39. Dickson, R. R., Meincke, J., Malmberg, S. & Lee, A. *Prog. Oceanogr.* **20,** 103–151 (1988).
40. Liu, W. T. & Niiler, P. P. *J. geophys. Res.* **95,** 9749–9753 (1990).
41. Schmitt, R. W., Bogden, P. S. & Dorman, C. E. *J. phys. Oceanogr.* **19,** 1208–1221 (1989).
42. Vonder Haar, T. H. in *Current Problems in Atmospheric Radiation, Proc. int. Radiation Symp, Lille, France* (eds Lenoble, J. & Geleyn, J.-F.) 342–344 (Deepak, Hampton, Virginia, 1989).
43. Bryden, H. L., Roemmich, D. H. & Church, J. A. *Deep Sea Res.* **38,** 297–324 (1991).
44. Peixoto, J. P. & Oort, A. H. in *Variations of the Global Water Budget* (eds Street-Perrott, A., Beran, M. & Ratcliffe, R.) 5–65 (Reidel, Dordrecht, 1983).
45. Stephens, G. L., Campbell, G. G. & Vonder Haar, T. H. *J. geophys. Res.* **86,** 9739–9760 (1981).
46. Cohen, J. & Rind, D. *J. Clim.* **4,** 689–706 (1991).
47. Shukla, J. & Mintz, Y. *Science* **215,** 1498–1500 (1982).
48. Manabe, S. *Mon. Weath. Rev.* **97,** 739–774 (1969).
49. Dooge, J. C. I. in *Land Surface Processes in Atmospheric General Circulation Models* (ed. Eagleson, P. S.) 243–288 (Cambridge Univ. Press, 1982).
50. Hansen, J. *et al. Mon. Weath. Rev.* **111,** 609–662 (1983).
51. Dickinson, R. E. *Am. geophys. Un. Geophys. Monogr.* **29,** 58–72 (1984).
52. Sellers, P. J., Mintz, Y., Sud, Y. C. & Dalcher, A. *J. atmos. Sci.* **43,** 505–531 (1986).
53. Moore, B. III *et al.* in *Global Ecology* (eds Rambler, E., Margulis, L. & Sagan, D.) 113–141 (Academic, Boston, Massachusetts, 1989).
54. Dickinson, R. E., Henderson-Sellers, A., Kennedy, P. J. & Wilson, M. F. *NCAR Tech. Note* NCAR/TN-275+STR (NCAR, Boulder, Colorado, 1986).
55. Deardorff, J. W. *J. geophys. Res.* **83,** 1889–1903 (1978).
56. Dickinson, R. E. & Henderson-Sellers, A. *Q. Jl R. met. Soc.* **114,** 439–462 (1988).
57. Lean, J. & Warrilow, D. A. *Nature* **342,** 411–413 (1989).
58. Shukla, J., Nobre, C. & Sellers, P. *Science* **247,** 1322–1325 (1990).
59. Rind, D., Rosenzweig, C. & Goldberg, R. *Nature* **358,** 119–122 (1992).
60. Sellers, P. J., Hall, F. G., Asrar, G., Strebel, D. E. & Murphy, R. E. *Bull. Am. met. Soc.* **69,** 22–27 (1988).
61. Press, F. *Opportunities in the Hydrological Sciences* (foreword) (Natn. Acad. Press, Washington DC, 1991).
62. National Research Council, *Opportunities in the Hydrologic Sciences* (Natn Acad. Press, Washington DC, 1991).

ACKNOWLEDGEMENTS. I thank D. Vane for discussions and comments and for her support in preparation of the manuscript. This work was sponsored by NASA.

Continuum of overlapping clones spanning the entire human chromosome 21q

Ilya Chumakov[*†], Philippe Rigault[†], Sophie Guillou[†], Pierre Ougen[*], Alain Billaut[*], Ghislaine Guasconi[†], Patricia Gervy[†], Isabelle LeGall[*], Pascal Soularue[*], Laurent Grinas[†], Lydie Bougueleret[*], Christine Bellanné-Chantelot[*], Bruno Lacroix[*], Emmanuel Barillot[†], Philippe Gesnouin[†], Stuart Pook[†], Guy Vaysseix[*†], Gerard Frelat[‡], Annette Schmitz[‡], Jean-Luc Sambucy[*], Assumpcio Bosch[§], Xavier Estivill[§], Jean Weissenbach[†‖], Alain Vignal[†], Harold Riethman[¶], David Cox[#], David Patterson[**], Kathleen Gardiner[**], Masahira Hattori[††], Yoshiyuki Sakaki[††], Hitoshi Ichikawa[‡‡], Misao Ohki[‡‡], Denis Le Paslier[*], Roland Heilig[§§], Stylianos Antonarakis[‖‖] & Daniel Cohen[*†¶¶]

* Centre d'Etude du Polymorphisme Humain (CEPH), 27 rue Juliette Dodu, 75010 Paris, France
† Genethon, 1 rue de l'Internationale, BP 59, 91002 Evry Cedex, France
‡ Centre d'Energie Atomique, 68 Avenue Division LeClerc, 92260 Fontenay aux Roses, France
§ Institut de Recerca Oncologica, Hospital Duran i Reynals Ctra Castelldefels, KM 2.7, Barcelona 08907, Spain
‖ CNRS URA 1445, Institut Pasteur, 28 rue Docteur Roux, 75724 Paris, France
¶ The Wistar Institute, 36th Street and Spruce, Philadelphia, Pennsylvania 19104, USA
# Neurogenetics Laboratory, University of California, 401 Parnassus Avenue, San Francisco, California 94143, USA
** Eleanor Roosevelt Institute, 1899 Gaylord Street, Denver, Colorado 80206, USA
†† The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai Minatoku, Tokyo 108, Japan
‡‡ Saitama Cancer Center Research Institute, Ina-machi, Kitaadachi-gun, Saitama-ken 362, Japan
§§ INSERM U184, Laboratoire de Génétique Moléculaire de Eucaryotes, 11 rue Humann, Strasbourg, France
‖‖ Center for Medical Genetics, The Johns Hopkins University School of Medicine, 600 North Wolfe Street CMSC 1003, Baltimore, Maryland 21205, USA

A continuous array of overlapping clones covering the entire human chromosome 21q was constructed from human yeast artificial chromosome libraries using sequence-tagged sites as landmarks specifically detected by polymerase chain reaction. The yeast artificial chromosome contiguous unit starts with pericentromeric and ends with subtelomeric loci of 21q. The resulting order of sequence-tagged sites is consistent with other physical and genetic mapping data. This set of overlapping clones will promote our knowledge of the structure of this chromosome and the function of its genes.

HUMAN genome mapping consists of ordering genomic DNA fragments on their chromosomes using several methods, such as fluorescence *in situ* hybridization (FISH), somatic cell hybrid analysis or random clone fingerprinting[1–10]. When the fragments correspond to polymorphic sites they can be ordered by genetic linkage analysis[11]. Distances between polymorphic loci are estimated by meiotic recombination frequencies. Such a genetic map allows the localization of any polymorphic trait gene.

Human chromosome 21 (HC21) represents a model for physical mapping of the human genome and is the smallest and one of the best-studied human chromosomes. Several genetic diseases are associated with this chromosome[12], including Down's syndrome (the most frequently occurring mental retardation in humans), some forms of Alzheimer's disease and other neurological diseases, such as progressive myoclonus epilepsy and amyotrophic lateral sclerosis. A map of contiguous units (contigs) covering this chromosome will speed the identification of the cause of these diseases. Indeed, it provides an immediate access to the genomic segment, including any pathological locus, as soon as it has been localized by genetic linkage or cytogenetic analysis.

The process of developing such a long-range contig map involves the identification and localization of landmarks in cloned genomic fragments. When there are enough landmarks for the size of the cloned fragments, contigs are formed, and the landmarks are simultaneously ordered[13]. Yeast artificial chromosome (YAC) cloning provides the means to isolate large, but manageable, DNA fragments of 100 to 2,000 kilobases (kb);

¶¶ To whom correspondence should be addressed.

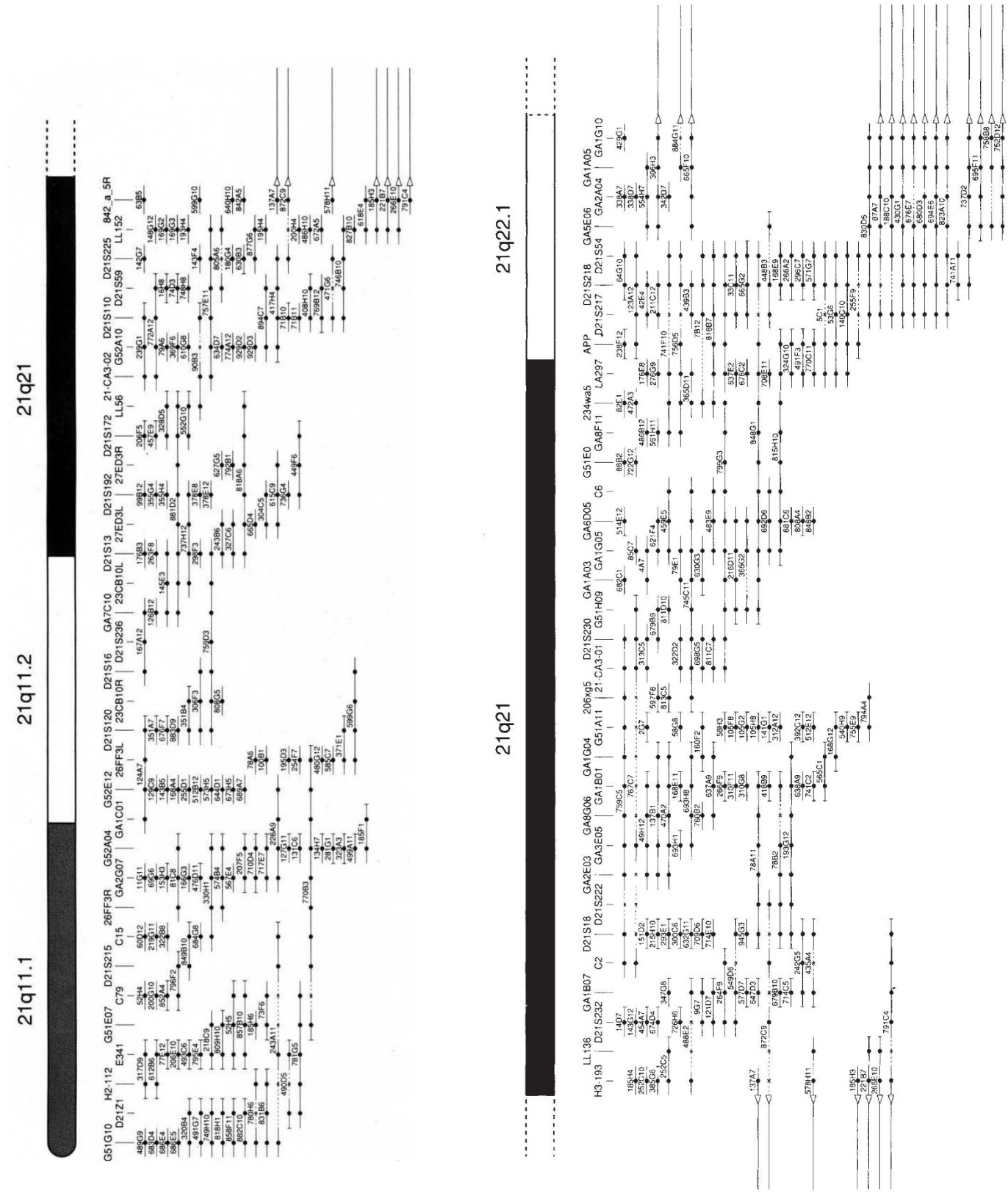


FIG. 1 Above and pages 382/3. STS content map of the contiguous array of YAC clones. Clones are presented as lines. Their size only reflects the number of included STSs. The physical distances between adjacent STSs are diverse. As the degree of chimaerism of each clone is unknown, only HC21-specific portions are represented. Filled circles indicate positive STS. Broken lines mean that a clone was not found to be positive with a given STS when only screening pools. A cross means that a clone, when checked individually, was found to be negative with an STS. Bars at the end of the clone indicate that a clone was found to be negative with an adjacent STS. For comparison, a drawing showing G-banded chromosome 21q is placed above the contig presentation. ►

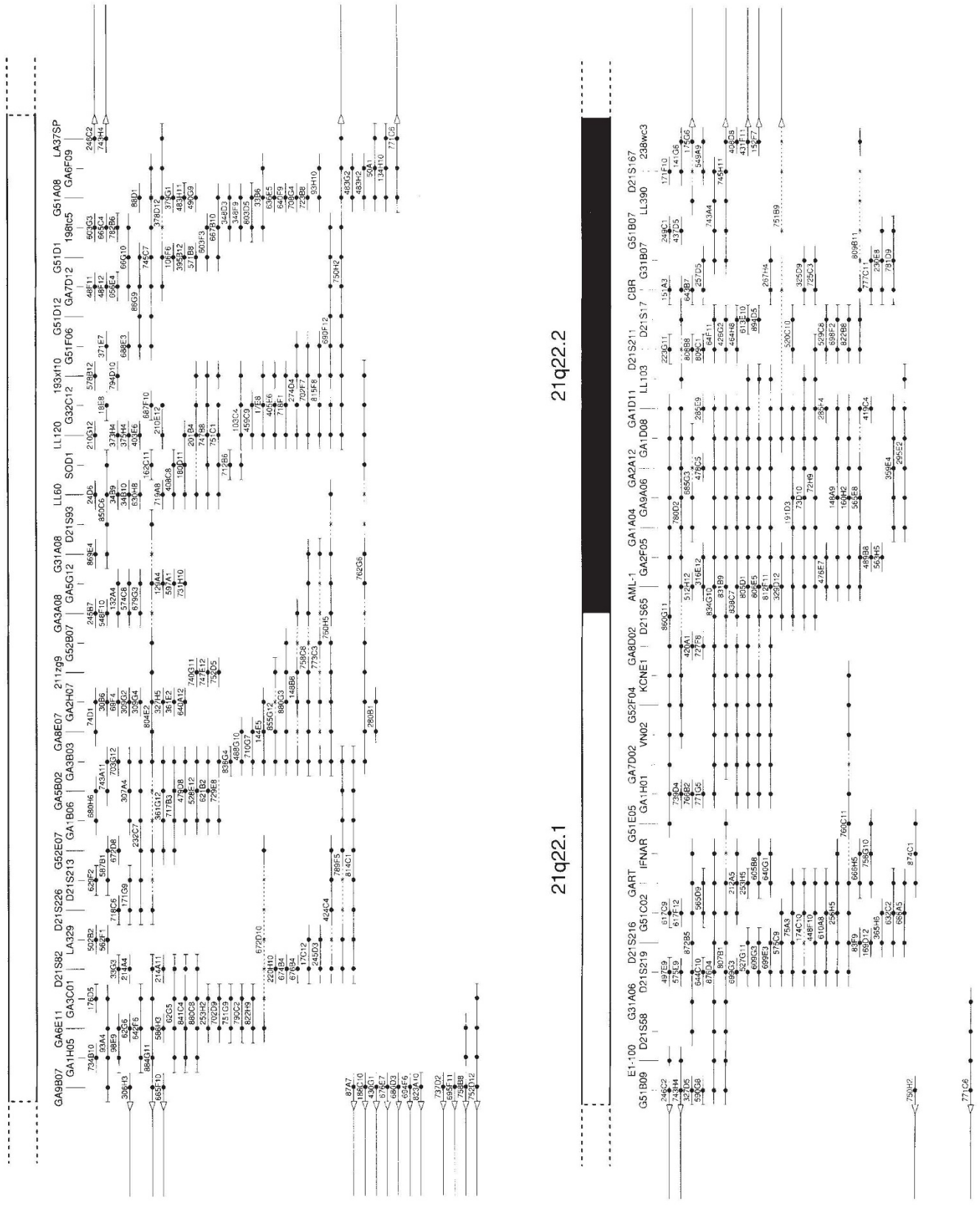


FIG. 1—continued.

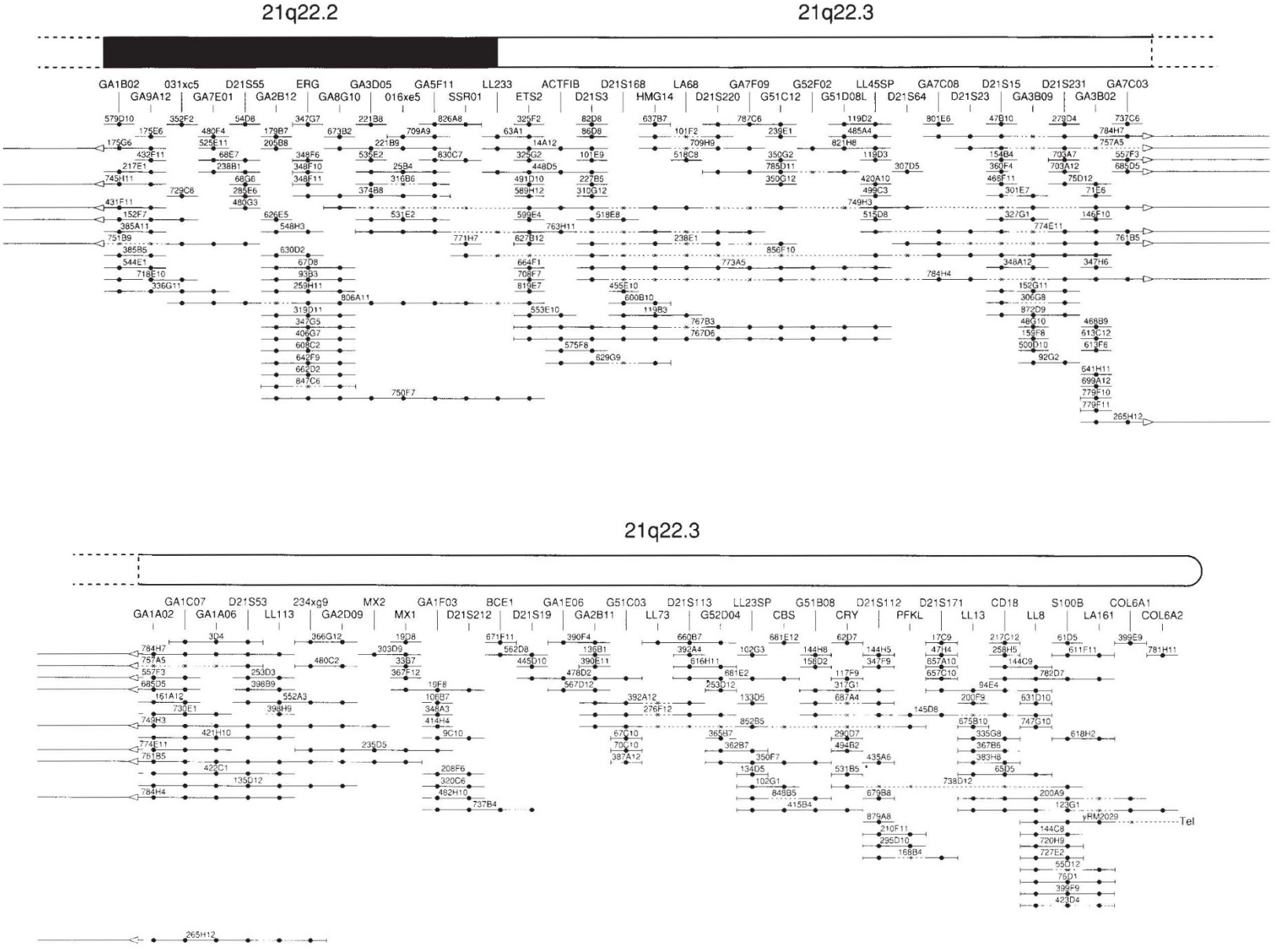


FIG. 1—continued.

these are much larger than those isolated using bacterial cloning systems[14–18]. This fills the gap between genetically and physically measurable distances. Moreover, the moderate size of YAC libraries make them well-adapted to screening by polymerase chain reaction (PCR). This allows the use, as landmarks, of sequence-tagged sites (STS)[19], which are short stretches of DNA sequence that can specifically be detected by PCR. Implicit in the definition is that the STS is operationally unique in the genome. A major advantage of this strategy is that it provides an ordered set of markers that can be stored as sequence information in a database. Other potentially informative physical landmarks can be readily converted into STS by DNA sequencing. They establish the common language for the comparison with other physical mapping data. Moreover, ordered collections of clones can be reconstituted from new libraries, when using ordered STS, providing an easy update as cloning technology improves[19]. We report here the first STS map for the entire human chromosome 21q (HC21q) that is now covered by a single array of overlapping YACs with a mean size of about 600 kb.

A set of 21q-specific sequence-tagged sites

Sequences of several HC21 genes are available from databases. These sequences can easily be converted into STS. Recently a number of STS, mainly polymorphic, were generated for this chromosome to aid genetic linkage studies. Other STS include markers from *Not*I-linking clones[20] and from the vicinity of naturally occurring chromosome translocations. Because there is a tendency of some of these markers to be nonuniformly distributed throughout the length of the chromosome, we also screened with other anonymous STS from different origins. One of these sources is the recent collection[21] of STS derived from previously mapped cloned fragments. We also converted some cloned PCR products from flow-sorted HC21 preparations into STS. This gave us 88 new anonymous markers, the largest single source of STS. All of them have been submitted to the Genome Data Base, the full list being available on request. In total, we have 21 STS derived from *Not*I-linking clones, 50 polymorphic STS and 21 STS derived from known genes, whereas the remainder are mainly anonymous markers. Some STS, assigned to HC21, could not be efficiently used for the screening because they produced nonspecific amplication from the yeast DNA or primer dimerization. All STS, whatever their source, were tested on a somatic cell hybrid containing only HC21 (ref. 22) before use and, in most cases, on another hybrid, containing only HC21q (refs 21–23).

Screening YAC libraries and contig assembly

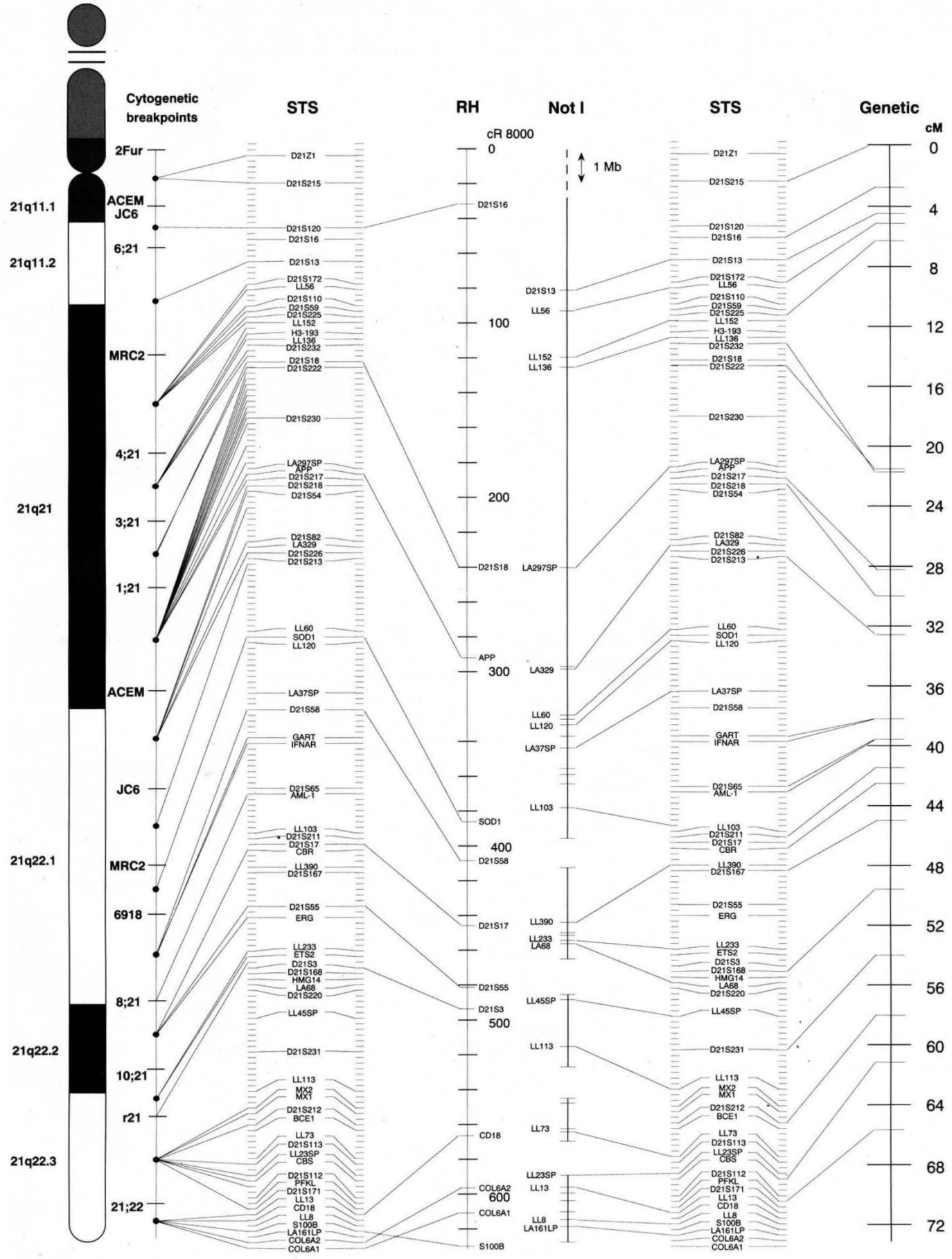
Three different YAC libraries were screened entirely by PCR for the presence of the HC21q-specific STS. The first one, containing about 70,000 YAC clones of average size 470 kb, corresponds to 9.4 human genome equivalents. It was made from the DNA of the same lymphoblastoid cell line and used to construct a subset of this library of about 50,000 clones already described[17]. This library, stored as an array of individual clones in 736 96-well plates, was screened by PCR in a two-step pooling procedure. We first tested 92 primary pools (8 plates each) representing all clones in the library. The candidate clone in a given positive primary pool was identified by testing 28 smaller pools. These smaller pools contained DNA from clones of individual plates (8 pools), rows (8 pools) and columns (12 pools) of the corresponding 8-plate set. On average, 355 tests were necessary to identify the clones positive for a given STS from this 9.4 genome-equivalent library. To exclude false positives, we usually checked candidate positive clones individually. Another YAC library, chromosome 21-specific, consisting of 180 clones was derived from 14,000 YACs of 1 megabase (Mb) mean size, as previously described[18]. This sublibrary corresponds to about four genome equivalents. Its clones were screened individually. Both total human genome libraries (P.O., manuscript in preparation) were constructed using *Eco*RI partially digested genomic DNA, sized through pulse-field agarose gel electrophoresis[17] and inserted into a pYAC4 vector[14]. The last library[24], consisting of human telomere-containing YACs, was only screened with subtelomeric STS. Analysis of PCR reactions was done by agarose gel electrophoresis and images were directly read by a CCD (charge-couple device) camera, then processed by appropriate software and transferred on-line to a specially designed SYBASE database. A special algorithm (available on request from the authors) was developed to produce and order contigs and landmarks. For a given STS order, an energy function was calculated, taking into account all inconsistencies in the clone overlaps. The best order was proposed by minimizing this energy function by simulated annealing[25]. This kind of ordering may occasionally lead to wrong solutions mainly due to the presence of false positives and negatives. But we assume that this problem can be overcome by using a sufficiently redundant library.

Continuum of YACs on 21q

In total, for 198 STS used to screen a 13.4 human genome-equivalent library we isolated 810 positive clones. The mean number of positive YACs per STS was 10.2. Although comparable to the library redundancy, this slightly lower number probably reflects overall efficiency of our screening procedure and also the presence of deleted clones. When corrected for the clones that are apparently deleted (at most 12%) this number increases to 11.2.

The array of overlapping clones is shown in Fig. 1. It contains the positive clones for all STS used for the screening. The interval between two STS is covered by an average of five clones, indicating the robustness of the contig. There are two main sources of error in contig assembly. We tried to exclude trivial false positives, by checking candidates individually. False merging could also be due to so-called chimaeric clones that contain more than one genomic region. Owing to the small size of HC21 (1.5% of the human genome), the frequency of such HC21 to HC21 chimaeric clones is only 0.7%, calculated on the basis of 40% chimaeric clones in the overall library (P.O., manuscript in preparation). But given the large redundancy of the library, problems due to potential chimaeras can be resolved (indeed we observed only one clear case of such a problem). Another source of false overlap can be inclusion of data from PCR landmarks that are not STS in a strict sense (not unique). This problem can be solved partially by assignment of STS to human chromosomes using a somatic cell hybrid mapping panel. We also considered a particular PCR landmark as a moderate repeat when it gave too many positive primary pools during the first step of screening. STS that gave more than 15 positives when tested on primary pools from the 9.4 genome-equivalent library were usually excluded from subsequent screening. Nevertheless, the six most proximal STS of our contig are also present on chromosomes other than HC21. PCR landmark D21Z1, a pericentromeric alpha satellite repeat[26], is present on chromosomes 21 and 13. We have shown that the STS E341 (ref. 27), derived from the sequence of minicircle DNA is present on chromosomes 21, 15 and 22. Polymorphic marker D21S215 also has a nonpolymorphic counterpart on chromosome 18. The rest of the markers in this region (G51G10, G51E07 and C79) are specific to other acrocentric chromosomes (13, 14, 15, 21 and 22). It is possible, therefore, that single STS-positive clones and even some of the STS-linking clones in this region do not

FIG. 2 Comparison of the STS map to the breakpoint panel map, irradiation hybrid map, *Not*I fragment PFGE map and genetic linkage map. STS are represented as bars, each corresponding to an individual site. Only some that are essential for comparison have been named. Scales are given in centirays (cR) for the radiation hybrid map, megabases (Mb) for PFGE map or centimorgans (cM) for genetic data. ►



originate from chromosome 21. But at least clones 243A11, 490D5, 781G5, 770B3, 796F2 and 849B10 establish the overlap of HC21-specific clones from the alpha satellite PCR marker D21Z1, localized at the centromere, to the HC21q-specific marker C15. From this point, the contig extends uninterruptedly and ends with the clone yRM2029, that contains active human telomere and three subtelomeric STS markers.

In some regions of our contig, the linkage is mediated only by a few or by unstable clones. We used repeat-containing restriction fragment fingerprint[9] (C.B.-C. *et al.*, manuscript in preparation) and Alu PCR product patterns[18] (I.C., unpublished results) of YACs to confirm these regional clone overlaps. Using these methods we were able to document the linkage in the regions between LL45SP and D21S64, LL56 and 21-CA3-02, LL136 and D21S232, LL103 and D21S211 (data not shown). The case of apparently uncertain linkage in the vicinity of LL73 is due to the low reproducibility of amplification obtained with the primers corresponding to this STS, the weakness disappearing when it is omitted.

It is difficult to estimate the effective genomic length of a contig produced by STS content analysis. Some of the STS used were derived from sequences around *Not*I sites. The lengths of the clones covering these intervals are compatible with the size of corresponding *Not*I fragments (data not shown). The sum of fragments, produced by *Not*I from the 21q suggests a size between 40 and 50 Mb[20,23]. A very rough statistical estimate for the size of our contig, derived from the total number of positive clones, their average insert size and corrected for the estimated 40% chimaeras gives a figure of the same order (42 Mb). Moreover, although it is difficult at present to define a minimal set of overlapping YACs, we estimated their number to be 55 with a mean size of 1 Mb.

Order of STS landmarks

Construction of the YAC contig by STS content analysis simultaneously produces the relative order of these landmarks across the chromosome. The resolution of this map is dependent not only on the number of STS tested, but also on the number of YACs containing informative combinations of STS. The most useful are those containing only pairs of adjacent landmarks. We found that, although the contig assembly was more efficient with larger YACs, the STS map was more accurate with smaller clones. The landmark order resolution is refined when the clone coverage increases. Obviously in certain regions, corresponding to an exceptionally high density of STS (for example in the vicinity of AML1 (acute myelogenous leukaemia)) no local order can be unequivocally deduced, given the lack of informative clones. Another factor influencing the order of STS is the presence of clones, not detected with a given STS, but detected with its neighbour. Such false negatives can create local errors when deducing order. They can be eliminated by checking clones for the presence of neighbouring STS. We found that in some regions these tests do change regional STS order. The present map reflects the number of such tests done. Another problem is the presence of rearranged, unstable clones creating errors which are difficult to eliminate when they occur at the same site. One such hot-spot for deletions is located between D21S3 and D21S15, another one being evident in the subtelomeric region. This latter could explain, for example, the more distal position proposed for the phosphofructokinase liver-type marker, relative to its immediate neighbour D21S112.

A few markers mapped at several places on our contig. For example, D21S11 has two possible locations: one close to D21S110, the other close to D21S232. LL54 is localized close to D21S18 and in the vicinity of β-amyeloid precursor protein. The simplest explanation is the existence of homologous regions in this part of chromosome 21. This also illustrates the potential difficulties that certainly will be encountered in the mapping of other genomic regions using the STS approach.

Comparison with other maps

The order of markers for chromosome 21 was derived by several other methods, including genetic linkage analysis[28] (M. G. McInnis *et al.*, manuscript in preparation), pulse-field electrophoresis analysis of fragments produced by rare-cutter endonucleases[19], the analysis of panels of somatic cell hybrids, containing naturally occurring deletions[23] or radiation-induced deletions[3]. Some of the markers used in these studies available as STS were mapped in our contig. The comparison of their position to known genetic and physical maps is given in Fig. 2. Most markers show the same relative order when all of these maps are integrated. The few discrepancies include the relative order of most telomeric markers on the radiation hybrid map. This can be explained both by the differential mode of retention of subtelomeric regions in radiation hybrids and by the abnormal nature of YACs derived from the telomere. Certainly we have a few clones containing genes COL6A1 and COL6A2, whose absence in human telomeric YAC yRM2029 argues for their more proximal position, concordant with radiation hybrid mapping. Our map fits with the orientation of LL390, LL233 and LA68 as it is drawn on the *Not*I map. Another possible inverted orientation of these markers has recently been suggested (H.I., M.O., unpublished results), but would be inconsistent with the integrated STS and genetic map in this region. Another *Not*I STS, LL102, is also located 100 kb from AML1 in the region of the above mentioned cluster of high-density STS.

Discussion

This work provides an essential tool for improving our knowledge of HC21, especially to find new genes and, more generally, to derive sequence data. But the choice of clones from which such information could be generated remains problematic owing to the frequency of chimaeric and rearranged inserts inherent in the cloning system. Size information, already obtained in 70 YACs did not yield a definite conclusion concerning these cloning artefacts. Therefore, future work on HC21 will certainly require checking by several methods, such as high-resolution restriction mapping. Moreover, new clones obtained from other cloning systems can rapidly be aligned to the present contig. Indeed, other libraries can be screened with these STS and other probes derived from this collection of YACs.

In total, 198 landmarks used were resolved in 191 discrete loci with an average spacing of roughly 220 kb. Only six loci contain more than one STS, five with two STS and one with three. This resolution probably exceeds the accuracy of most mapping approaches. Indeed, from such arrays of YAC clones, new polymorphic markers can be generated directly from YACs to create a high-resolution genetic map, allowing mapping of genes involved in multifactorial traits.

Despite its small size, chromosome 21 might not be much different in structure from other human autosomes. Construction of a YAC contig for its entire length indicates clearly that STS content mapping could also be applied to other chromosomes. It also demonstrates that this strategy can be used directly, without regional assignment of landmarks by other means. It is tempting, therefore, to suggest that such an approach could be applied to the entire human genome using completely random landmarks. □

Received 1 September; accepted 4 September 1992.

1. Montanaro, V. *et al. Am. J. hum. Genet.* **48,** 183–194 (1991).
2. Korenberg, J. R. & Tykowski, M. C. *Cell* **53,** 391–400 (1988).
3. Cox, D. R. *et al. Science* **250,** 245–250 (1990).
4. Gardiner, K. *et al. Somat. Cell molec. Genet.* **14,** 623–638 (1988).
5. Coulson, A., Sulston, J., Brenner, S. & Kam, J. *Proc. natn. Acad. Sci. U.S.A.* **83,** 7821–7825 (1986).
6. Olson, M. *et al. Proc. natn. Acad. Sci. U.S.A.* **83,** 7826–7830 (1986).
7. Evans, G. A. & Lewis, K. A. *Proc. natn. Acad. Sci. U.S.A.* **86,** 5030–5034 (1989).
8. Craig, G. *et al. Nucleic Acids Res.* **18,** 2653–2660 (1990).
9. Stallings, R. L. *et al. Proc. natn. Acad. Sci. U.S.A.* **87,** 6218–6222 (1990).
10. Schlessinger, D. *et al. Genomics* **11,** 783–793 (1991).
11. Botstein, D. *et al. Am. J. hum. Genet.* **32,** 314–331 (1980).

12. Cox, D. R. *et al. Cytogenet. Cell Genet.* **58,** 800–826 (1991).
13. Green, E. D. & Green, P. *PCR Meth. Appl.* **1,** 77–90 (1991).
14. Burke, D. T., Carle, G. F. & Olson, M. V. *Science* **236,** 806–812 (1987).
15. Anand, R., Villasente, A. & Tyler-Smith, C. *Nucleic Acids Res.* **17,** 4325–4333 (1989).
16. Larin, Z., Monaco, A. P. & Lehrach, H. *Proc. natn. Acad. Sci. U.S.A.* **88,** 4123–4127 (1991).
17. Albertsen, H. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **87,** 4256–4260 (1990).
18. Chumakov, I. M. *et al. Nature Genet.* **1,** 222–225 (1992).
19. Olson, M. V. *et al. Science* **245,** 1434–1435 (1989).
20. Ichikawa, H. *et al. Proc. natn. Acad. Sci. U.S.A.* **89,** 23–27 (1992).
21. Tanzi, R. E. *et al. Genomics* (in the press).
22. Van Keuren, M. *et al. Am. J. med. Genet.* **38,** 793–804 (1989).
23. Gardiner, K. *et al. EMBO J.* **9,** 25–34 (1990).
24. Riethman, H. *et al. Proc. natn. Acad. Sci. U.S.A.* **86,** 991–995 (1989).
25. Rigault, P. *Int. J. Genome Res.* (in the press).
26. Warburton, P. E. *et al. Genomics* **11,** 324–333 (1991).
27. Assum, G. *et al. Genomics* **11,** 397–404 (1991).
28. Petersen, M. B. *et al. Genomics* **9,** 407–419 (1991).

ACKNOWLEDGEMENTS. We thank G. Peirano, D. Grausz, R. Bahouayila, J. Beckmann, P. Millasseau, J.-M. Delabar, N. Creau-Goldberg, H. Bui, C. Soravito, M. Belova, A. C. Warren, M. G. McInnis, H. Chen, J.-L. Blouin and D. Avramopoulos for their help. S.A. thanks A. Chakravarti and J. Blaschak for communicating unpublished results for the analysis of the genetic linkage map. We also warmly thank B. Barataud. This work was supported by the French Ministry of Research and Technology and Association Française contre les Myopathies through the Genethon Program. This work was also supported in part by the National Insitute of Child Health and Human Development (D.P.), the National Center for Human Genome Research (USA) (K.G.), the NIH (D.C., S.A. and H.R.), FISS, CICYT and the MEC (Spain) (A.B. and X.E.), and a Grant-in-Aid from the Ministry of Education, Science and Culture of Japan (Y.S. and M.O.).

Crystal structure of the *met* repressor–operator complex at 2.8 Å resolution reveals DNA recognition by β-strands

William S. Somers* & Simon E. V. Phillips†

Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK

The crystal structure of the *met* repressor–operator complex shows two dimeric repressor molecules bound to adjacent sites 8 base pairs apart on an 18-base-pair DNA fragment. Sequence specificity is achieved by insertion of double-stranded antiparallel protein β-ribbons into the major groove of B-form DNA, with direct hydrogen-bonding between amino-acid side chains and the base pairs. The repressor also recognizes sequence-dependent distortion or flexibility of the operator phosphate backbone, conferring specificity even for inaccessible base pairs.

GENE regulation both in prokaryotes and in eukaryotes is achieved largely by proteins that bind to specific sequences in the DNA. The source of this specificity is recognition by the protein of the pattern of functional groups on the edges of the base pairs in the DNA major groove, mediated by hydrogen-bonding, together with a contribution from the sequence-dependent conformational preferences of the DNA backbone[1,2]. Structural studies of a number of prokaryotic repressors and activators[3–13], and eukaryotic homeodomain complexes[14,15], have demonstrated that insertion of one of the α-helices of a conserved helix-turn-helix motif into the major groove is a frequent method of interaction. The structures of two types of zinc-finger–DNA complexes[16,17] also show α-helices lodged in the major groove. We report here the crystal structure of the *Escherichia coli met* repressor–operator complex, where β-strands rather than α-helices interact with the DNA bases to mediate sequence recognition.

Expression of many of the structural genes in the methionine biosynthetic pathway of *E. coli* (reviewed in ref. 18) is controlled largely by the *met* repressor protein, the product of the *metJ* gene. One of the products of the pathway, *S*-adenosylmethionine (SAM), acts as a corepressor *in vitro* and, presumably, *in vivo*. The repressor is a dimer (M_r 23,992) of identical 104-amino-acid subunits, which binds two molecules of SAM non-cooperatively; its three-dimensional structure has already been reported in the presence and absence of corepressor[19]. The repressor binds cooperatively to a number of operators, all of which share sequence homology to an underlying 8-base-pair (bp) repeating unit (AGACGTCT), referred to as the '*met* box', in two to five tandem copies[20–22]. We proposed[21] that arrays of dimeric *met* repressor molecules bind to these extended operator regions, with a stoichiometry of one repressor per *met* box, to form a left-handed superhelix around the DNA. The structure of the specific repressor–operator complex described below is consistent with this proposal. Each repressor molecule interacts directly with the bases through a pair of antiparallel β-strands inserted into the major groove of B-form DNA. The structure of the complex also shows evidence for recognition by the repressor of sequence-dependent distortions or flexibility of the phosphodiester backbone, which could mediate sequence specificity where the base pairs themselves are inaccessible. Studies of operator binding *in vitro* and repression efficiency *in vivo* indicate that the structure observed in the crystal represents the specific repression complex in solution[23].

Structure determination

The *met* repressor protein was overexpressed and purified as before[24]. A self-complementary 19-base oligonucleotide, with sequence 5′-TTAGACGTCTAGACGTCTA-3′ (that is, two tandem consensus *met* boxes flanked by T·A base pairs and one unpaired 5′-T), was synthesized on an Applied Biosystems 381A oligonucleotide synthesizer, purified by anion exchange and reversed-phase high-performance liquid chromatography (HPLC), and annealed by heating to 80 °C for 5 min, followed by slow cooling to room temperature, for use in crystallization trials. Oligonucleotide fragments recovered from these trials are bound specifically by repressor in gel retardation assays (I. Manfield, personal communication). Crystals were grown by hanging-drop vapour diffusion over wells containing 10 mM sodium cacodylate buffer, pH 7.0, 1 mM sodium azide and 30–35% (v/v) freshly distilled 2-methyl-2,4-pentanediol. The drops contained equal volumes of well solution and solutions containing 5 mg ml^{-1} oligonucleotide, 6–12 mg ml^{-1} *met* repressor, 15–30 mM calcium chloride, 10 mM sodium cacodylate buffer, pH 7.0, 6 mM sodium chloride and 1 mg ml^{-1} SAM (*p*-toluenesulphonate salt; Sigma). Crystals grew in space group

* Present address: Genentech Inc., 460 Point San Bruno Boulevard, South San Francisco, California 94080, USA.
† To whom correspondence should be addressed.