5. M. Raveendran *et al.*, *Genomics* **88**, 706 (2006).
6. E. Delson, in *The Macaques: Studies in Ecology, Behavior, and Evolution*, D. D. Lindburg, Ed. (van Nostrand Rheinhold, New York, 1980), pp. 10–30.
7. C. Abegg, B. Thierry, *Biol. J. Linn. Soc.* **75**, 555 (2002).
8. D. G. Smith, J. McDonough, *Am. J. Primatol.* **65**, 1 (2005).
9. J. Viray, B. Rolfs, D. G. Smith, *Comp. Med.* **51**, 555 (2001).
10. B. Ferguson *et al.*, *BMC Genom.* **8**, 43 (2007).
11. ENCODE Project Consortium, *Science* **306**, 636 (2004).
12. ENCODE regions were chosen because they have been widely studied across several mammals, including rhesus and baboon.
13. Materials and methods are available as supporting material on *Science* Online.
14. D. Falush, M. Stephens, J. K. Pritchard, *Genetics* **164**, 1567 (2003).
15. A. L. Price *et al.*, *Nat. Genet.* **38**, 904 (2006).
16. K. Hayasaka, K. Fujii, S. Horai, *Mol. Biol. Evol.* **13**, 1044 (1996).
17. J. C. Morales, D. J. Melnick, *J. Hum. Evol.* **34**, 1 (1998).
18. M. A. Eberle, M. J. Rieder, L. Kruglyak, D. A. Nickerson, *PLoS Genet.* **2**, 1319 (2006).
19. L. Kruglyak, *Nat. Genet.* **22**, 139 (1999).
20. J. Fooden, in *The Macaques: Studies in Ecology, Behavior, and Evolution*, D. D. Lindburg, Ed. (van Nostrand Rheinhold, New York, 1980), pp. 1–9.
21. HapMap, *Nature* **437**, 1299 (2005).

22. We thank the Yerkes, Oregon, and California National Primate Research Centers for contributing samples, and D. G. Torgerson for comments. Funded by NIH grant RR05090 to D.G.S., NIH RR00163 to B.F., NIH RR015383 to J.R., NSF0516310 to C.D.B., and 1R01HG003229 to C.D.B., R.N., A. G. Clark, and T. Mattise. Trace Index numbers are consecutively numbered from 1664051535 to 1664070335 and can be retrieved using the following query: PROJECT_NAME='ENCODE' STRATEGY='Re-sequencing' TRACE_TYPE_CODE='PCR' SPECIES_CODE='MACACA MULATTA'.

26 January 2007; accepted 16 March 2007
10.1126/science.1140462

REPORT

Evolutionary Formation of New Centromeres in Macaque

Mario Ventura,[1]* Francesca Antonacci,[1]* Maria Francesca Cardone,[1] Roscoe Stanyon,[2] Pietro D'Addabbo,[1] Angelo Cellamare,[1] L. James Sprague,[3] Evan E. Eichler,[3] Nicoletta Archidiacono,[1] Mariano Rocchi[1]†

A systematic fluorescence in situ hybridization comparison of macaque and human synteny organization disclosed five additional macaque evolutionary new centromeres (ENCs) for a total of nine ENCs. To understand the dynamics of ENC formation and progression, we compared the ENC of macaque chromosome 4 with the human orthologous region, at 6q24.3, that conserves the ancestral genomic organization. A 250-kilobase segment was extensively duplicated around the macaque centromere. These duplications were strictly intrachromosomal. Our results suggest that novel centromeres may trigger only local duplication activity and that the absence of genes in the seeding region may have been important in ENC maintenance and progression.

Evolutionary new centromeres (ENCs) can appear during evolution in a novel chromosomal region with concomitant inactivation of the old centromere. The new centromere then becomes fixed in the species while inevitably progressing toward the complexity typical of a mature mammalian centromere, with intra- and interchromosomal pericentromeric segmental duplications and a large core of satellite DNA (*1*). Unambiguous examples of ENCs were initially reported in primates (*2*) and then described in various other mammalian orders (*3*). A similar phenomenon, well known from clinical cases, is the mitotic rescue of an acentric chromosomal fragment by the opportunistic de novo emergence of a neocentromere (*4*). Recently, two cases of neocentromeres in normal individuals with otherwise normal karyotypes were fortuitously discovered (*5*, *6*). These two "in progress" centromeres can be regarded as ENCs at the initial stage, thus reinforcing the opinion that ENCs and clinical neocentromeres are two faces of the same coin. The goal of the research presented here was to gain insight into the processes and mechanisms of ENC evolution. First, we systematically compared macaque and human synteny organization in search of ENCs. Then, we characterized in detail a macaque ENC and compared it to the orthologous domain in humans, which represents the ancestral genomic structure before ENC seeding.

Multicolor hybridization on rhesus macaque chromosomes [*Macaca mulatta* (MMU) $2n$ = 42, where n is the haploid number of chromosomes] of about 500 evenly spaced human bacterial artificial chromosome (BAC) clones revealed that seven macaque/human homologs (chromosomes 6/5, 8/8, 11/12, 17/13, 19/19, 20/16, and X/X, respectively) were colinear when the position of the centromere was excluded . However, human chromosomes 7/21, 14/15, and 20/22 form syntenic associations as part of three compound macaque chromosomes (3, 7, and 10, respectively). Differences in marker order between macaque and humans were accounted for by 20 chromosome rearrangements. Reiterative fluorescence in situ hybridization (FISH) experiments with additional BAC clones more precisely defined rearrangement breakpoints (table S1). A summary of all results is graphically displayed at www.biologia.uniba.it/macaque. Tables S2 and S3 provide a comprehensive list of the ~600 clones that were used in FISH experiments, from the perspective of both macaque and human chromosomes, respectively.

This comprehensive marker-order comparison revealed that the centromeres of many orthologous chromosomes were embedded in different genomic contexts. To distinguish whether ENC events had occurred in the human or macaque ortholog, or in both, we took into account previous reports that attempted to establish the ancestral form for each chromosome (*2*, *3*, *6*–*15*). (Most of these papers use a different macaque chromosome nomenclature; here we follow the nomenclature used by the macaque genome sequencing consortium. For a comparison, see www.biologia.uniba.it/macaque.) The

Table 1. Macaque chromosomes with neocentromeres. The two noncontiguous positions defining, in human, the ENC of chromosome 1 are due to the colocalization of the ENC with a macaque-specific inversion breakpoint.

MMU (HSA)	Clones	Position of the neocentromere on the human sequence	Reference
1 (1)	RP4-621O15	chr1:226,810,735–226,866,653	Present study
	RP11-572K18	chr1:160,918,751–161,035,790	
2 (3)	RP11-355I21/RP11-418B12	chr3:163,822,353–164,707,155	(*6*)
4 (6)	RP11-474A9	chr6:145,651,644–145,845,896	(*9*)
12 (2q)	RP11-343I5/RP11-846E22	chr2:138,659,884–138,908,673	Present study
13 (2p)	RP11-722G17	chr2:86,622,638–86,827,260	Present study
14 (11)	RP11-625D10/RP11-661M13	chr11:5,667,339–6,043,020	(*10*)
15 (9)	RP11-542K23/RP11-64P14	chr9:124,189,785–124,493,134	Present study
17 (13)	RP11-543A19/RP11-527N12	chr13:61,111,769–62,699,203	(*3*)
18 (18)	RP11-61D1/RP11-289E15	chr18:50,155,761–50,526,34	Present study

[1]Department of Genetics and Microbiology, University of Bari, 70126 Bari, Italy. [2]Department of Animal Biology and Genetics, University of Florence, Florence 50125, Italy. [3]Howard Hughes Medical Institute, Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: rocchi@biologia.uniba.it

results of this analysis confirmed the previously published results and exposed five macaque ENCs (Fig. 1 and Table 1). In total between macaque and human, there are 14 ENCs; 9 ENCs occurred in the macaque lineage [MMU1 (1), 2 (3), 4 (6), 12 (2q), 13 (2p), 14 (11), 15 (9), 17 (13), and 18 (18) (corresponding human chromosomes in parentheses)], and 5 occurred in the human lineage [HSA3 (2), 6 (4), 11 (14), 14 (7a), and 15 (7b) (corresponding macaque chromosomes in parentheses), where HSA denotes *Homo sapiens*]. The newly discovered macaque ENCs were found on MMU1 (1), 12 (2q), 13 (2p), 15 (9), and 18 (18) (corresponding human chromosomes in parentheses). In this context, all macaque centromeres, including the nine ENCs, harbor very large arrays of alpha satellite DNA (*16*) (fig. S1). One possibility is that after their emergence, new macaque centromeres were rapidly stabilized by acquiring alpha satellite DNA.

Human chromosome 6 and the macaque homolog, MMU4, both have ENCs. The ancestral centromere for both species was located at HSA6p22.1 (*9*) (Fig. 2), and the new macaque centromere is located at HSA6q24.3. A comparison of the HSA6q24.3 region [chr6: base pair 139,100,001 to 149,100,000; University of California Santa Cruz (UCSC) March 2006 release] with the orthologous regions of dog, rat, mouse, and opossum genomes, by careful inspection of the specific alignment "Net" in the UCSC genome browser (http://genome.ucsc.edu), showed that a reasonable assumption was that the human region closely resembled the ancestral condition. We reasoned that a detailed comparison of the organization of the MMU4 centromeric-pericentromeric region with the organization of the human counterpart at 6q24.3 might allow us to examine hypotheses of the formation and progression of ENCs.

Human BAC RP11-474A9 (L2 in Fig. 2) mapping at chr6:145,651,644 to 145,845,896 yielded an apparently splitting signal around the MMU4 centromere (*9*). It was therefore considered to be the probable seeding point, and the FISH analysis of flanking markers was consistent with this conclusion (Fig. 3). The construction of a BAC contig spanning the MMU4 centromere started, therefore, from the L2 locus and is reported in detail in the supporting online material (SOM) text. To briefly summarize this construction, appropriate human sequence tagged sites (STSs), mapping within 1 megabase from both sides of the MMU4 centromere, were used to screen high-density filters of the macaque BAC library CH250, segment 1 (http://bacpac.chori.org). FISH analysis of these BACs showed that some of them were duplicated on both sides of the MMU4 centromere. The sequencing of BAC ends and of appropriate polymerase chain reaction (PCR) products and FISH

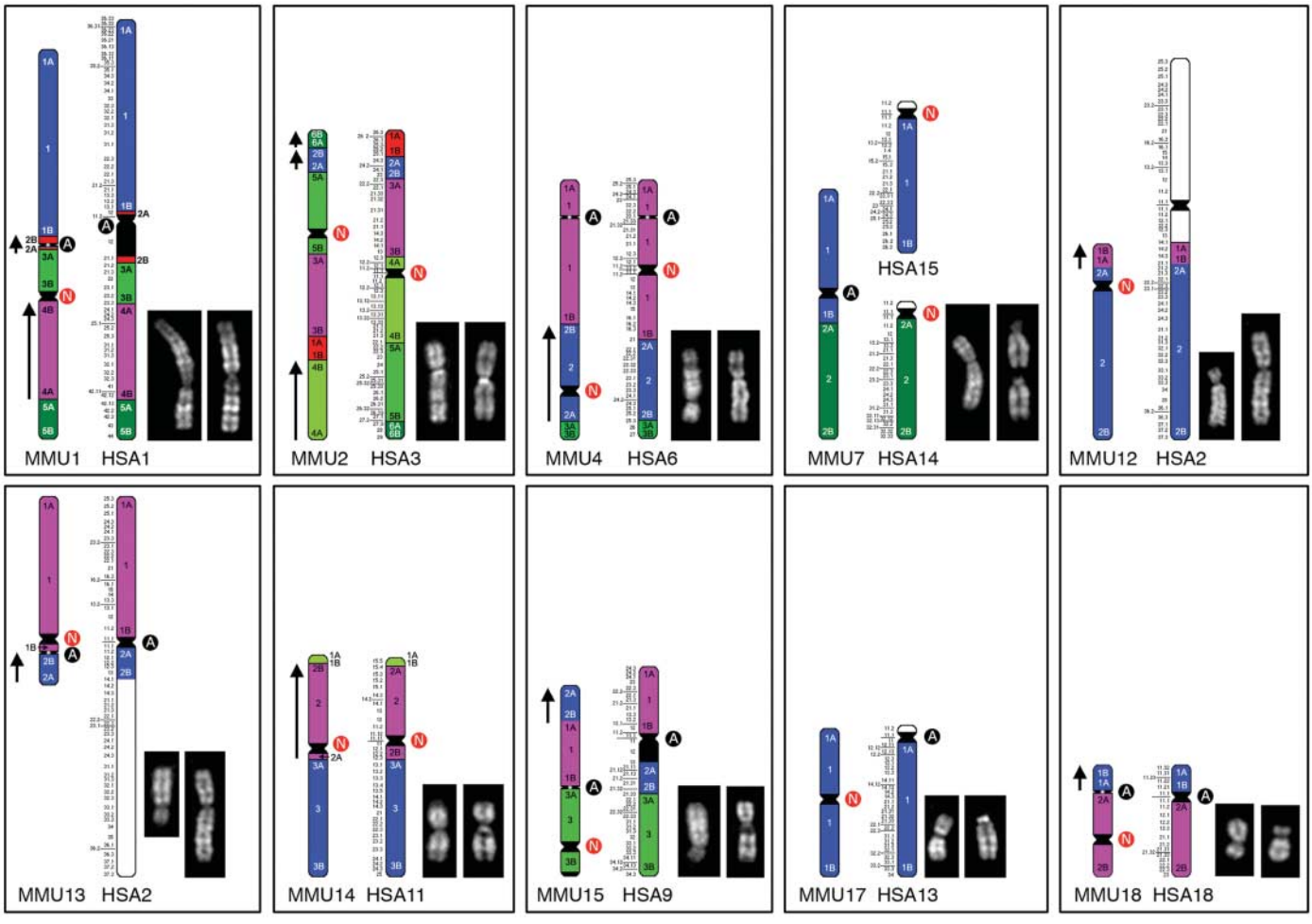

Fig. 1. All macaque (left) or human (right) chromosomes harboring ENCs. Quinacrine mustard banded macaque (left) and human (right) chromosomes are also displayed. Details are reported at www.biologia.uniba.it/macaque. Some macaque chromosomes are shown upside down to facilitate comparison in accordance with the orientation in the whole-genome assemblage. The Ns in red circles flank ENCs; the As in black circles flank ancestral centromeres. Synteny block conservation is indicated by the color. The arrows on the left of some macaque synteny blocks indicate that these blocks have opposite sequence polarity, with respect to the corresponding block in humans. These annotations, obviously, are affected by the upside down position of some chromosomes. The Human Genome Project has assumed as "+" sequence polarity the 5′->3′ strand that starts from the tip of the short arm of each chromosome. The assumption that centromeres were conserved leads to a violation of this rule for macaque chromosomes 1 (1), 2 (3), 4 (6), 10 (20/22), 13 (2p), and 18 (18) (corresponding human chromosomes in parentheses).

experiments on stretched chromosomes (Fig. 3D) allowed for the construction of a contig defining the duplicated pericentromeric region. In summary, seven imperfect copies of a 250-kb segment, mapping at the seeding point, were duplicated both proximally and distally to the MMU4 centromere. The global tiling path and the detailed organization of the central duplicated region are shown (Fig. 4, A and B). Duplicated regions appear to be co-oriented with respect to each other and to the human sequence assembly. Unexpectedly, a high proportion of BAC ends of duplicated clones were alphoid in nature (30 out of 46). These alphoid sequences showed a monomeric structure that is typical of peripherally located alpha satellite sequences (table S8). Small, repeated inversions in the duplicated regions might be hypothesized to account for these findings. This hypothesis, however, clashes with the apparent co-orientation of the duplicated blocks.

Data from human pericentromeric regions have shown that the ratio of inter- versus intrachromosomal duplications is about 6:1 (*1*). We previously suggested that duplications of the ENC of macaque chromosome 17 (human 13) were intrachromosomal only (*3*), but in that case only human probes were used, which precluded any firm conclusion on the absence of interchromosomal duplications. Our current results show that the MMU4 pericentromeric duplications detected by FISH were strictly intrachromosomal and originated only from the ENC seeding point. We have also shown that centromeres of human chromosomes 3, 6, 11, 14, and 15 are ENCs (*6*, *9*–*11*). Chromosomes HSA3 and HSA6 match the pattern we found on MMU4, whereas human chromosomes 11, 14, and 15 accommodate large blocks of interchromosomal duplications (*1*). A careful analysis of the evolutionary history of the latter chromosomes, however, showed that it was very likely that large blocks of segmental duplications were already present or simultaneously seeded in the ENC region (*10*, *11*). It could therefore be hypothesized that a novel centromere triggers only local duplication activity, whereas interchromosomal duplications are triggered by distinct forces, probably linked to intrinsic properties of specific sequences (*17*, *18*). However, until further cases are studied we cannot rule out that the duplications we detected on MMU4 are simply macaque-specific.

The corresponding human region in proximity to the L2 marker was investigated for gene content. A relatively large region (780 kb) harboring the MMU4 centromere has not been annotated in the UniProt, RefSeq, and GenBank mRNA databases. The two closest genes on opposite sides, UTRN (chr6:144,654,658 to 145,209,657) and EPM2A (chr6:145,988,133 to 146,098,299), are 778 kb apart. Heterochromatin supposedly silences embedded genes (*19*). Genes mapping to regions where a centromere repositioning occurred might be at risk of silencing, but recent reports have indicated that a neocentromere by itself does not repress gene expression (*20*–*22*). The gene silencing might be attributed to the successive heterochromatization of the region. The average gene content in the human genome is about 1 gene per 100 kb (www.ncbi.nlm.nih.gov). Human chromosome 6 contains about 1272 genes, on average 1 gene every 131 kb (*23*). Consequently, six genes would be expected in the 778-kb gene-desert area. A similar gene-desert area was also found around the ENC of the Old World monkey chromosome homologous to human chromosome 13 (*3*). Our data appear to support the hypothesis that the absence of genes in the ENC seeding region can play an important role in ENC maintenance and progression. Analyses of additional ENCs and their corresponding regions in the human genome will be required to determine whether this is a stochastic occurrence or whether it represents a prerequisite for novel centromere survival. This hypothesis initially appears to be contradicted by the presence of active genes at the centromeres of rice chromosomes 8 (*24*) and 3 (*25*). However, Nagaki *et al.* and Yan *et al.* suggest that these two rice centromeres may represent ENCs that are still acquiring the full heterochromatic organization that is typical of normal centromeres, and the analysis of the fully sequenced *Arabidopsis* genome strongly supports the view that the absence of gene expression in centromeres is also a general rule in plants. Alternatively, it could be hypothesized that the heterochromatization process pushes the surrounding genes to pericentromic regions without affecting their expression.

Ferreri *et al.* (*26*) reviewed the various hypotheses formulated to explain ENC and clinical-neocentromere emergence. One hypothesis proposes that the centromere seeding event is essentially epigenetic in nature and is sequence independent (*27*). Another hypothesis considers the seeding regions to be domains with inherent latent centromere-forming potentiality (*11*, *28*). A third hypothesis suggests that rearrangements trigger neocentromere seeding through chromatin repatterning (*11*). Roizes (*29*) has suggested that damage to a centromere, like retroposon insertion, could trigger the emergence of evolutionary neocentromeres. All of these hypotheses consider clinical neocentromeres and ENCs to be strictly related.

An unexpected finding is the high number of ENCs in recent human and Old World monkey evolution. In the 25 million years since macaque and human divergence, 14 ENCs have arisen and become fixed in either the human or the macaque lineage. It is difficult to escape the conclusion that ENCs had a considerable impact on shaping the primate genome and that they are fundamental to our understanding of genome evolution. Knowledge of centromere repositioning, for instance, provides a cogent explanation for the unusual clustering of human clinical neocentromeres at 15q25, the domain of an inactivated ancestral centromere (*11*). Despite their relevance, ENCs have never been identified on the basis of sequence analysis alone. Indeed, the extensive pericentromeric duplication we report has not been identified in the macaque genome assembly, reinforcing the opinion

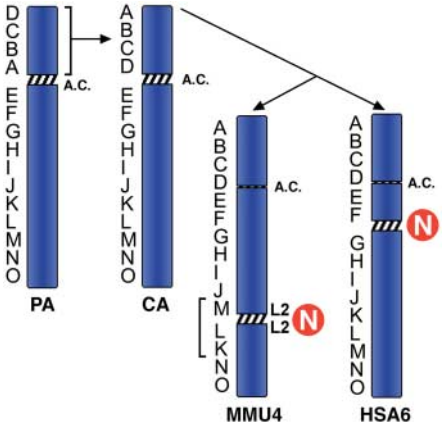

Fig. 2. A reconstruction of chromosome 6 evolution in primates [modified from (*9*)]. The letters indicate the specific BAC clones used in the study (*9*) and reported in table S9. PA, primate ancestral; CA, Catarrhini ancestral; and A.C., ancestral centromere. The letter N in a red circle represents a novel centromere. A square bracket encompasses the Old World monkey–specific (macaque included) inversion of K-L-M markers, with respect to the human form. MMU4 is upside down, with respect to the correct position as reported in Fig. 3, to allow for an easy synteny comparison. L2, human BAC RP11-474A9.

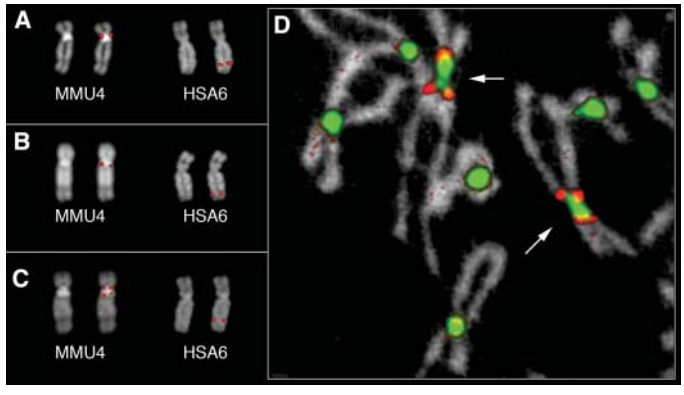

Fig. 3. (**A** to **C**) Partial metaphases showing examples of FISH experiments on macaque and human chromosomes, with the use of the nonduplicated BAC clones (A) CH250-209i5 and (B) CH250-215i15 and the duplicated clone (C) CH250-284C24. (Mapping details are given in Fig. 4, A and B.) (**D**) Partial metaphases showing FISH on stretched chromosomes with the use of the duplicated BAC clones CH250-188G18 (red) and CH250-67B18 (green), containing alphoid sequences. The arrows indicate MMU4.

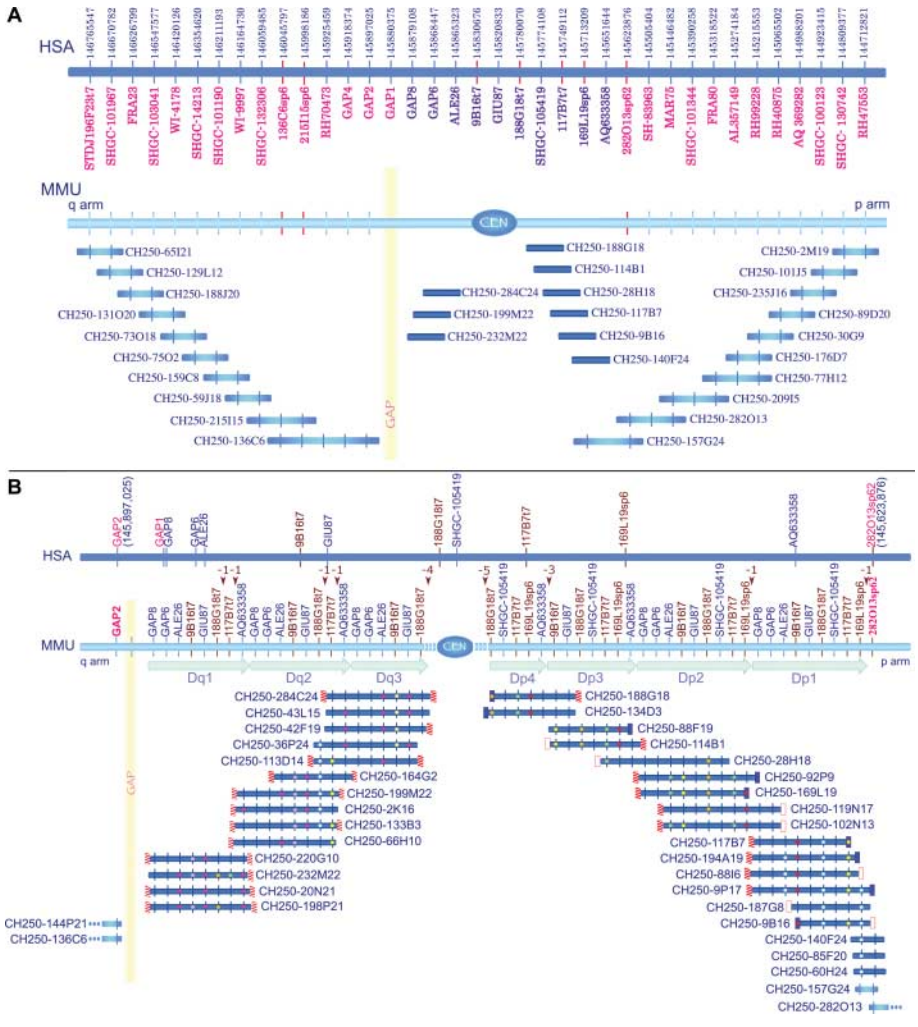

Fig. 4. (**A**) BAC clone contig map of the MMU4 region corresponding to the HSA6 DNA segment from bp 144,712,821 (right) to 146,765,547 (left) (UCSC March 2006 release). Human and macaque STSs (table S4) were used to screen high-density filters of the macaque BAC library CH250. Positively hybridizing BACs were then used in FISH experiments on macaque and human metaphases, and a contig tiling path of the BACs (blue segments) was defined. The analysis discovered single and duplicated regions around the MMU4 pericentromeric region. The marker positions of the STSs in the human sequence are reported in the top line. STSs in red are localized in the MMU4 single-copy region; blue STSs markers are within the region that is duplicated around the MMU4 centromere. This duplicated region is illustrated in detail in (B). Vertical red lines represent macaque STSs derived from BAC-end sequences. The yellow vertical region represents a gap that is probably composed of sequences refractory to cloning. (**B**) Details of the BAC contig assembly of the duplicated pericentromeric region of MMU4. FISH analysis of the macaque BAC clones indicated that duplicated blocks were present on both sides of the centromere. To further characterize these duplicated pericentromeric regions, we sequenced and aligned appropriate STS-amplified products from different clones with the MegAlign software (*31*). The sequence analysis allowed for the classification of the duplicated STSs and the corresponding BACs into seven separate blocks: three on the q arm (Dq1-3) and four on the p arm (Dp1-4) (table S7). The top blue line shows the mapping of the STS markers in human. The first nonduplicated STS, on both sides, is shown in red. The MMU line depicts all of the STS that were used in the study, arranged according to their inferred position in macaque. Macaque STSs derived from BAC-end sequences are shown in brown. Dq1, Dq2, and Dq3 and Dp1, Dp2, Dp3, and Dp4 indicate the duplicated subsets on the q-arm side (Dq-block) and on the p-arm side (Dp-block) of the centromere, respectively. Some STSs are apparently missing in some blocks. Their positions are indicated by brown arrowheads above the macaque STS line. The number of missing STSs is also indicated. Sequenced BAC ends are represented by small rectangles at the borders of the BACs. White and red diagonally striped rectangles stand for monomeric alpha satellite sequences; unfilled red rectangles stand for repetitive nonsatellite elements; and blue rectangles stand for single-copy sequences. The absence of a rectangle indicates that the end sequencing failed twice. Sequenced PCR products of specific STS primers are represented by colored circles. These sequences were used for the classification of duplicated blocks. The colored circles indicate STSs whose PCR products have been sequenced. Circles related to the same STS are in the same color only if their sequences are perfectly matched. A full description of the BAC contig assembly is reported in the SOM text.

that an integrated, multidisciplinary approach is needed for high-quality genome assembly and for comparative genomics (*30*).

The present data extend the link between segmental-duplication bias and centromeres to additional primate species. The homology and shuffling of sequences creates substrates for evolutionary innovation (the birth of new genes) and instability (via non-allelic homologous recombination). Lastly, the contig assembly we have constructed represents a framework for the complete sequencing of the pericentromeric region of MMU4 ENC through a direct sequencing of BAC templates, as opposed to whole-genome shotgun sequencing.

References and Notes

1. X. She *et al.*, *Nature* **430**, 857 (2004).
2. G. Montefalcone, S. Tempesta, M. Rocchi, N. Archidiacono, *Genome Res.* **9**, 1184 (1999).
3. M. F. Cardone *et al.*, *Genome Biol.* **7**, R91 (2006).
4. D. J. Amor, K. H. Choo, *Am. J. Hum. Genet.* **71**, 695 (2002).
5. D. J. Amor *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6542 (2004).
6. M. Ventura *et al.*, *Genome Res.* **14**, 1696 (2004).
7. W. J. Murphy, L. Fronicke, S. J. O'Brien, R. Stanyon, *Genome Res.* **13**, 1880 (2003).
8. R. Marzella *et al.*, *Genomics* **63**, 307 (2000).
9. V. Eder *et al.*, *Mol. Biol. Evol.* **20**, 1506 (2003).
10. M. F. Cardone *et al.*, *Genomics*, in press.
11. M. Ventura *et al.*, *Genome Res.* **13**, 2059 (2003).
12. J. Wienberg, R. Stanyon, A. Jauch, T. Cremer, *Chromosoma* **101**, 265 (1992).
13. S. Muller, J. Wienberg, *Hum. Genet.* **109**, 85 (2001).
14. W. J. Murphy, R. Stanyon, S. J. O'Brien, *Genome Biol.* **2**, REVIEWS0005 (2001).
15. R. Stanyon *et al.*, *Am. J. Primatol.* **50**, 95 (2000).
16. X. She *et al.*, *Genome Res.* **16**, 576 (2006).
17. J. E. Horvath *et al.*, *Genome Res.* **15**, 914 (2005).
18. J. A. Bailey, E. E. Eichler, *Nat. Rev. Genet.* **7**, 552 (2006).
19. K. S. Weiler, B. T. Wakimoto, *Annu. Rev. Genet.* **29**, 577 (1995).
20. R. Saffery *et al.*, *Mol. Cell* **12**, 509 (2003).
21. N. C. Wong *et al.*, *PLoS Genet.* **2**, e17 (2006).
22. A. L. Lam, C. D. Boivin, C. F. Bonney, M. K. Rudd, B. A. Sullivan, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 4186 (2006).
23. A. J. Mungall *et al.*, *Nature* **425**, 805 (2003).
24. K. Nagaki *et al.*, *Nat. Genet.* **36**, 138 (2004).
25. H. Yan *et al.*, *Plant Cell* **18**, 2123 (2006).
26. G. C. Ferreri, D. M. Liscinsky, J. A. Mack, M. D. Eldridge, R. J. O'Neill, *J. Hered.* **96**, 217 (2005).
27. A. Alonso *et al.*, *Hum. Mol. Genet.* **12**, 2711 (2003).
28. K. H. A. Choo, *Am. J. Hum. Genet.* **61**, 1225 (1997).
29. G. Roizes, *Nucleic Acids Res.* **34**, 1912 (2006).
30. M. Rocchi, N. Archidiacono, R. Stanyon, *Genome Res.* **16**, 1441 (2006).
31. Materials and methods are available as supporting material on *Science* Online.

32. We acknowledge the Ministero della Universita' e della Ricerca (MUR) and the European Commission (grant QLRI-CT-2002-01325) for financial support. This work was also supported in part by NIH grants GM58815 and HG002385 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute. We would like to thank G. Herrick for critical reading. R.S. was supported by the MUR grant "Mobility of Italian and Foreign Researchers Residing Abroad."

Supporting Online Material

www.sciencemag.org/cgi/content/full/316/5822/243/DC1
Materials and Methods
SOM Text
Figs. S1 and S2
Tables S1 to S9
References

31 January 2007; accepted 15 March 2007
10.1126/science.1140615