

A High-Resolution Metric HAPPY Map of Human Chromosome 14

Paul H. Dear,¹ Alan T. Bankier, and Michael B. Piper

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom

Received October 9, 1997; accepted November 17, 1997

We have mapped 1001 novel sequence-tagged sites on human chromosome 14. The mean spacing between markers is ~90 kb, most markers are mapped with a resolution of better than 100 kb, and physical distances are determined. The map was produced using HAPPY mapping, a simple and widely applicable *in vitro* approach that is analogous to linkage or to radiation hybrid mapping, but that circumvents many of the difficulties and potential artifacts associated with these methods. We show also that the map serves as a robust scaffold for building physical maps using large-insert clones. © 1998 Academic Press

INTRODUCTION

Several methods are now available for mapping chromosomes or entire genomes, each with advantages and drawbacks. Physical mapping, in which a “contig” of overlapping cloned fragments is assembled, has the advantage of providing not only a map but also a sequencing substrate (Gregory *et al.*, 1997). It is limited, however, by the cloning process on which it relies: regions recalcitrant to cloning lead to uncloseable gaps, while rearranged or coligated fragments, or repeated regions larger than the size of the clones, can lead to distortions (Green *et al.*, 1991; Gregory *et al.*, 1997; Little, 1992). Hence, physical maps are most effective if built over an independently constructed sequence-tagged site (STS) “scaffold.”

Genetic linkage analysis relies on meiotic recombination and segregation to determine distances between polymorphic markers (Ott, 1986). Closely linked markers will rarely be separated by an intervening recombination and hence will consistently cosegregate; unlinked markers will not. Genetic linkage analysis is the only general means for mapping phenotypic traits (reflecting the segregation of alleles of an as yet unknown gene) to a chromosomal location. However, it is ill-suited to detailed mapping (its practical limit of resolution in human is ~1 Mb), is applicable only to polymorphic loci, and gives genetic distances that do

not reliably reflect physical distances (Djilalisaiah *et al.*, 1996; Hubert *et al.*, 1994; Wang and Smith, 1994).

Radiation hybrid (RH) mapping also relies on segregation analysis, but the genome is fragmented by irradiation rather than by recombination and markers segregate among hybrid cells made by fusing the irradiated donor with a suitable host (Stewart and Cox, 1997). In contrast to linkage mapping, RH mapping is applicable to monomorphic markers and offers a higher resolution, depending on the frequency of radiation-induced breaks. Making panels of radiation hybrids, however, is not trivial, and biological activity of the donor fragments (particularly centromeres and telomeres) can bias their segregation and retention among hybrids, complicating analysis and distorting the resulting maps (Jones, 1996; Orphanos *et al.*, 1995; Raeymakers *et al.*, 1995; Sapru *et al.*, 1994; Wang and Smith, 1994). It has also been suggested (Teague *et al.*, 1996) that local variations in chromosome structure may affect the frequency of radiation-induced breakage.

HAPPY mapping (Dear, 1997; Dear and Cook, 1989, 1993; Walter *et al.*, 1993) is also a form of segregation analysis, but is an entirely *in vitro* process (Fig. 1). Genomic DNA is broken at random by irradiation and size-selected using pulsed-field gels, and the fragments are aliquotted into 96 samples (the “mapping panel”). Each member of the panel contains less than one genome’s worth of fragments (1–2 pg) and therefore contains a random subset of the genome.

The members of the mapping panel are preamplified using polymerase chain reaction (PCR), to amplify all of the markers that they contain. (This step is analogous to growing up the members of an RH panel for bulk DNA isolation). They are then screened for specific STS markers (which can be either polymorphic or monomorphic), and cosegregation frequencies between markers reflect the distance between them: the more closely linked the markers, the less likely they are to be broken apart and so the more often they cosegregate. In this way, the order and distances between markers can be determined.

The range and resolution afforded by the mapping panel depend on the size of the fragments from which it is made: small fragments allow high-resolution map-

¹ To whom correspondence should be addressed. Telephone: [44] (1223) 402190. Fax: [44] (1223) 412178. E-mail: phd@mrc-lmb.cam.ac.uk.

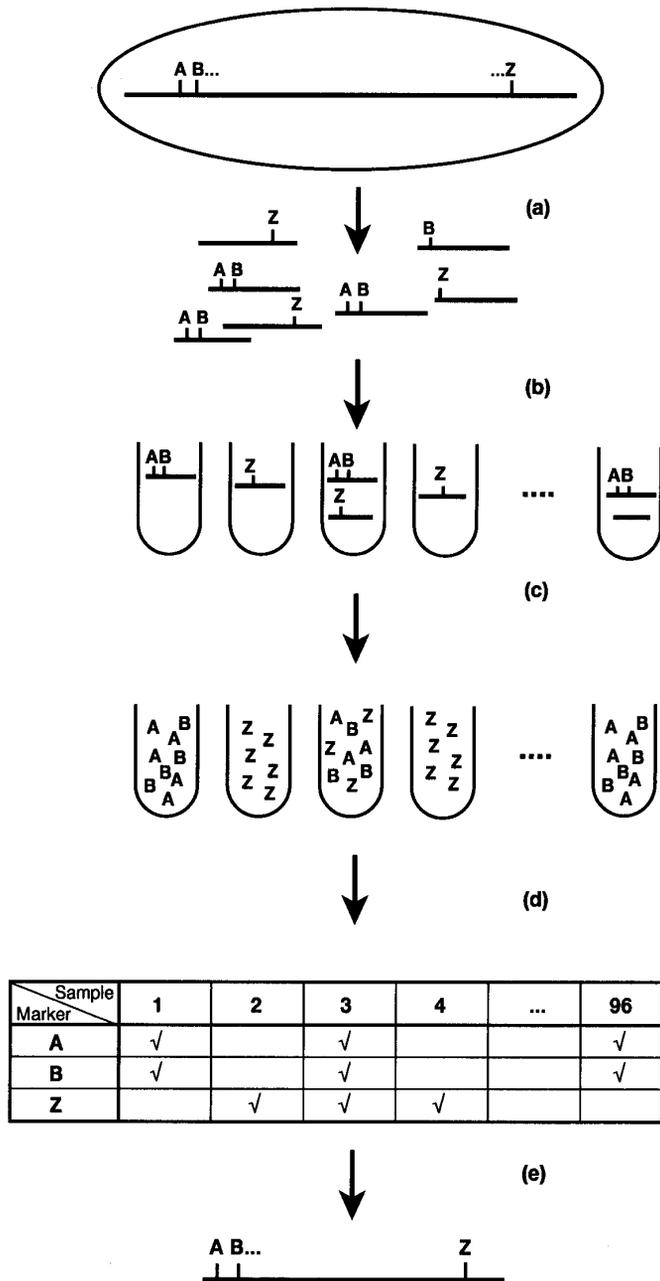


FIG. 1. How HAPPY mapping works. (a) Genomic DNA is prepared (for example from blood lymphocytes), broken by irradiation, and size-selected to give a pool of random fragments. (b) 96 aliquots (the mapping panel) are taken, each containing only 1–2 pg of fragments and hence each containing a random subset of the genome. (c) The panel is preamplified (for example, by repeat element-mediated PCR); this gives enough material for subsequent marker typing, but preserves the marker content of each aliquot. (d) The panel is screened for STS markers (A, B, ..., Z). (e) Linked markers (such as A and B) are found to cosegregate frequently; unlinked markers (e.g., B and Z) do not. Analysis of these results allows marker order and distance to be determined.

ping over short distances; larger fragments allow long-range mapping, but at the expense of resolution. Hence, the choice of fragment size is a compromise between reliable resolution of closely spaced markers (down to approximately $\frac{1}{10}$ of the fragment size) and the require-

ment to detect robust linkage between more widely spaced markers (which is possible only over distances less than approximately 0.5–0.7 of the fragment size).

A HAPPY mapping panel is simple to produce (requiring days or weeks), can provide any chosen level of resolution (depending on the size of the fragments), and is immune to artifacts caused by biological activity of the fragments, to cloning artifacts, or to effects of chromosome structure.

Chromosome 14 is acrocentric, with a short arm and pericentromeric region composed largely of repeated sequences and a long-arm of ~90 Mb (Hudson *et al.*, 1995; Pandit *et al.*, 1995). To produce our map, we made a total of three mapping panels. A short-range panel (containing genomic fragments of <1.5 Mb) allows high-resolution mapping over relatively short distances (<1 Mb). Mid- and long-range panels (fragment sizes of ~2 and >2.5 Mb, respectively) were used to bridge gaps and to assemble the complete map. These panels were prepared from normal human male genomic DNA and hence can be used to map any human chromosome.

For markers, we chose to map STSs consisting of unique sequence flanked by repeat element motifs (LINE, *Alu*, or $[CA]_n$; Fig. 2). This format means that all marker sequences can be preamplified using PCR with a cocktail of repeat-element primers or any individual marker can be amplified using its specific primers (see below).

We have produced a high-resolution, densely populated map spanning the entire long arm, in which physical distances between markers are determined. We have also shown that this map is an effective scaffold on which to build contigs of large-insert clones.

MATERIALS AND METHODS

Chromosome-wide marker sequences and primer design. Chromosome-wide marker sequences were isolated from flow-sorted chromosome 14 (FSCs, from cell line HRC160) or from human–mouse or human–hamster hybrids (NA10479 and NA11535, respectively; Coriell Cell Repositories). Sixty nanograms of hybrid DNA, or 1000 FSCs, was amplified in 50 μ l [50 mM KCl, 10 mM Tris–HCl, pH 8.3, 1 mM $MgCl_2$, 200 μ M each dNTP, PCR primer(s) at 1 μ M each, and either 1.25 U (for hybrid DNAs) or 10 U (for FSCs) *Taq* polymerase; 93°C for 5 min followed by 22 cycles (hybrid DNAs) or 35 cycles (FSCs) of 94°C for 20 s, 65°C for 30 s, 72°C for 2 min; final extension at 72°C for 4 min]. Hybrid DNAs were amplified either with primers Alu1 (GGATTACAGGYRTGAGCCA; Liu *et al.*, 1993) plus Alu2 (RCCAYTGCCTCCAGCCTG; Liu *et al.*, 1993) or with these primers plus LINE2 (CACGTTGTGCACATGTACC; Scott *et al.*, 1987). FSCs were amplified either with Alu1 plus Alu2 or with the dinucleotide repeat primer $[TG]_{10}$. Products of 250–800 bp were selected by electrophoresis through a 0.8% low melting temperature agarose minigel in TBE (90 mM Tris borate, 2 mM EDTA, pH 8.3), followed by digestion with β -agarase, phenol extraction, and ethanol precipitation.

Fragments were end-repaired using T4 DNA polymerase and kinased using T4 polynucleotide kinase, essentially as described in Wang *et al.* (1994). The fragments were then cloned into the *HincII* site of m13mp18 (Messing and Bankier, 1989). High-throughput phage growth, single-stranded DNA purification, and sequencing, using Perkin–Elmer ABI dye primers and Amersham ThermoSequenase, were carried out in 96-well microtiter trays using a Biomek



FIG. 2. STS marker format. Each of our markers consists of a stretch of unique sequence flanked by repeat elements (*Alu*, LINE, or [TG]_n; shaded rectangles). Hence, all marker sequences can be preamplified using repeat element-mediated PCR with a cocktail of repeat primers (r1, r2). Any one marker can be selectively amplified using its specific primers (s1, s2) in a conventional PCR.

1000 workstation (Bankier, 1993; Smith *et al.*, 1990). The sequences, obtained from an ABI 373 automated sequencer, were trimmed to exclude poor-quality data, but were not subjected to detailed editing. A database of the sequences was maintained using the Staden package (Dear and Staden, 1991).

Specific PCR primers were designed where possible to amplify within the unique region of each nonredundant sequence, either manually or by our own software (P.H.D., unpublished results).

Anchor markers. To align our map to existing maps, we generated anchor markers. Each anchor marker is an inter-*Alu* sequence (which can be placed on our map) derived from a PAC or cosmid clone known to contain a particular genetic or RH marker (for example, a D14S sequence). Hence, each anchor marker lies within 190 kb (the size of the largest PAC clone; mean size 110 kb) of the corresponding genetic or RH marker.

For isolating anchor marker sequences, PAC clones containing the relevant genetic or RH marker were found by PCR screening the RPC11 human genomic PAC library (Ioannou and de Jong, 1996). Cosmids from the V_h region were provided by Ian Tomlinson. Clones were toothpicked into a 50- μ l reaction (1.5 mM MgCl₂, 1 μ M concentrations of each primers Alu1 + Alu2, 1.25 U *Taq* polymerase, other components as above). After cycling (93°C for 5 min, then 33 cycles of 94°C for 20 s, 62°C for 30 s, 72°C for 2 min, final extension of 72°C for 4 min), products were purified using Qiagen PCR purification columns. Sequencing and specific primer design were performed as for the chromosome-wide markers (above). Anchor marker primers were tested to verify that they amplified a product of the correct size from the clone from which they originated.

Isolation of genomic DNA. Peripheral blood lymphocytes were isolated from fresh, citrated normal male human blood (Ficoll-Paque, Pharmacia), resuspended in PBSG [phosphate-buffered saline (Sigma) plus 1% w/v glucose] at 10⁷ cells/ml, and mixed with low-melting-point agarose in PBSG at 37°C to a final concentration of 1% agarose and 2.5 \times 10⁶, 3 \times 10⁵, or 1 \times 10⁵ cells/ml. This mixture was taken up into \sim 40 glass capillaries (130 mm \times 1.2 mm internal diameter; Supracaps, 100 μ l, Scientific Laboratory Supplies) and chilled at 4°C for 10–15 min. The set agarose “strings” were allowed to fall under gravity from the capillaries into \sim 150 ml lysis solution (10 mM Tris–HCl, pH 7.5, 1 mM EDTA, 1% w/v lithium dodecyl sulfate) and incubated at 4°C. Lysis solution was replaced at intervals of 15 min, 30 min, and then three times at hourly intervals, and finally after overnight incubation; strings were stored at 4°C in lysis solution for up to several months.

Preparation of mapping panels. Three mapping panels were prepared, differing in the size of DNA fragments they contain and hence in the resolution and range which they afford.

For the A-panel (short-range), agarose strings containing 10⁵ and 2.5 \times 10⁶ cells/ml were irradiated (35 J/kg) using a Gravatom RX30/55M ¹³⁷Cs source. The low-concentration string was placed in a 140 mm \times 3 mm well in a 3-mm thick gel (LKB Gene Navigator system) of 1% chromosomal grade agarose (Bio-Rad) in 0.5 \times TBE, flanked by a short segment of the irradiated high-concentration string and by *Saccharomyces cerevisiae* chromosomal markers (Bio-Rad). Samples were sealed into the gel using 1% agarose in 0.5 \times TBE prior to electrophoresis (180 V, 100-s pulse time for 12 h).

For the B-panel (midrange), strings as above were irradiated with 10 J/kg. After a first stage of electrophoresis [Bio-Rad CHEF-DRIII system; 0.8% chromosomal grade agarose; TAE buffer (40 mM Tris acetate, 1 mM EDTA); 1200-s pulse, 96° switch angle/1500 s, 100°/

1800 s, 106°; each for 14 h at 2 V/cm, 14°C], the upper part of the gel including sample wells was removed, and electrophoresis was repeated as before. This two-stage electrophoresis ensures that small fragments trapped in the irradiated string cannot continue to leach slowly into the gel during the second stage of electrophoresis (P.H.D., unpublished results). *Schizosaccharomyces pombe* size standards (Bio-Rad) were substituted for *S. cerevisiae*.

Strings for the long-range C-panel (3 \times 10⁵ cells/ml, plus a high-density control of 2.5 \times 10⁶/ml) were not irradiated, as unavoidable breakage occurring during DNA preparation was sufficient to provide large fragments. Electrophoresis was again performed in two stages (LKB Gene Navigator system; 0.8% chromosomal grade agarose; 0.5 \times TBE; 35 V, 6000-s pulse), with the sample well being excised between the first stage (71 h) and the second (96 h). Markers were *S. pombe* chromosomes.

After electrophoresis, the sides of the gel, containing yeast standards and the high-concentration human DNA fragments, were excised, stained with ethidium bromide, and visualized with UV light. The central portion was incubated for >3 h in TE (10 mM Tris–HCl, pH 7.5, 1 mM EDTA), followed by 30 min in 0.1 \times TE, both at 4°C. The gel may be stored in 0.1 \times TE for several weeks.

The gel was removed from 0.1 \times TE in a DNA-free environment, and a glass capillary, connected through a filter to a mouthpiece, was used to remove 96 plugs of agarose from the gel, at a point containing fragments of the desired size as judged by reference to the size standards (<1.5 Mb for the A-panel, \sim 2 Mb for the B-panel, and >2.5 Mb for the C-panel). The internal diameter of the capillary was either 0.56 mm (A- and B-panels) or 0.70 mm (C-panel). One plug was transferred to each well of a thermocycler-compatible microtiter plate and overlaid with \sim 30 μ l of light mineral oil (Sigma).

Panel preamplification and evaluation. Preamplification of the panels was performed using two rounds of repeat-element-mediated (REM)-PCR, under conditions designed to rigorously exclude contaminating DNA. (Subsequent marker screening does not require such precautions.) Reagents were added to the agarose plugs to give a total volume of 5 μ l containing 50 mM KCl, 10 mM Tris–HCl, pH 8.3, 200 μ M each dNTP, PCR primers at 1 μ M each, and 0.25 U *Taq* polymerase. For the A-panel, Alu1, Alu2, LINE, and [TG]₁₀ primers were used with 1 mM MgCl₂. For the B- and C-panels, Alu1 and Alu2 primers were used with 1.5 mM MgCl₂. Amplification conditions were 93°C for 5 min, then 22 cycles of 94°C for 20 s, 65°C (A-panel) or 62°C (B- and C-panels) for 30 s, 72°C for 2 min, and a final extension of 72°C for 4 min. Immediately upon completion of the last cycle (before the agarose in the reactions set), each sample was diluted to 100 μ l with water and transferred to a fresh microtiter plate, overlaid with light mineral oil, and stored at -70° C.

For the second-round preamplification, first-round products were diluted a further fivefold in water; 4- μ l samples of this were reamplified in 10 μ l (as above, but for 25 cycles), diluted to 600 μ l with water, and stored at -70° C. Four microliters of this material was used for each routine marker typing (below).

Three-round preamplification was also tested, by reamplifying 4 μ l of the diluted second-round products in a 10- μ l reaction for 15 cycles (1.5 mM MgCl₂, 62°C annealing; other conditions as for second-round preamplification), then diluting again to 600 μ l with water. As before, 4 μ l of this material was used as template in typing reactions (below).

Panels were evaluated initially by typing 12–20 markers (below), to determine the mean DNA content of the panel members.

Marker typing. Markers were typed in 10- μ l reactions containing 4 μ l of the preamplified panel, 50 mM KCl, 10 mM Tris-HCl, pH 8.3, 200 μ M each dNTP, 1 mM MgCl₂, marker-specific primers at 1 μ M each, 0.25 U *Taq* polymerase; 93°C for 5 min then 33 cycles of 94°C for 20 s, 60°C for 30 s, 72°C for 1 min. In some cases conditions were modified (1.5 mM MgCl₂ and/or annealing temperature of 52–66°C) to improve amplification. Eight microliters of SyBr dye [15% Ficoll 400 (Pharmacia), 0.15 mg/ml bromophenol blue, 4 \times SyBr Green I (FMC Bioproducts) in 1 \times TBE] was added, and 14- μ l aliquots were analyzed on high-density minigels (Hybaid Electro-4; 3% Nu-Sieve 3:1 agarose in 0.5 \times TBE; 200 V, 18 min). Results were entered manually using a computer program (P.H.D., unpublished results).

Pairwise LOD and distance calculations. Pairwise lod scores were calculated using an essentially standard algorithm (P.H.D., unpublished results). Briefly, the log of the probability of obtaining the observed pattern of segregation between two markers is calculated for all hypothetical values of θ (the probability of one or more breaks falling between the two markers) between 0 and 1 in increments of 0.01. (Breaks are assumed to occur at random). The lod score between the two markers is the highest log likelihood, relative to that at $\theta = 1$; the θ at which it occurs is the estimate of the breakage frequency between them.

If the size selection of fragments composing the mapping panel is perfect, then θ is linearly related to physical distance, up to a limiting value of $\theta = 1$ (corresponding to a physical distance equal to the fragment size). In practice, imperfections in the size selection (particularly the presence of fragments smaller than expected) cause a departure from this linear relation (Dear and Cook, 1993; P.H.D., unpublished results). θ values were therefore rescaled to be linearly proportional to physical distance, using a mapping function derived previously (Dear and Cook, 1993).

Contig assembly. The RPCII human PAC library (Ioannou and de Jong, 1996) was screened by PCR for markers in the region to be covered, following standard procedures. Additional clones were isolated from the ICI human YAC library (Anand *et al.*, 1990) where necessary. Contig assembly was performed manually, based on the STS content of each clone. Clone sizes were determined by pulsed-field gel electrophoresis of either linearized PAC clones prepared using alkaline lysis (Ioannou and de Jong, 1996) or (for YACs) of yeast DNA prepared essentially as in Bautsch *et al.* (1997).

RESULTS

Markers and Mapping Panels

The majority of marker sequences were obtained by PCR amplification of rodent-human hybrids containing chromosome 14, with either *Alu* primers (1520 sequences) or *Alu* + LINE primers (103 sequences), followed by cloning in M13 and sequencing as described. A further 433 sequences were obtained following amplification of flow-sorted chromosome 14 with various combinations of *Alu* and [TG]₁₀ primers, again followed by cloning and sequencing of amplification products. A total of 191 anchor marker sequences were also derived from *Alu* PCR products of PAC or cosmid clones containing known genetic or RH markers. Of 2247 unique sequences derived in these ways, 1626 were suitable for PCR primer design.

Preliminary evaluation of the mapping panels revealed retention rates (the mean probability of finding a given marker in any given panel member) of 0.47, 0.34, and 0.18 for the A-, B-, and C-panels, respectively. Two or more copies of a marker in a panel member are indistinguishable from a single copy so, assuming a Poisson distribution of markers, these values corre-

spond to mean DNA contents of 0.64, 0.42, and 0.20 genome equivalents per panel member for the three panels, respectively.

Marker Typing and Map Assembly

All markers were initially typed against the short-range A-panel (preamplified by two rounds of REM-PCR). Of 1626 primer pairs tested, 947 were typed successfully (as "first-rate" markers) against the A-panel. (Of these, approximately 850 were typed successfully on the first attempt; the remainder were retyped with altered magnesium concentrations or annealing temperatures to produce successful amplification). A further 290 pairs gave partial or ambiguous results (anomalously low or high numbers of positives, faint amplification products, or nonspecific products in addition to those expected) even after optimization of PCR conditions; these were designated "second-rate" markers and were excluded from the initial analysis. A total of 294 failed to amplify under standard conditions, were detectable in too few panel members for analysis, or gave only nonspecific products. Ninety-five were found to amplify multicopy sequences (as judged by their being present in all or most members of the panel).

Duplicate typing of 20 of the first-rate markers on the A-panel revealed an average error rate of 1.9%, an error being defined as a marker being typed as positive against a given panel member on one occasion and negative on another, or vice versa. This was true regardless of whether both typings were performed on the same portion of the panel or on portions derived from two different second-round preamplifications. Hence, it was not considered worthwhile to perform duplicate typings routinely, particularly as errors of comparable frequency might be expected to arise during initial creation of the panel (due to contamination or failure in the initial stages of preamplification) and would be undetectable by duplicate typings.

The effectiveness of the three-round preamplification protocol was tested by retyping 30 first-rate and 12 second-rate markers against the A-panel after preamplification with three rounds of REM-PCR. The mean error rate for first-rate markers (i.e., discrepancies between these typings and those done after only two rounds of preamplification) was 1.5%, i.e., no greater than that observed after two rounds of preamplification. There was no deterioration in the quality or clarity of the results. Of the 12 second-rate markers tested in this way, 5 gave better (i.e., clearer) results, 3 gave worse results, and 4 gave results comparable to those seen after two rounds of preamplification.

From the two-point lod scores, all first-rate markers were sorted into 57 linkage groups (A-groups, with each member of a linkage group being linked to at least one other member by a lod score of ≥ 6) and 61 unlinked markers ("singletons"). (A lod threshold of 6 was chosen because with 947 markers there are $\sim 5 \times 10^5$ pairwise

lods to consider; a lod threshold of 6 makes it unlikely that any two markers will appear linked by chance alone.) At least one member of each linkage group, and each singleton, was tested by PCR against the NIGMS2 panel of rodent-human monochromosomal hybrids (Coriell Cell Repositories) to determine its chromosomal location and hence that of any others in the same A-group. In this way 120 markers were found to lie on chromosome 16 (these having originated from the hybrid line 10479, which is known to contain part of chromosome 16 as well as chromosome 14; Sunden *et al.*, 1996); 27 markers (mostly derived from flow-sorted chromosome 14) lay on various other human chromosomes or could not be assigned. Non-chromosome 14 markers accounted for most of the singletons. These non-chromosome 14 markers were rejected from analysis, leaving 800 first-rate markers in 42 linkage groups plus 11 singletons. Maps were then produced for each of these linkage groups using a distance geometry algorithm, *DGMap* (Newell *et al.*, 1995).

Singletons and one or more markers lying at or close to the ends of each A-group (102 markers in total) were then typed on the midrange B-panel, and pairwise lod scores were calculated as before. These results were then used to link the A-groups into larger linkage groups (B-groups), at a lod threshold of 4.

Finally, the long-range C-panel was used to link together markers from the ends of the B-groups (a total of 14 markers being typed on the C-panel), again at a lod threshold of 4. Only one gap (between markers a2155 and a1891; Fig. 3) was not rigorously closed by this method (lod scores 2.1 and 2.3 with the B- and C-panels, respectively).

The majority of the gap closures made by the B-panel were supported by A-panel LODs of between 2 and 6, increasing our confidence in the reliability of the map. In the same way, most gaps closed by the C-panel showed B-panel LODs of between 1 and 3. There were no conflicts in marker order (as determined by each of the three panels) at a LOD of >2 .

To unify distances *between* A-groups (i.e., the gaps that were closed using the B- or C-panels) with those *within* A-groups (measured using the A-panel), two approaches were used. Where the lod score measured by the A-panel was between 2 and 6, the B- or C-panel data were used to confirm linkage but distance was based on A-panel data. Where there was no significant A-panel linkage (LOD <2), the distance measured by the B- or C-panel was rescaled by extrapo-

lating from those distances where the A- and B- or C-panels overlap.

Placement of Additional Markers

The second-rate markers, which were rejected from initial mapmaking, were reexamined. Those having a lod score of >3 to one or more first-rate markers were assigned to the likeliest interval between first-rate markers, using a breakage minimization algorithm analogous to the recombination minimization approach to genetic linkage analysis. Two hundred one such markers were placed in this way (Fig. 3).

Map Verification, Contig Construction, and Alignment with Other Maps

Previous studies (Dear and Cook, 1993) showed that the A-panel should offer a resolution of better than 100 kb. Support for this comes from the observation that whenever two or more anchor markers were derived for the same genetic or RH marker, they mapped to within <180 kb (in most cases <50 kb) of one another. Since the anchors are derived from PAC clones with a maximum size of ~ 190 kb, the maximum *true* distance between anchors is 190 kb. Hence, these results are consistent with a mapping error of <100 kb.

To verify further the short-range accuracy of the map, we constructed a contig of PAC clones (supplemented where necessary by YACs) spanning a region of ~ 1 Mb, by screening the appropriate clone libraries with the first-rate STSs from this region (Fig. 4). Conflicts between the marker order as determined by HAPPY mapping and that determined by the contig are all at distances that are <100 kb (as measured on the HAPPY map) and within a distance spanned by a single PAC clone.

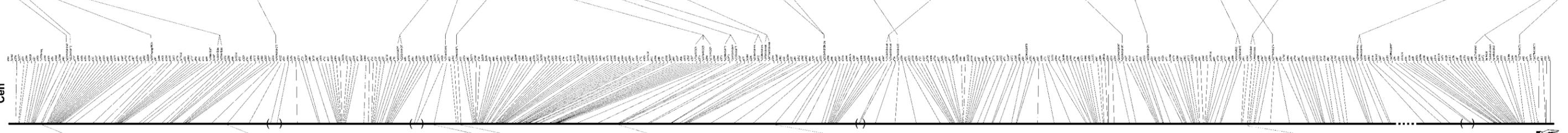
Long-range accuracy of the map is verified by alignment of anchor markers with genetic, RH, and STS-content maps of this chromosome (Fig. 3).

DISCUSSION

The HAPPY map covers the long arm of chromosome 14 at a mean marker spacing of ~ 90 kb. Moreover, the resolution of ordering most of these markers (the first-rate markers) is shown to be better than 100 kb. The accuracy of placing the second-rate markers varies from case to case, depending on the quality of the

FIG. 3. HAPPY map of human chromosome 14. The heavy line represents the HAPPY map, connected to the STS content map (**above**) and RH map (**below**). (Only those markers shared by the HAPPY map are shown on the RH and STS maps.) Marker names prefixed by "2" are second-rate markers; they are assigned only to an interval between flanking first-rate markers. Markers prefixed by an asterisk are anchor markers; the name of the corresponding RH/genetic marker follows the marker name; only first-rate anchors are shown connected to the RH and STS content maps. Parentheses on the HAPPY map delimit short segments that can be inverted at odds of $<1000:1$. Resolution elsewhere is approximately 100 kb (approximately 1/1000 of the total map length). The dashed segment represents a gap that was not rigorously closed (see text). "Cen" and "Tel" indicate centromere and telomere, respectively. The shaded rectangle indicates the region spanned by the contig in Fig. 4. Marker names in GenBank have the form h14a#, where # represents the marker number. For example, marker *a2194 has GenBank name h14a2194; marker 2a1304.34 has GenBank name h14a1304.34.

STS content map



RH map

01-1023

01-1024

01-1025

01-1026

01-1027

01-1028

01-1029



Cen

()

()

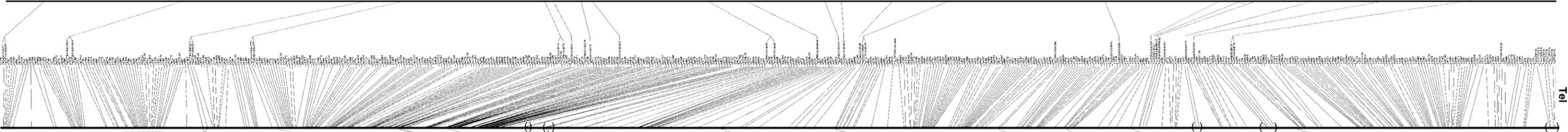
()

()

()

()

STS content map



tel



RH map

140000 140001 140002 140003 140004 140005 140006 140007 140008 140009 140010 140011 140012 140013 140014 140015 140016 140017 140018 140019 140020 140021 140022 140023 140024 140025 140026 140027 140028 140029 140030 140031 140032 140033 140034 140035 140036 140037 140038 140039 140040 140041 140042 140043 140044 140045 140046 140047 140048 140049 140050 140051 140052 140053 140054 140055 140056 140057 140058 140059 140060 140061 140062 140063 140064 140065 140066 140067 140068 140069 140070 140071 140072 140073 140074 140075 140076 140077 140078 140079 140080 140081 140082 140083 140084 140085 140086 140087 140088 140089 140090 140091 140092 140093 140094 140095 140096 140097 140098 140099 140100

marker typing. However, second-rate anchor markers generally map to within one or two intervals of corresponding first-rate anchor markers; this suggests that most second-rate markers are placed to an accuracy of better than 2–300 kb. Because both the mean marker spacing and the marker resolution are of the same order as the size of typical PAC or BAC clones (~100 kb), assembly and verification of physical maps are greatly simplified.

The distribution of markers along the map is clearly not uniform. In part this is due to the random scattering of markers, but there are evidently regions of higher than average marker density. These we presume to reflect the distribution of *Alu* repeat elements, as most of our markers are derived from inter-*Alu* sequences.

Fortuitously, 45 of the mapped markers contain simple sequence repeats (mostly di- and tetranucleotide motifs in approximately equal proportions) long enough to be likely to be polymorphic. (A further 5 markers, which were originally obtained by PCR of flow-sorted chromosomes using the primer [TG]₁₀, have the dinucleotide repeat [TG]_n at one or both ends).

On a large scale, the HAPPY map generally confirms the order of anchor markers on existing RH and STS-content maps (Hudson *et al.*, 1995) of this chromosome (Fig. 3) and also that of genetic linkage maps (not shown). There are two conflicts between our map and the STS-content map. One involves D14S66 and D14S274; here we agree with the genetic map, where these markers are well resolved, but disagree with the STS map. The other concerns D14S1068 and D14S1055; again, we disagree with the STS map. These markers lie in a region where there is considerable disparity between the STS and the RH maps and between the STS and the genetic maps. The expected resolution of the STS content map leads us to believe that our marker order is correct in these cases.

Our distance estimates are (like all measurements based on linkage analysis) subject to statistical error and also to any errors in creating or typing the mapping panels. They are also sensitive to inaccuracies in the mapping function by which we relate breakage frequencies (θ) between markers to physical distance. Although we are confident that our mapping function is approximately correct, there may be a slight expansion or contraction of marker-dense regions relative to sparsely populated areas of the map. Because HAPPY mapping is an entirely *in vitro* process, however, we do not expect any gross distortion of the map due to biological activity of centromeres, telomeres, or other sequences nor to the effects of chromosome structure. The availability of physical distances between markers (as well as marker order) further facilitates the assembly of overlying contigs. Distance information can show whether a gap is likely to be bridged by a single clone or whether chromosome walking will be required. It can also draw attention to major insertions or deletions in large-insert clones such as YACs.

One of the major distinctions between HAPPY mapping and RH mapping is that fragments are selected and preamplified *in vitro* in the former, but by cloning and propagation in mammalian cells in the latter. Our two-step preamplification (a total of 47 cycles of REM-PCR) provides sufficient material for approximately 19,000 marker typings. However, the three-round preamplification (a total of 62 cycles) provides enough material for approximately 2.8 million typings, with no apparent deterioration in the quality of the results. Hence, a HAPPY mapping panel can be considered as an effectively unlimited resource.

The use of REM-PCR (chosen because of the robust, though selective, amplification that it provides), however, limits our choice of markers to those that are flanked by repeat elements, whereas most of the human STSs now available are not. We have previously shown (Dear and Cook, 1993) that HAPPY mapping panels may be preamplified using “whole-genome” (rather than REM) PCR, enabling any sequence to be mapped. Future panels will be made using an enhanced version of this whole-genome approach.

In contrast to panels of radiation hybrids or clone libraries, a HAPPY mapping panel contains no background or host DNA. This simplifies the typing of markers and should also make it possible to map markers based on arbitrary primers (such as RAPDs or AFLPs). This may be of particular advantage in mapping plant or animal genomes for which the cost of large numbers of sequence-specific PCR primers is prohibitive.

Whereas most mapping methods are limited in the resolution that they can achieve, the reverse is true of HAPPY mapping. Resolution down to a few kilobases is possible by using small DNA fragments (Dear and Cook, 1993; Walter *et al.*, 1993), but the *maximum* distance between markers is limited by the size of the largest fragments that can be size-selected using pulsed-field gels. In practice, this means that no two consecutive markers may be further than ~2 Mb apart, and their mean spacing must be several-fold less than this to allow for the random distribution of markers. Hence, very sparse maps (which are readily achieved by genetic linkage mapping or fluorescence *in situ* hybridization in most species) cannot be produced.

We believe that HAPPY mapping will be of use both in human genome analysis (where it can resolve ambiguities or distortions in RH and genetic maps or provide a reliable metric scaffold for physical maps) and in mapping nonhuman genomes (particularly those for which RH panels are difficult to produce) at moderate or high resolution.

Map information, PCR primer sequences, amplification conditions, and amplicon sequences are available on our web site (<http://www.mrc-lmb.cam.ac.uk/happy/happy-home-page.html>). STS sequences have also been submitted to GenBank (see legend to Fig. 3). We are willing to make our mapping panels (which can be used to map any human chromosome) and limited quantities of PCR primers available to interested parties.

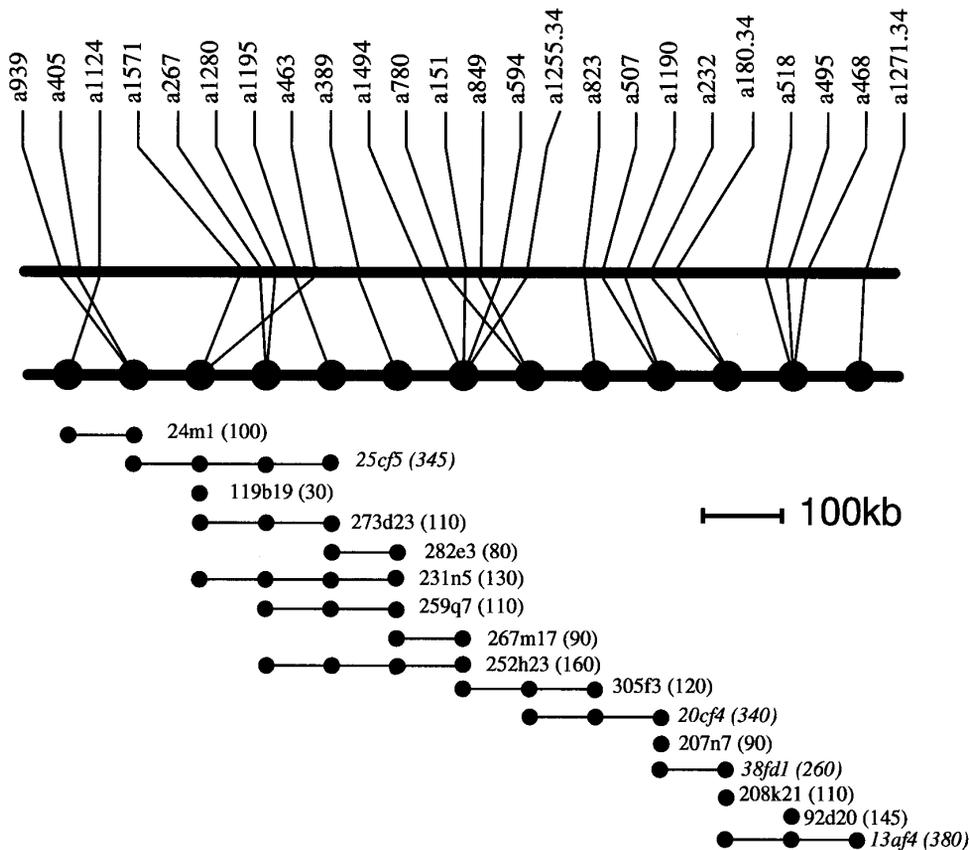


FIG. 4. Contig assembly using the HAPPY map as a scaffold. A portion of the HAPPY map, spanning ~1 Mb, is shown uppermost. Below it is the contig map of the corresponding region; large filled circles represent STS markers (not all STSs can be ordered using the contig map), connected to the corresponding markers on the HAPPY map. Below the contig map are shown the PAC clones and YACs (italic names) forming the contig. Clone sizes in kb are indicated in parentheses. A small filled circle on a clone indicates that it contains all of the STS markers in the large filled circle above. The scale bar refers to distances on the HAPPY map.

ACKNOWLEDGMENTS

We thank Cordelia Langford and Nigel Carter for supplying flow-sorted chromosomes and Ian Tomlinson and Anil Menon for providing clones from which anchor markers were derived. Human genomic PAC and YAC clone libraries were provided by the MRC HGMP Resource Centre. P.H.D. was funded in part by Techné (UK) Ltd. and by the Isaac Newton Trust.

REFERENCES

- Anand, R., Riley, J. H., Butler, R., Smith, J. C., and Markham, A. F. (1990). A 3.5 genome equivalent multi access YAC library: Construction, characterisation, screening and storage. *Nucleic Acids Res.* **18**: 1951–1956.
- Bankier, A. T. (1993). The use of robotic workstations in DNA sequencing. In "DNA Sequencing Protocols" (H. G. Griffin and A. M. Griffin, Eds.), Vol. 23, pp. 373–383, Humana, Totowa, NJ.
- Bautsch, W., Römmling, U., Schmidt, K. D., Samad, A., Schwartz, D. C., and Tümmler, B. (1997). Long-range restriction mapping of genomic DNA. In "Genome Mapping—A Practical Approach" (P. H. Dear, Ed.), pp. 281–313, IRL Press, Oxford.
- Dear, P. H. (1997). HAPPY mapping. In "Genome Mapping—A Practical Approach" (P. H. Dear, Ed.), pp. 95–124, IRL Press, Oxford.
- Dear, P. H., and Cook, P. R. (1989). Happy mapping: A proposal for linkage mapping the human genome. *Nucleic Acids Res.* **17**: 6795–6807.
- Dear, P. H., and Cook, P. R. (1993). Happy mapping: Linkage mapping using a physical analogue of meiosis. *Nucleic Acids Res.* **21**: 13–20.
- Dear, S., and Staden, R. S. (1991). A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.* **19**: 3907–3911.
- Djilalisaiah, I., Benini, V., Daniel, S., Assan, R., Bach, J. F., and Caillatzucman, S. (1996). Linkage disequilibrium between HLA class-II (DR, DQ, DP) and antigen-processing (LMP TAP, DM) genes of the major histocompatibility complex. *Tissue Antigens* **48**: 87–92.
- Green, E. D., Riethman, H. C., Dutchik, J. E., and Olsen, M. V. (1991). Detection and characterization of chimeric yeast artificial-chromosome clones. *Genomics* **11**: 658–669.
- Gregory, S. G., Soderlund, C. A., and Coulson, A. (1997). Contig assembly by fingerprinting. In "Genome Mapping—A Practical Approach" (P. H. Dear, Ed.), pp. 228–254, IRL Press, Oxford.
- Hubert, R., MacDonald, M., Gusella, J., and Arnheim, N. (1994). High-resolution localization of recombination hot-spots using sperm typing. *Nat. Genet.* **7**: 420–424.
- Hudson, T. J., Stein, L. D., Gerety, S. S., Ma, J., Castle, A. B., Silva, J., et al. (1995). An STS-based map of the human genome. *Science* **270**: 1945–1954.
- Ioannou, P. A., and de Jong, P. J. (1996). Construction of bacterial artificial chromosome libraries using the modified P1 (PAC) system. In "Current Protocols in Human Genetics" (N. C. Dracopoli, Ed.), Vol. 5, pp. 1–24, Wiley, New York.
- Jones, H. B. (1996). Pairwise analysis of radiation hybrid mapping data. *Ann. Hum. Genet.* **60**: 351–357.

- Little, P. (1992). Mapping the way ahead. *Nature* **359**: 367–368.
- Liu, P., Siciliano, J., Seong, D., Craig, J., Zhao, Y., and de Jong, P. J. (1993). Dual *Alu* polymerase chain-reaction primers and conditions for isolation of human-chromosome painting probes from hybrid cells. *Cancer Genet. Cytogenet.* **65**: 93–99.
- Messing, J., and Bankier, A. T. (1989). The use of single-stranded DNA phage in DNA sequencing. In "Nucleic Acids Sequencing—A Practical Approach" (C. J. Howe and E. S. Ward, Eds.), pp. 1–37, IRL Press, Oxford.
- Newell, W. R., Mott, R., Beck, S., and Lehrach, H. (1995). Construction of genetic maps using distance geometry. *Genomics* **30**: 59–70.
- Orphanos, V., Greaves, M., Santibanezkoref, M., Fox, M., Edwards, Y. H., and Boyle, J. M. (1995). A radiation hybrid panel for human chromosome 6q. *Mamm. Genome* **6**: 285–290.
- Ott, J. (1986). A short guide to linkage analysis. In "Human Genetic Diseases—A Practical Approach" (K. E. Davies, Ed.), pp. 19–32, IRL Press, Oxford.
- Pandit, S. D., Wang, J. C., Veile, R. A., Mishra, S. K., Warlick, C. A., and Donis-Keller, H. (1995). Index, comprehensive microsatellite, and unified linkage maps of human chromosome 14 with cytogenetic tie points and a telomere microsatellite marker. *Genomics* **29**: 653–664.
- Raeymakers, R., van Zand, K., Jun, L., Höglund, M., Cassiman, J.-J., van den Berghe, H., *et al.* (1995). A radiation hybrid map with 60 loci covering the entire short arm of chromosome 12. *Genomics* **29**: 170–178.
- Sapru, M., Gu, J., Gu, X., Smith, D., Yu, C. E., Wells, D., *et al.* (1994). A panel of radiation hybrids for human chromosome 8. *Genomics* **21**: 208–216.
- Scott, A. F., Schmeckpeper, B. J., Abdelrazik, M., Comey, C. T., O'Hara, B., Rossitter, J. P., *et al.* (1987). Origin of human L1 elements: Proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* **1**: 113–125.
- Smith, V., Brown, C. M., Bankier, A. T., and Barrell, B. G. (1990). Semi-automated preparation of DNA templates for large-scale DNA sequencing projects. *DNA Seq.* **1**: 73–78.
- Stewart, E. A., and Cox, D. R. (1997). Radiation hybrid mapping. In "Genome Mapping—A Practical Approach" (P. H. Dear, Ed.), pp. 73–93, IRL Press, Oxford.
- Sunden, S. L. F., Businga, T., Beck, J., McClain, A., Gastier, J. M., Pulido, J. C., *et al.* (1996). Chromosomal assignment of 2900 tri- and tetranucleotide repeat markers using NIGMS somatic cell hybrid panel 2. *Genomics* **32**: 15–20.
- Teague, J. W., Collins, A., and Morton, N. E. (1996). Studies on locus content mapping. *Proc. Natl. Acad. Sci. USA* **93**: 11814–11818.
- Walter, G., Tomlinson, I. M., Cook, G. P., Winter, G., Rabbits, T. H., and Dear, P. H. (1993). HAPPY mapping of a YAC reveals alternative haplotypes in the human immunoglobulin Vh locus. *Nucleic Acids Res.* **21**: 4524–4529.
- Wang, D., and Smith, C. L. (1994). Large-scale structure conservation along the entire long arm of human chromosome 21. *Genomics* **20**: 441–451.