably, even within this limited dataset, four of these assemblies contain previously undescribed families of higher-order  $\alpha$  satellite (see supplementary data online), suggesting that the complete set of centromeric repeats in the human genome has yet to be revealed.

In our analysis of  $\alpha$  satellite DNA, in the current genome assembly, we found two categories of higher-order  $\alpha$  satellite DNA that differ in the degree and extent of sequence homogeneity (Figure 1). Within a region of higher-order  $\alpha$  satellite DNA on any one chromosome arm, the most homogeneous higher-order repeat units are 97%–100% identical. The mean percentage identity



TRENDS in Genetics

**Figure 2.** Alpha satellite and genes located in the first megabase that is adjacent to the centromere gap for each chromosome arm. Data are from the July 2003 genome assembly (Build 34). RefSeq genes (purple) are shown below the black line. Monomeric  $\alpha$  satellite (blue), typical higher-order  $\alpha$  satellite (red) and more divergent higher-order  $\alpha$  satellite (pink, with asterisks) are shown above the black line. Broken lines indicate centromere gaps, whereas other gaps in the reported assembly are shown as black boxes. Acrocentric short arms (omitted from the assembly) are indicated by grey lines. Chromosomes 1, 9 and 16 have heterochromatic q arm regions adjacent to the centromere, and the corresponding heterochromatin gaps are included in the centromere gaps to simplify the figure. For a more detailed version of this figure, see the supplementary data online.

among tandem higher-order repeat units within such regions is  $98.4\% \pm 0.5\%$  (Figure 1a), reflecting their concerted evolution and consistent with previous estimates of intra-array sequence homogeneity in the human genome [10,22,23]. However, other higher-order repeat units in the assembly lack the regular organization and consistent higher-order repeat length that is characteristic of highly homogeneous tandem arrays. Their less highly homogenized repeats, although clearly multimeric, are more divergent in both sequence

and structure, with a pairwise mean identity of  $93.6\% \pm 2.6\%$  (Figure 1b).

The nature of this second category of higher-order  $\alpha$  satellite DNA in the genome is itself probably heterogeneous. In some cases, these repeats correspond to diverged copies at the edges of an otherwise homogeneous array [10]; in other cases, they might represent vestiges of ancient arrays that are no longer present in the genome (or at least not represented in the assembled portion). In addition to higher-order repeats contained in long arrays,

several assemblies contain evidence of short (<1–10 kb) ‘islands’ that are characterized by local homogeneity (>98% identity between tandem multimers) within a region of otherwise monomeric  $\alpha$  satellite DNA (Figure 1d). Presumably, such regions reflect the small numbers of recent homogenizing events, as predicted by evolutionary models [24], and might thus represent the earliest stages of the appearance of new arrays.

Only two chromosome assemblies have reached arrays of highly homogeneous higher-order  $\alpha$  satellite DNA on both p and q arm contigs (Figure 2). Because all chromosomes are known to contain higher-order  $\alpha$  satellites at their centromeres [5,6], the fact that only the assemblies of chromosome 8 and the X chromosome have had this level of success indicates that most current assemblies probably terminate some distance from the functional centromere. In the two cases where there is higher-order  $\alpha$  satellite DNA on both p and q arms, the repeats are oriented in the same direction on both arms (supplementary Figure 1 online), consistent with them being part of the same homogeneous tandem array [5]. By contrast, within the heterogeneous monomeric arrays, the orientation of  $\alpha$  satellite DNA typically switches several times within each arm contig [10] (supplementary Figure 1 online).

In addition to the blocks of  $\alpha$  satellite adjacent to the centromere gaps in the genome assembly, there are also smaller regions of  $\alpha$  satellite that do not appear to be near centromeres. In total, we found 133 blocks of  $\alpha$  satellite located >5Mb from the centromere gaps. Although the largest of these could represent ancient inversions or other chromosomal rearrangements involving centromere regions [25,26], there are 60 blocks containing <1kb of  $\alpha$  satellite DNA. Such  $\alpha$  satellite blocks could represent assembly errors, library contamination or real occurrences of  $\alpha$  satellite far away from the centromere. Two lines of evidence argue for the legitimacy of at least some of these small blocks of  $\alpha$  satellite. First, we validated a subset of the blocks by PCR and sequencing in 20 unrelated individuals (M.K. Rudd, unpublished), indicating that at least these segments of non-centromeric  $\alpha$  satellite are legitimate. Second, 39 of the 60 blocks lie within 10 bp of a transposable element, consistent with their spread to non-centromeric locations via a transduction mechanism or an unequal crossover event involving *Alu* elements or L1 repetitive sequences adjacent to  $\alpha$  satellite [27] (M.K. Rudd, unpublished). Interestingly, 13 such blocks of non-centromeric  $\alpha$  satellite were found within the introns of validated genes, demonstrating that at least small stretches of  $\alpha$  satellite are not detrimental to gene expression.

### Centromeric landscape

Centromeric or pericentromeric regions have been described historically as home of the ‘junk DNA’ of the genome [28,29]. Recent studies have indicated a relatively sharp transition between the euchromatin of chromosome arms and the satellite-containing region near the centromere [10,18,21], raising the possibility that some genes are located close to  $\alpha$  satellites [20]. Indeed, there are 104 genes listed in the Reference Sequence collection

(<http://www.ncbi.nih.gov/RefSeq>) within the most proximal 1-Mb regions of the 43 chromosome arm contigs (Figure 2), an average gene density of 2.5 genes per Mb. Although this density is lower than the genome-wide average of  $\sim$ 7.5 genes per Mb, it is not substantially different from densities reported for some (gene-poor) chromosomes, such as chromosomes 13 [30] and 21 [31].

The most proximal segments of the chromosome arms are also full of segmental duplications [32], potentially accounting in part for the difficulty of assembling these regions [33]. In the current assembly, 14.9% of the most proximal 1-Mb regions are part of segmental duplications and are >98% identical to another region of the genome (supplementary Figure 1 online). An emerging model is that segments rich in segmental duplications define some of the pericentromeric regions of the genome distal to  $\alpha$  satellites, whereas the centromeric region itself consists of  $\alpha$  satellite DNA and is expected to be largely devoid of such duplications.

Other repeats besides  $\alpha$  satellites are also enriched at the centromere. Using RepeatMasker (<http://repeatmasker.genome.washington.edu>), we examined the repeat content of the combined 43 most proximal 1-Mb regions adjacent to the centromere gaps, and compared them with the genome average. The genome as a whole and the most centromere proximal regions were 49% and 64% repetitive, respectively (supplementary Table 2 online). This enrichment in repeat content near the centromere gaps is almost entirely due to a >40-fold increase in satellite DNA. Although the majority of these satellite sequences consist of  $\alpha$  satellite DNA, other satellites are also significantly more frequent at the edges of the contigs, compared with the genome average (supplementary Table 2 online). These other types of satellite sequence lie just distal of  $\alpha$  satellite or in some cases are interspersed among blocks of monomeric  $\alpha$  satellite [10,19] (supplementary Figure 1 online). Similar to the situation with segmental duplications, a high density of these non- $\alpha$  satellites might be features of the pericentromere rather than the centromere *sensu stricto*.

### Concluding remarks and future outlook

The centromere is a crucial functional part of our genome; however, its complex repetitive organization and the assumption that it contains nothing but ‘junk DNA’ made it a logical region to omit from assembly strategies, which can be frustrated by high levels of sequence homogeneity and/or the extensive polymorphism that has been described for  $\alpha$  satellite arrays [5,10,17]. Given the current status of the assembly as analyzed here, the next phase of genome and centromere annotation might consider a targeted strategy to complete the contigs of each chromosome arm until they reach higher-order arrays of  $\alpha$  satellite associated with centromere function [10,20]. Such a strategy, similar in some respects to that used successfully to assemble the highly complex and repetitive Y chromosome sequence [34], could build on the evident heterogeneity of monomeric repeats [10,11,18] and the periodic variants that punctuate the otherwise homogeneous arrays of higher-order  $\alpha$  satellites [5]. As in the case of the Y chromosome, the sequencing of a single

haplotype would probably facilitate the correct assembly of the centromeric regions because polymorphisms in  $\alpha$  satellite DNA could confound the assembly process. Such an assembly, in concert with parallel analyses at human telomeres [35], will provide the underlying sequence data that are necessary for complete annotation of elements required for human chromosome structure and function and will move the genome one step closer to true completion.

### Acknowledgements

We thank E. Eichler for providing access to unpublished data. We thank E. Eichler, M. Schueler and D. Ledbetter for helpful discussions, and Patrick McConnell for assistance. This work was supported by a research grant from the March of Dimes Birth Defects Foundation and by the Duke University Institute for Genome Sciences and Policy. M.K.R. was a predoctoral student at Case Western Reserve University, Cleveland, OH, USA.

### Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2004.08.008

### References

- Cleveland, D.W. *et al.* (2003) Centromeres and kinetochores. From epigenetics to mitotic checkpoint signaling. *Cell* 112, 407–421
- Sullivan, B.A. *et al.* (2001) Determining centromere identity: cyclical stories and forking paths. *Nat. Rev. Genet.* 2, 584–596
- Willard, H.F. and Wayne, J.S. (1987) Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* 3, 192–198
- Willard, H.F. (1991) Evolution of alpha satellite. *Curr. Opin. Genet. Dev.* 1, 509–514
- Warburton, P. and Willard, H. (1996) Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In *Human Genome Evolution* (Jackson, M.S.T. and Dover, G. eds), pp. 121–145, BIOS Scientific Publishers
- Alexandrov, I. *et al.* (2001) Alpha-satellite DNA of primates: old and new families. *Chromosoma* 110, 253–266
- Collins, F.S. *et al.* (1998) New goals for the U.S. Human Genome Project: 1998–2003. *Science* 282, 682–689
- Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
- Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
- Schueler, M.G. *et al.* (2001) Genomic and genetic definition of a functional human centromere. *Science* 294, 109–115
- Wevrick, R. *et al.* (1992) Structure of DNA near long tandem arrays of alpha satellite DNA at the centromere of human chromosome 7. *Genomics* 14, 912–923
- Spence, J.M. *et al.* (2002) Co-localization of centromere activity, proteins and topoisomerase II within a subdomain of the major human X alpha-satellite array. *EMBO J.* 21, 5269–5280
- Vafa, O. and Sullivan, K.F. (1997) Chromatin containing CENP-A and alpha satellite DNA is a major component of the inner kinetochore plate. *Curr. Biol.* 7, 897–900
- Ando, S. *et al.* (2002) CENP-A, -B, and -C chromatin complex that contains the I-type alpha-satellite array constitutes the prekinetochore in HeLa cells. *Mol. Cell. Biol.* 22, 2229–2241
- Harrington, J.J. *et al.* (1997) Formation of *de novo* centromeres and construction of first-generation human artificial microchromosomes. *Nat. Genet.* 15, 345–355
- Ikeno, M. *et al.* (1998) Construction of YAC-based mammalian artificial chromosomes. *Nat. Biotechnol.* 16, 431–439
- Wevrick, R. and Willard, H.F. (1989) Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: High frequency array-length polymorphism and meiotic stability. *Proc. Natl. Acad. Sci. U. S. A.* 86, 9394–9398
- Horvath, J.E. *et al.* (2000) Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11. *Hum. Mol. Genet.* 9, 113–123
- Guy, J. *et al.* (2003) Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. *Genome Res.* 13, 159–172
- Eichler, E.E. *et al.* (2004) An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* 5, 345–354
- Rudd, M.K. *et al.* (2004) Sequence organization and functional annotation of human centromeres. *Cold Spring Harb. Symp. Quant. Biol.* 68, 141–149
- Durfy, S.J. and Willard, H.F. (1989) Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: Evidence for short range homogenization of tandemly repeated DNA sequences. *Genomics* 5, 810–821
- Schindelbauer, D. and Schwarz, T. (2002) Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous alpha-satellite DNA array. *Genome Res.* 12, 1815–1826
- Dover, G. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* 299, 111–117
- Baldini, A. *et al.* (1993) An alphoid DNA sequence conserved in all human and great ape chromosomes: evidence for ancient centromeric sequences at human chromosomal regions 2q21 and 9q13. *Hum. Genet.* 90, 577–583
- Yunis, J.J. and Prakash, O. (1982) The origin of man: a chromosomal pictorial legacy. *Science* 215, 1525–1530
- Deininger, P.L. *et al.* (2003) Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* 13, 651–658
- Doolittle, W.F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601–603
- Orgel, L.E. and Crick, F.H. (1980) Selfish DNA: the ultimate parasite. *Nature* 284, 604–607
- Dunham, A. *et al.* (2004) The DNA sequence and analysis of human chromosome 13. *Nature* 428, 522–528
- Hattori, M. *et al.* (2000) The DNA sequence of human chromosome 21. *Nature* 405, 311–319
- She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* (in press)
- Bailey, J.A. *et al.* (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* 11, 1005–1017
- Skaletsky, H. *et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423, 825–837
- Riethman, H. *et al.* (2004) Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome Res.* 14, 18–28
- Ikeno, M. *et al.* (1994) Distribution of CENP-B boxes reflected in CREST centromere antigenic sites on long-range alpha-satellite DNA arrays of chromosome 21. *Hum. Mol. Genet.* 3, 1245–1257
- Smith, G.P. (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* 191, 528–553