A physical map of human chromosome 14

Thomas Brüls*, Gabor Gyapay*, Jean-Louis Petit*, François Artiguenave*, Virginie Vico*, Shizen Qin†, Aye Mon Tin-Wollam‡, Corinne Da Silva*, Delphine Muselet*, Delphine Mavel*, Eric Pelletier*, Michael Levy*, Asao Fujiyama§, Fumihiko Matsuda‖, Richard Wilson‡, Lee Rowen†, Leroy Hood†, Jean Weissenbach*, William Saurin* & Roland Heilig*

** Genoscope and CNRS UMR 8030, 2 Rue Gaston Cremieux, CP 5706, 91057 Evry Cedex, France*
† Multimegabase Sequencing Center, The Institute for Systems Biology, 4225 Roosvelt Way, NE Suite 200, Seattle, Washington 98105, USA
‡ Washington University Genome Sequencing Center, Box 8501, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA
§ RIKEN Genome Sciences Center, RIKEN, 1-7-22, Tsurumi-ku, Yokohama-city, Kanagawa 230-0045, Japan
‖ Centre National de Genotypage, 2 Rue Gaston Cremieux, CP 5721, 91057 Evry Cedex, France

We report the construction of a tiling path of around 650 clones covering more than 99% of human chromosome 14. Clone overlap information to assemble the map was derived by comparing fully sequenced clones with a database of clone end sequences[1,2] (sequence tag connector strategy). We selected homogeneously distributed seed points using an auxiliary high-resolution radiation hybrid map comprising 1,895 distinct positions. The high long-range continuity and low redundancy of the tiling path indicates that the sequence tag connector approach compares favourably with alternative mapping strategies.

We have constructed a map of contiguous DNA fragments cloned in *Escherichia coli* and covering more than 99% of human chromosome 14. Unlike the 'map first and sequence second' approach used for previous chromosome sequencing efforts[3,4], the construction of this map was mainly based on a sequence tag connector (STC) strategy[1,2] in which the map progresses in parallel with the sequencing project. In this iterative approach, fully sequenced bacterial artificial chromosomes (BACs) are searched against a database of clone end sequences to identify minimally overlapping clones and select the next BACs to enter the sequencing pipeline. We combined this 'map as you go' strategy with a dense high-resolution radiation hybrid map. This conjunction conferred a high degree of flexibility on the project: it allowed us to increase the number of relatively evenly distributed seed points and to tailor our efforts in specific chromosomal regions when needed. The fingerprint-based mapping strategy reported by McPherson[5] also successfully faced challenging schedules and integrated alternative sources of map and sequence data; but we feel that the STC-based approach required less manual editing and resulted in a more accurate and complete set of overlapping clones.

The establishment of ordered sets of clones (contigs) relied on a public database of BAC end sequences[2] (ftp://ftp.tigr.org/pub/data/h_sapiens/bac_end_sequences/) augmented with end sequences generated in house. Auxiliary mapping resources included a high-resolution chromosome 14 radiation hybrid map and a chromosome 14 enriched library of about 10,000 BACs.

Initially, we chose around 50 seed markers, ordered at high odds and distributed as evenly as possible on the chromosome, using radiation hybrid data from the Genebridge4 panel[6]. We used probes corresponding to the seeds to isolate clones by hybridization against high-density filters of the chromosome 14 BAC library. To cope with the increasing sequencing throughput, the number of seeds was subsequently tripled by choosing additional BACs that could be mapped unambiguously. We assessed the identity and integrity of selected clones by analysing their restriction fingerprints using the FPC software[7]. Walking could then proceed iteratively and bidirectionally from each fully sequenced seed clone by querying the BAC end sequence database. The walking process entailed systematic checks consisting of re-sequencing of candidate clone ends as well as comparison of their fingerprints.

We monitored the progress of chromosome coverage on a radiation hybrid map built with the TNG high-resolution panel (http://www-shgc.stanford.edu/Mapping/rh/RH_poster/). Roughly 2,350 markers were analysed on this panel, including 640 markers derived from the ends of sequenced clones and 616 entries downloaded from the RHdb database (http://www.ebi.ac.uk/RHdb). Marker ordering was treated as being analogous to the well studied 'travelling salesman problem' for which powerful heuristic tools are available[8] (http://www.caam.rice.edu/keck/concorde.html). We then estimated the distances between markers adjacent in the resulting orders using standard maximum likelihood techniques[9]. Information from lower resolution maps was used to improve long-range continuity. The TNG map enabled us to order and orient clone contigs, select additional candidate seed points and determine the gap sizes in the late stages of the project.

Contigs established earlier and covering around 15% of the chromosome were also included in the BAC map. Overall, we selected 162 seed clones (~0.30 chromosome 14 equivalents). About 650 clones were assembled in three clone contigs covering more than 99% of chromosome 14 and totalling around 88 megabases (Mb). The largest contig (~85 Mb) physically connects the most proximal genetic marker (D14S261) to the second most telomeric cluster of markers (D14S292), encompassing almost the entire genetic map of the chromosome. We estimate that fewer than five BACs remain to be incorporated to reach full chromosomal coverage.

Publicly available fingerprint data (ref. 5 and http://genome.wustl.edu/gsc/human/Mapping) provided a consistency check for our clone map, which was built independently. Fingerprinted contigs were anchored on our map in several ways. First, fully sequenced clones were digested '*in silico*' and the resulting theoretical restriction pattern aligned against the experimental patterns of the fingerprint clone map[5]. Second, anchoring of fingerprinted contigs was based on the map data associated with markers (including end sequences) contained in the digested clones. These map comparisons indicated that the consistency between fingerprint contigs and our clone scaffold were consistent at a coarse level (some contig junctions were predicted). However, the present clone scaffold was notably more accurate at the finer resolution of local clone ordering and clone overlaps.

The overlap between consecutive BACs resulting from the walking process was estimated to average 20 kilobases (kb) per walking step, and the overlap resulting from random gap closure was estimated to average 52 kb per contig junction. The fraction of redundant sequence was estimated to be 22%, approximately half of which was attributable to suboptimal gap closures.

Note that: (1) most of the redundancy in the clone tiling path is a consequence of the high number of seeds that were required to complete the map in a short time frame[10]; (2) this amount of redundancy is still significantly smaller than that computed for typical draft chromosomes (http://genome.ucsc.edu/); and (3) the fact that fewer than 1% of the selected clones originated from chromosomes other than 14 shows that the strategy is robust with respect to false links of various origins. This contrasts with the fingerprint-based selection of clones, which led to incorrect chromosomal assignment more frequently.

The distance from the telomere was estimated to be about 5 kb on the basis of fragment restriction data involving a yeast artificial

chromosome clone containing the 14q telomere (F.M., unpublished data). The distance to the alphoid centromeric repeats is unknown. However, the current most centromeric BAC already extends 1,200 kb beyond the most proximal marker of previously reported maps[11–13]. Interestingly, this clone contains two markers from our TNG map that exhibit extremely high retention rates in the hybrid lines, indicating that they may be close to the centromere.

The clone coverage of chromosome 14 that has been achieved using essentially an STC strategy is very satisfactory, and compares favourably with the coverage obtained for the human chromosomes that have been completely sequenced[3,4]. The two remaining gaps are located in the subtelomeric part of 14q; subtelomeric regions of many chromosomes are under-represented in most genomic libraries and hence often contain cloning gaps. The largest gap, estimated to be around 600 kb, was subsequently divided into two smaller gaps following the identification of clones through library screening with probes mapped within the gap. The second and more distal gap (20 kb) occurs in the immunoglobulin heavy-chain constant gene region, which contains a number of nearly identical genes and pseudogenes.

Considerations of the optimum design of an STC strategy should include theoretical aspects which correlate the effective depth of the clone end library and the number of seed points to the level of sequence redundancy[10,14], as well as practical aspects such as the sequencing capacity and costs[15], the time schedule for the project and the resolution of the available mapping data. However, a centralized repository of BAC end sequences is the only prerequisite for the construction of a tiling path based on the STC approach. Other mapping resources used in this project were auxiliary and provided useful information for seed selection and validation of map extension. Such a strategy is therefore generally portable to any large-scale sequencing project and is readily compatible with partitioning of the project. □

Received 20 October; accepted 21 December 2000.

1. Venter, J. C., Smith, H. O. & Hood, L. A new strategy for genome sequencing. *Nature* **381,** 364–366 (1996).
2. Mahairas, G. G. *et al.* Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. *Proc. Natl Acad. Sci. USA* **96,** 9739–9744 (1999).
3. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
4. The chromosome 21 mapping and sequencing consortium. The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
5. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409,** 934–941 (2001).
6. Gyapay, G. *et al.* A radiation hybrid map of the human genome. *Hum. Mol. Genet.* **5,** 339–346 (1996).
7. Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13,** 523–535 (1997).
8. Agarwala, R., Applegate, D. L., Maglott, D., Schuler, G. D. & Schaffer, A. A. A fast and scalable radiation hybrid map construction and integration strategy. *Genome Res.* **10,** 350–364 (2000).
9. Lange, K., Boehnke, M., Cox, D. R. & Lunetta, K. L. Statistical methods for polyploid radiation hybrid mapping. *Genome Res.* **5,** 136–150 (1995).
10. Roach, J. C., Siegel, A. F., van den Engh, G., Trask, B. & Hood, L. Gaps in the Human Genome Project. *Nature* **401,** 843–845 (1999).
11. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380,** 152–154 (1996).
12. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282,** 744–746 (1998).
13. Dear, P. H., Bankier, A. T. & Piper, M. B. A high-resolution metric HAPPY map of human chromosome 14. *Genomics* **48,** 232–241 (1998).
14. Batzoglou, S., Berger, B., Mesirov, J. & Lander, E. S. Sequencing a genome by walking with clone-end sequences: a mathematical analysis. *Genome Res.* **9,** 1163–1174 (1999).
15. Siegel, A. F. *et al.* Analysis of sequence-tagged-connector strategies for DNA sequencing. *Genome Res.* **9,** 297–307 (1999).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank D. Cox, P. Dear and D. Cox for unpublished data and helpful discussions.943

Correspondence and requests for materials should be addressed to R.H. (e-mail: heilig@genoscope.cns.fr).

Integration of telomere sequences with the draft human genome sequence

H. C. Riethman*, Z. Xiang*, S. Paul*, E. Morse*, X.-L. Hu*, J. Flint†, H.-C. Chi‡, D. L. Grady‡ & R. K. Moyzis‡

* *The Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsylvania 19104, USA*
† *Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, UK*
‡ *Department of Biological Chemistry, College of Medicine, University of California, Irvine, California 92697, USA*

Telomeres are the ends of linear eukaryotic chromosomes. To ensure that no large stretches of uncharacterized DNA remain between the ends of the human working draft sequence and the ends of each chromosome, we would need to connect the sequences of the telomeres to the working draft sequence. But telomeres have an unusual DNA sequence composition and organization that makes them particularly difficult to isolate and analyse. Here we use specialized linear yeast artificial chromosome clones, each carrying a large telomere-terminal fragment of human DNA, to integrate most human telomeres with the working draft sequence. Subtelomeric sequence structure appears to vary widely, mainly as a result of large differences in subtelomeric repeat sequence abundance and organization at individual telomeres. Many subtelomeric regions appear to be gene-rich, matching both known and unknown expressed genes. This indicates that human subtelomeric regions are not simply buffers of nonfunctional 'junk DNA' next to the molecular telomere, but are instead functional parts of the expressed genome.

Telomeres are essential for genome stability and faithful chromosome replication. The chromatin structures associated with telomeric DNA mediate the many biological activities associated with telomeres, including cell-cycle regulation, cellular ageing, movement and localization of chromosomes within the nucleus, and transcriptional regulation of subtelomeric genes[1,2]. Specialized functions involving telomeric and subtelomeric DNA have evolved in several eukaryotes. For example, frequent subtelomeric gene conversion provides diversity for surface antigens in trypanosomes[3], and rapidly evolving subtelomeric gene families confer selective advantages for closely related yeast strains[4].

Human telomeres end with a stretch of the conserved simple repeat sequence $(TTAGGG)n$[5]. This tract is present at the end of all telomeres and therefore cannot be used to distinguish one telomere from another. To capture single-copy human DNA regions linked to telomeres that are useful for this purpose, we isolated large telomere-terminal fragments of human chromosomes using specialized yeast artificial chromosome (YAC) cloning vehicles called half-YACs[6]. Each half-YAC clone contains a large segment of subtelomeric DNA flanked by the cloning vector sequence at one end and the human telomere repeat sequence, which has been modified to operate as a functional yeast telomere *in vivo*, at the other. Characterization of these clones revealed low-copy subtelomeric repeats adjacent to the $(TTAGGG)n$ sequence[6,7]. Physical mapping experiments on a large group of these half-YAC clones showed that, in most cases, they can stably maintain faithful copies of human telomere-terminal DNA fragments in yeast[8]. By contrast, bacterial artificial chromosome (BAC) libraries used to prepare the human working draft sequence are not expected to contain sequences extending to the telomere, owing to the absence of restriction sites in $(TTAGGG)n$, the effects of length associated with the construction of size-selected DNA recombinant clones, and the genomic instability of these regions[9].