

low recombination. Nucleotide and haplotype diversity will also probably parallel recombination rates. Although our baseline long-range recombination rates will be useful, they should be recalculated when the human genomic sequences are finished and as higher resolution genetic maps become available. In the more distant future, genotyping greater numbers of reference families at much higher polymorphism densities will lead to short-range maps of recombination hot spots. □

Methods

Connection of genetic and physical maps

We used short, single-pass genomic sequences and/or PCR primer sequences for STRPs to identify draft or finished bacterial artificial chromosome (BAC) or cosmid sequences within GenBank that encompass the STRPs using BLAST²⁷ and ePCR²⁸. Blast criteria were score (bits) > 200, expect (E) value < e⁻⁵⁰, and ratio of matched bases to marker sequence length > 85%. ePCR criteria were no more than one base mismatch in each primer and size of PCR product within allele size range for the STRP. About 75% of the STRPs were connected to the long genomic sequences. The reasons for failure of the remaining 25% are not fully understood, but include absence of the corresponding sequence in GenBank and poor quality of the STRP sequences. As the genetic maps are marker rich, the absence of 25% was not a serious limitation. Tables of STRPs with GenBank sequence accession numbers for encompassing BACs, genetic map positions and recombination rates are available from the Marshfield web site.

Determination of recombination rates

For each sequence assembly we built new female, male and sex-average genetic maps, using the marker order provided by the assemblies and using the genotyping data from the eight CEPH reference families⁵. We fitted cubic splines to plots of genetic versus physical distance, and from these curves we obtained recombination rates as first derivatives¹⁵. The statistical significance of the recombination rates was estimated by computer simulation of 1,000 iterations of recombination within each interval between markers, assuming a constant level of recombination across the genome for each sex. The constant levels of recombination were taken as the total genetic lengths of all the assemblies analysed divided by the total physical lengths of these assemblies.

Computation of marker and sequence parameters

We calculated STRP heterozygosities using genotypes of individuals within the eight CEPH families. We obtained STRP positions relative to centromeres and telomeres as the fractional sex-average genetic map distances from the centromeres to the telomeres (value of 0 for a STRP at the centromere and 1.0 for a STRP at the telomere)⁵. GC content and STR densities were obtained from programs written and tested at Marshfield²⁹. STR densities were measured as numbers of runs of non-interrupted repeats rather than total numbers of repeats. Minimum values of *n* for (A)_{*n*}, (AC)_{*n*}, (AGAT)_{*n*}, (AAN)_{*n*}, and (AAAN)_{*n*} sequences were 12, 11 or 19 ((AC)_{*n*}), 5, 7 and 5, respectively. We obtained interspersed repetitive element densities using the program Repeat Masker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>). SINES and LINES were defined by Repeat Masker and consist primarily of Alu and L1 elements, respectively. We computed all DNA sequence parameters over 250-kb windows centred about each STRP. For markers ≤ 125 kb from the ends of the sequence assemblies, we defined the window as the 125 kb of proximal sequence plus all available distal sequence. Unknown bases in the sequence assemblies were excluded from analysis. All parameters were corrected for reduced window size owing to unknown bases or proximity to ends.

Measurement of linkage disequilibrium

Recombination deserts and jungles were selected as those chromosomal regions with sex-average recombination rates of <0.3 or >3.0, respectively. We measured linkage disequilibrium for all pairs of STRPs within the deserts (449 pairs) and jungles (467 pairs) using Fisher's exact test³⁰. Only disequilibrium results that were significant at *P* ≤ 0.01 were plotted in Fig. 2. An overall *P*-value was obtained by a permutation test treating the regions as units in order to account for the dependence between marker pairs within a region.

Received 27 October; accepted 8 December 2000.

1. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
2. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
3. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
6. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
7. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
8. Fain, P. R., Kort, E. N., Yousry, C., James, M. R. & Litt, M. A high resolution CEPH crossover mapping panel and integrated map of chromosome 11. *Hum. Mol. Genet.* **5**, 1631–1636 (1996).
9. Bouffard, G. G. *et al.* A physical map of human chromosome 7: An integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* **7**, 673–692 (1997).
10. Nagaraja, R. *et al.* X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res.* **7**, 210–222 (1997).

11. Mohrenweiser, H. W., Tsujimoto, S., Gordon, L. & Olsen, A. S. Regions of sex-specific hypo- and hyper-recombination identified through integration of 180 genetic markers into the metric physical map of human chromosome 19. *Genomics* **47**, 153–162 (1998).
12. Nicolas, A. Relationship between transcription and initiation of meiotic recombination: toward chromatin accessibility. *Proc. Natl Acad. Sci. USA* **95**, 87–89 (1998).
13. Wahls, W. P. Meiotic recombination hotspots: shaping the genome and insights into hypervariable minisatellite DNA change. *Curr. Top. Dev. Biol.* **37**, 37–75, (1998).
14. Faris, J. D., Haen, K. M. & Gill, B. S. Saturation mapping of a gene-rich recombination hot spot region in wheat. *Genetics* **154**, 823–835 (2000).
15. Kliman, R. M. & Hey, J. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol. Biol. Evol.* **10**, 1239–1258 (1993).
16. Chen, T. L. & Manuelidis, L. SINES and LINES cluster in distinct DNA fragments of Giemsa band size. *Chromosoma* **98**, 309–316 (1989).
17. Payseur, B. A. & Nachman, M. W. Microsatellite variation and recombination rate in the human genome. *Genetics* **156**, 1285–1298 (2000).
18. Nachman, M. W., Bauer, V. L., Crowell, S. L. & Aquadro, C. F. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141 (1998).
19. Nachman, M. W. & Crowell, S. L. Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. *Genetics* **155**, 1855–1864 (2000).
20. Majewski, J. & Ott, J. GT repeats are associated with recombination on human chromosome 22. *Genome Res.* **10**, 1108–1114 (2000).
21. Bernardi, G. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3–17 (2000).
22. Eisenbarth, L., Vogel, G., Krone, W., Vogel, W. & Assum, G. An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am. J. Hum. Genet.* **67**, 873–880 (2000).
23. Yu, J. *et al.* Individual variation in recombination among human males. *Am. J. Hum. Genet.* **59**, 1186–1192 (1996).
24. Lien, S., Szyda, J., Schechinger, B., Rappold, G. & Arnheim, N. Evidence for heterogeneity in recombination in the human pseudoautosomal region: High resolution analysis by sperm typing and radiation-hybrid mapping. *Am. J. Hum. Genet.* **66**, 557–566 (2000).
25. Carrington, M. Recombination within the human MHC. *Immunol. Rev.* **167**, 245–256 (1999).
26. Jeffreys, A. J., Ritchie, A. & Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hot spot. *Hum. Mol. Genet.* **9**, 725–733 (2000).
27. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389 (1997).
28. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* **7**, 541–550 (1997).
29. Zhao, C., Heil, J., & Weber, J. L. A genome wide portrait of short tandem repeats. *Am. J. Hum. Genet.* **65**, (Suppl.) A102 (1999).
30. Huttley, G. A., Smith, M. W., Carrington, M. & O'Brien, S. J. A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711–1722 (1999).

Supplementary information is available from *Nature's* World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*, and from the Marshfield Web site (<http://research.marshfieldclinic.org/genetics>).

Acknowledgements

This work was supported by contracts from the US National Institutes of Health and Department of Energy. Assistance was also provided by the chromosome 6 and 20 project groups at the Sanger Centre, supported by the Wellcome Trust.

Correspondence and requests for materials should be addressed to J.L.W. (e-mail: weberj@cmg.mfldclin.edu).

.....
Integration of cytogenetic landmarks into the draft sequence of the human genome

The BAC Resource Consortium*

* *Authorship of this paper should be cited using the names of authors that appear at the end.*

.....
 We have placed 7,600 cytogenetically defined landmarks on the draft sequence of the human genome to help with the characterization of genes altered by gross chromosomal aberrations that cause human disease. The landmarks are large-insert clones mapped to chromosome bands by fluorescence *in situ* hybridization. Each clone contains a sequence tag that is positioned on the genomic sequence. This genome-wide set of sequence-anchored clones allows structural and functional analyses of the genome. This resource represents the first comprehensive integration of cytogenetic, radiation hybrid, linkage and sequence maps of the

human genome; provides an independent validation of the sequence map^{1,2} and framework for contig order and orientation; surveys the genome for large-scale duplications, which are likely to require special attention during sequence assembly; and allows a stringent assessment of sequence differences between the dark and light bands of chromosomes. It also provides insight into large-scale chromatin structure and the evolution of chromosomes and gene families and will accelerate our understanding of the molecular bases of human disease and cancer.

With the draft of the human genome available², scientists can conduct global analyses of its gene content, structure, function and variation. One important challenge is to define the genetic contribution to human diseases. For many developmental disorders, inherited conditions and cancers, gross chromosomal aberrations provide clues to the locations of the causative molecular defects. These aberrations are visible as alterations in chromosomal banding patterns³ or in the number or relative positions of DNA sequences labelled by fluorescence *in situ* hybridization (FISH)⁴. Although tracing gross abnormalities to the level of DNA sequence⁵ has revealed the genetic causes of many diseases, molecular characterization of chromosomal aberrations has lagged far behind their discovery⁶. To proceed from cytogenetic observation to gene discovery and mechanistic explanation, scientists will need access to a resource of experimental reagents that effectively integrates the cytogenetic and sequence maps of the human genome.

We describe here the results of a concerted effort to assemble such a genome-wide resource of well mapped, large-insert DNA clones. Each clone has been localized directly to chromosomal band(s) by FISH (Fig. 1a) and assigned one or more unique sequence tags, which can anchor the clone to the emerging draft sequence. We used complementary strategies to amass the current set of 8,877 clones.

The set, which consists primarily of bacterial artificial chromosome (BAC) clones, includes clones targeted to contain sequence-tagged sites (STSs) ordered along the genome by genetic linkage or radiation hybrid mapping (for well ordered and distributed coverage); clones randomly selected for end sequencing from the RPCI-11 library (for coverage of regions low in STSs); clones identified during intense mapping efforts that preceded sequencing of some chromosomes (for denser coverage); and clones suspected of being partially

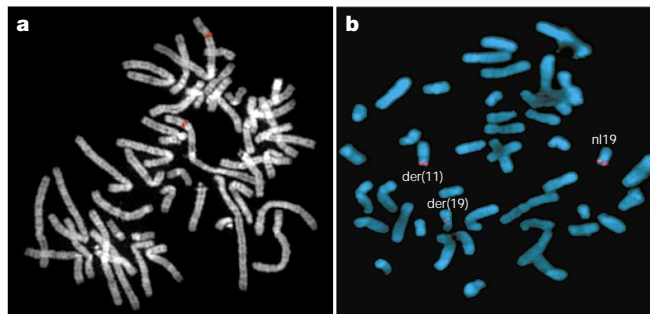


Figure 1 Cytogenetic analyses of sequence-integrated clones. **a**, Using FISH, fluorescent signals are observed at cytogenetic bands (grey) where fragments of a sequence-tagged BAC hybridize (red). **b**, Clones selected on the basis of band location were used in FISH analyses to map the breakpoint of a translocation involving chromosomes 11 and 19 in a patient with multiple congenital malformations and mental retardation (DGAP012, <http://dgap.harvard.edu>). Clone CTD-3193o13 spans the breakpoint on chromosome 19; red signal is split between the derivative chromosome 11 and derivative 19 chromosomes and is also present on the normal chromosome 19. The GTG-banded karyotype for this patient is 46,XY,t(11;19)(p11.2;p13.3).

Table 1 A cytogenetic resource of FISH-mapped, sequence-tagged clones

| Chromosome | Number of FISH-mapped clones | | | | | Connections to the draft sequence | | | | |
|-----------------|------------------------------|-------------|---------|----------|---------------|---|--------------|------|------|--------------|
| | Type of sequence tag | | | Coverage | | No. of clones anchored to draft sequence§ | % discordant | | | % concordant |
| | Accession* | STS or gene | BAC end | Total† | Avg. density‡ | | Chrom. | Site | | |
| 1 | 868 | 318 | 355 | 1,248 | 4.9 | 1,127 | 2 | 2 | 95 | |
| 2 | 43 | 180 | 189 | 297 | 1.2 | 241 | 3 | 2 | 95 | |
| 3 | 128 | 222 | 178 | 308 | 1.5 | 233 | 5 | 6 | 90 | |
| 4 | 42 | 253 | 227 | 341 | 1.7 | 275 | 7 | 1 | 92 | |
| 5 | 35 | 237 | 168 | 296 | 1.6 | 255 | 3 | 4 | 93 | |
| 6 | 653 | 212 | 176 | 909 | 5.1 | 801 | 3 | 1 | 96 | |
| 7 | 25 | 254 | 151 | 324 | 1.9 | 274 | 2 | 1 | 96 | |
| 8 | 31 | 181 | 161 | 245 | 1.6 | 203 | 5 | 3 | 92 | |
| 9 | 208 | 169 | 252 | 384 | 2.7 | 324 | 4 | 2 | 94 | |
| 10 | 191 | 302 | 288 | 454 | 3.2 | 382 | 4 | 4 | 93 | |
| 11 | 119 | 243 | 225 | 378 | 2.7 | 324 | 6 | 2 | 92 | |
| 12 | 109 | 251 | 178 | 304 | 2.2 | 266 | 7 | 2 | 91 | |
| 13 | 182 | 101 | 175 | 278 | 2.4 | 252 | 3 | 1 | 96 | |
| 14 | 48 | 167 | 167 | 222 | 2.1 | 196 | 3 | 2 | 96 | |
| 15 | 109 | 117 | 154 | 224 | 2.2 | 189 | 5 | 2 | 94 | |
| 16 | 72 | 237 | 196 | 267 | 2.8 | 222 | 4 | 2 | 95 | |
| 17 | 21 | 71 | 77 | 119 | 1.3 | 93 | 10 | 1 | 89 | |
| 18 | 9 | 73 | 76 | 105 | 1.3 | 86 | 2 | 0 | 98 | |
| 19 | 7 | 55 | 49 | 76 | 1.1 | 56 | 14 | 0 | 86 | |
| 20 | 228 | 107 | 112 | 388 | 5.6 | 333 | 1 | 1 | 98 | |
| 21 | 4 | 64 | 52 | 85 | 1.8 | 72 | 1 | 0 | 99 | |
| 22 | 217 | 123 | 108 | 343 | 6.5 | 303 | 3 | 1 | 96 | |
| X | 641 | 274 | 150 | 872 | 5.5 | 782 | 2 | 2 | 96 | |
| Y | 7 | 13 | 15 | 17 | 0.3 | 14 | 7 | 0 | 93 | |
| Subtotal | 3,997 | 4,224 | 3,879 | 8,484 | 2.6 | 7,303 | 3.6 | 1.9 | 95 | |
| Multiple sites¶ | 209 | 100 | 220 | 393 | n.a. | 297 | n.a. | n.a. | n.a. | |
| Total | 4,206 | 4,324 | 4,099 | 8,877 | n.a. | 7,600 | | | | |

All clones are associated with a sequence-tag; localized directly to cytogenetic bands by FISH; BACs, P1, or PACs; archived as single-colony-purified stocks; and publicly available.

n.d., not done. n.a., not applicable.

* Clones whose draft or finished sequence is deposited in GenBank.

† Total is less than sum of preceding three columns because some clones have >1 type of sequence tag.

‡ In clones per Mb, that is, number of FISH-mapped clones/chromosome size in Mb²⁰.

§ Sequence tags of 8,325 single-site and 352 multisite clones were used to search the 7 October 2000 draft. Clones whose sole tag consisted of a Unigene accession (Hs.) and some multisite clones have not yet been evaluated.

|| Discordant site refers to clones mapped by FISH to location >1 band away from, but on same chromosome as, neighbours on draft.

¶ Because most clones in the resource were not selected randomly, fraction of multi-site clones does not accurately reflect frequency of low-copy duplications in the genome.

duplicated at more than one location in the genome (to flag regions of the genome that might complicate sequence assembly⁷). The molecular signatures are STSs (many corresponding to genes or expressed sequence tags (ESTs)), BAC end sequences, or the actual draft or final sequence of the clone (Table 1). Earlier publications have described genome-wide and chromosome-specific subsets of this collection^{8–12}.

Each clone is publicly available as single-colony-purified bacterial stocks and is ready for distribution. Each clone can each be obtained from one of three stock centres by e-mail: mapped-clones@mail.cho.org, libraries@resgen.com and clonerequest@sanger.ac.uk. The website <http://www.ncbi.nlm.nih.gov/genome/cyto> provides information about all clones in this collection, including how to obtain each clone. (Additional information can be obtained at the websites listed in Supplementary Information 1).

The 8,877 clones provide excellent coverage of the human genome (Table 1), with at least one clone on average per megabase (Mb) for 23 of the 24 chromosomes. Clone density ranges from greater than ~5 clones per Mb for chromosomes 1, 6, 20, 22 and X to about 0.3 clones per Mb for chromosome Y.

Our study provides an assessment of the representation of the human genome in the RPCI-11 BAC library¹³, which serves as the intermediate template for most sequencing efforts² and the foundation of genome-wide contig assembly by fingerprint analyses¹. We

randomly selected 1,243 clones from this library for FISH analysis. The number of clones assigned to each chromosome correlated well with chromosome size, with no significant bias in the distribution of clones between Giemsa (G)-dark and G-light bands of chromosomes (see Supplementary Information 2 and 3).

Cytogenetic mapping is one of several methods that can produce a framework of ordered clones upon which the human sequence can be assembled. The resource provides an opportunity to cross-check these critical framework maps, because over 3,300 FISH-mapped clones have STSs that reference the radiation hybrid¹⁴ or linkage maps^{15,16}. Overall, the concordance between cytogenetic map order and marker order established by radiation hybrid and linkage mapping is very high for clones with single cytogenetic locations (94–98%, depending on the map; Table 2). Significant discrepancies were observed for only around 140 of these clones and are probably due to errors in clone tracking. Integration of cytogenetic and linkage maps also aids efforts to map disease genes. The location of the cytogenetic abnormality in one patient can guide the choice of polymorphic markers to assess linkage in other families that have similar phenotypes, but no visible chromosomal aberrations.

At present, 7,303 clones that map to single cytogenetic locations are positioned by their sequence tags on the draft sequence assembly of 7 October 2000 (Table 1). The fraction of clones located on the draft sequence ranges from 76% to 91% across different

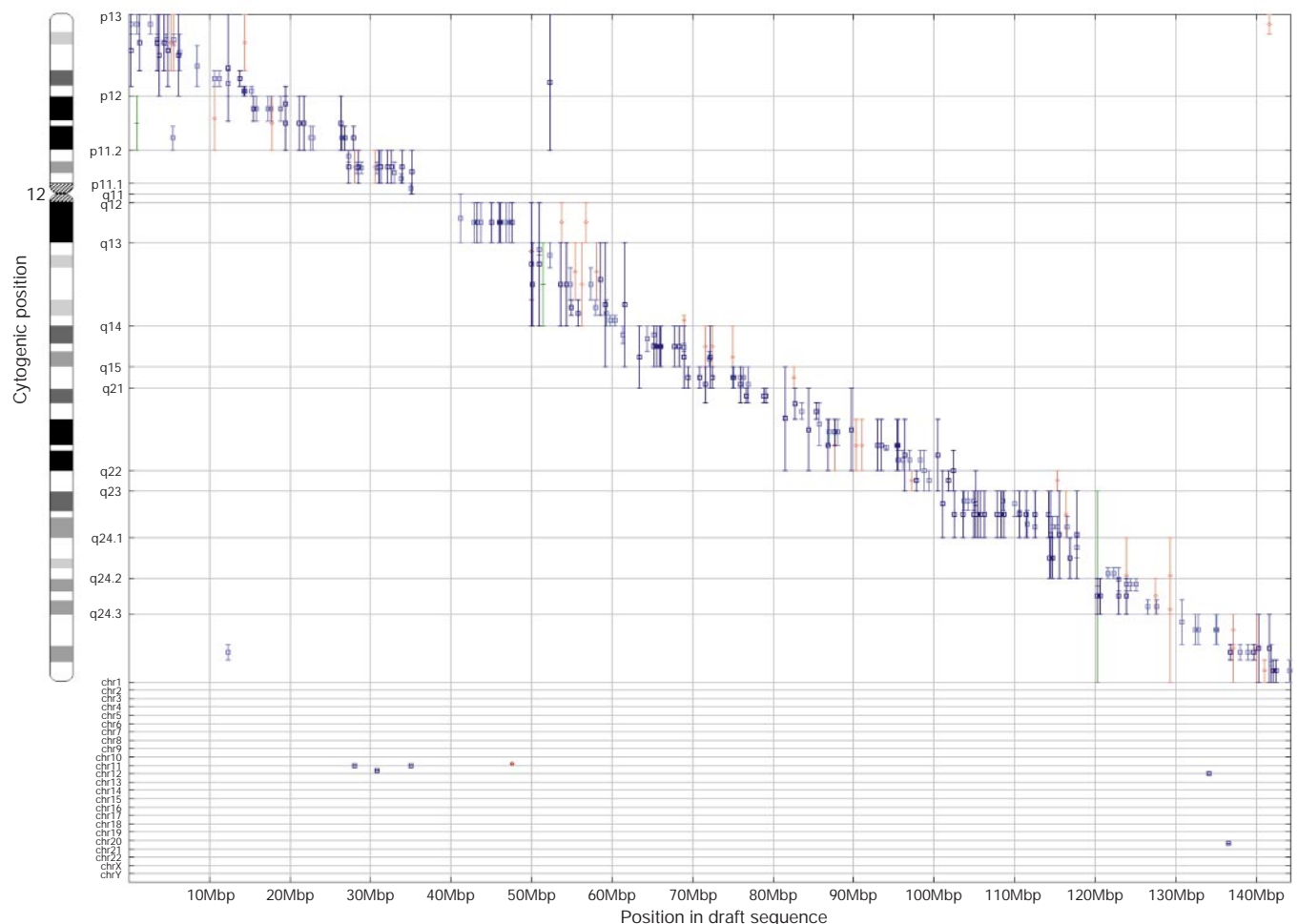


Figure 2 The correspondence between cytogenetic location and position on the 7 October 2000 draft sequence for chromosome 12. The band location of each clone is indicated by a range on the y-axis. Clones mapping to chromosomes other than 12 are indicated at the bottom. Colours differentiate assignments made in different laboratories. Each clone is anchored on the draft sequence by one or more sequence tags. Plots for the other chromosomes and the 5 September, 2000 assembly can be found at <http://>

genome.ucsc.edu/goldenPath/mapPlots/. Genome browsers that assist researchers in navigating from cytogenetic location to other maps and detailed, annotated sequence information are available at http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch (NCBI Mapviewer, which includes chromosomal aberrations associated with cancer and inherited disorders), <http://www.ensembl.org/> and <http://genome.ucsc.edu>.

Table 2 Clones connecting the cytogenetic map and other maps of the human genome

| Map type | Version | Number | % concordant | % discordant | |
|------------------|--------------|--------|--------------|--------------|------|
| | | | | Chrom. | Site |
| Genetic | Genethon | 1,686 | 98 | 1.4 | 0.7 |
| | Marshfield | 1,757 | 98 | 1.4 | 0.9 |
| Radiation hybrid | GM99-GB4 | 1,433 | 98 | 1.3 | 1.2 |
| | GM99-G3 | 1,654 | 96 | 2.5 | 1.4 |
| | TNG | 908 | 94 | 2.9 | 2.5 |
| Draft sequence | 5 Sept. 2000 | 7,364 | 94 | 3.7 | 2.2 |
| | 7 Oct. 2000 | 7,303 | 95 | 3.6 | 1.9 |

Many clones have markers positioned on more than one map. Only clones assigned to single chromosome locations by FISH are considered above. An additional 91 clones that map by FISH to more than one location contain STSs placed on other maps. STSs are by definition unique, single-copy markers, so each is assigned to a single genomic location by other mapping approaches. In 88% of these 91 cases, the STS location corresponds to one of the FISH-detected locations.

chromosomes (see Supplementary Information 4). We expect these percentages to rise as more sequence is merged into the draft and algorithms for locating tags are refined.

The connections between the cytogenetic map and the draft sequence are well distributed across the genome, and the correspondence in position on the two maps is excellent for these 7,303 clones (Fig. 2 shows chromosome 12 as an example). Of the 943 contigs of overlapping clones in the 7 October 2000 draft sequence, 660 are connected to the cytogenetic map by at least one clone, and 531 by two or more clones. Thus, many contigs can be oriented on the chromosome on the basis of FISH results of constituent clones. Relatively few discrepancies between cytogenetic location and position in the draft sequence are apparent at this level of resolution (~5% of the clones map either to other chromosomes or more than one band away from the expected position; Table 1). We found only eight locations where the cytogenetic data indicated that portions of the sequence were misplaced within an earlier draft assembly (5 September 2000). The sequencing centres used these cytogenetic findings to locate errors in the assembly and produce the later draft of improved quality (Table 2).

FISH analyses of this clone collection reveal abundant paralogous relationships among sites dispersed across the human genome. Of 1,243 clones randomly selected from the RPCI-11 library, 5.4% hybridize to more than one chromosomal location (see Supplementary Information 3). The entire collection includes 393 clones that together identify over 150 bands containing at least one segment with significant homology to one or more (up to 25) other sites in the genome (see Supplementary Information 5). These data provide clues to duplications and exchanges that have occurred within and between chromosomes. Among the 393 clones, 111 contain blocks duplicated within the same chromosome; 282 hybridize to more than one chromosome. Paralogous relationships involving pericentromeric and subtelomeric regions of multiple chromosomes are particularly frequent and complex. Clones in the collection also identify low-copy duplications specific to chromosomes 1, 7, 11 and 16, the pseudoautosomal regions of X and Y, and sites of the olfactory receptor gene family¹⁷. Many previously undescribed patterns were also observed; some were confirmed with two or more clones, but others require further study to verify that they reflect true duplications.

Many of these duplications are functionally significant, as some have generated multigene families, and some are potential sites of recombination events, which can result in chromosome abnormalities. The cytogenetic data should greatly facilitate analyses of these regions, which are likely to pose challenges to sequence assembly. The sequence tags of 84% of the clones that hybridize to more than one site were placed in the 7 October 2000 draft assembly, and the location(s) were roughly consistent with at least one FISH observation for 88% of these clones. Collectively, the multisite clones highlight regions that are more likely to become entangled with other regions of the genome during sequence assembly than clones with single FISH locations. Indeed, global BLAST analyses show that

regions encompassing sequence tags of multi-site clones (either the sequence of the FISH-mapped clone or a surrogate clone from the assembly) contain blocks of homology found at an average of around 3.9 chromosomal locations (compared to around 1.3 for the regions underlying clones with single FISH signals). The regions observed by FISH and revealed through homology searches are not fully congruent, however (not shown). These findings indicate that both FISH and sequence analyses may underestimate large-scale duplications and that these complex, inter-related regions of the genome will require special attention during the finishing stages of genome sequencing.

The extensive integration of cytogenetic and primary sequence data gives investigators access to fine-structure information—including details on predicted genes—for cytogenetic locations of interest. Tools such as NCBI's MapViewer and the UCSC and ENSEMBL genome browsers (see Fig. 2 for URLs) allow researchers to navigate readily between chromosomal location and annotated sequence.

This integration provides insight into the sequence differences underlying cytogenetic banding patterns. Sequence analyses of 200-kilobase (kb) regions surrounding the sequence tags of 338 clones mapped with the finest band resolution reveal more striking differences in the base-pair composition between Giemsa-positive and -negative bands than were predicted from earlier studies¹⁸. These clones were mapped with high precision to 850-level bands of varying staining intensity¹⁹ on seven chromosomes. The AT content of 58 of the 59 clones in the darkest G-bands exceeds the genome-wide average of 0.59 (mean 0.63), whereas the AT content of only 22 of the 143 clones in G-negative bands is higher than average (mean 0.55; $\chi^2 = 43$, $P < 0.005$). These data confirm that dark G-bands are more AT-rich than G-negative bands.

The utility of a sequence-integrated cytogenetic resource is illustrated by two examples. In the first, clones are applied in conventional FISH assays to rapidly narrow the search for candidate genes disrupted or deregulated by translocations causing developmental disorders. The process is expedited by selection of clones assigned to the regions implicated by banding analyses. In a patient with multiple congenital malformations and mental retardation (DGAP012, <http://dgap.harvard.edu>), a breakpoint-spanning clone was identified (Fig. 1b). This clone spans a 170-kb interval containing the gene for MKK7, a human mitogen-activated protein kinase, and a novel sequence with homology to the *trc-2* oncogene, both plausible candidate genes. More typically, breakpoints will be mapped to an interval between neighbouring clones. For example, a translocation implicated in mental retardation in another patient maps to an interval containing at least 12 genes, including protocadherin 8, a promising candidate given its exclusive expression in fetal and adult brain²⁰.

In the second example, an array of around 2,000 BAC clones from the collection is used to perform a genome-wide scan for segmental aneuploidy by comparative genomic hybridization (CGH) (Fig. 3 and A. Snijders *et al.*, in preparation). The array format offers better sensitivity and resolution^{21,22} than metaphase chromosomes, the traditional target for CGH²³, and, because the arrayed clones are integrated into the draft, copy-number abnormalities can be related directly to sequence information. To illustrate the power of array CGH, the ML-2 cancer cell line was 'karyotyped' using the array. Array CGH revealed relative copy-number losses on 1p, 6q, 11q and 20p and gains of 12, 13 and 20q (Fig. 3). Copy-number abnormalities on chromosomes 6, 11 and 20 were subsequently confirmed by FISH using clones predicted by array CGH to be included in the region of loss. Several of these alterations were noted in previous banding analyses (1p-, 6q-, 11q-, +12, +13q+)²⁴, but array CGH locates the breakpoints precisely relative to BACs that reference specific locations in the sequence.

More than 7,500 clones now link the cytogenetic map and sequence of the human genome. Application of these reagents in

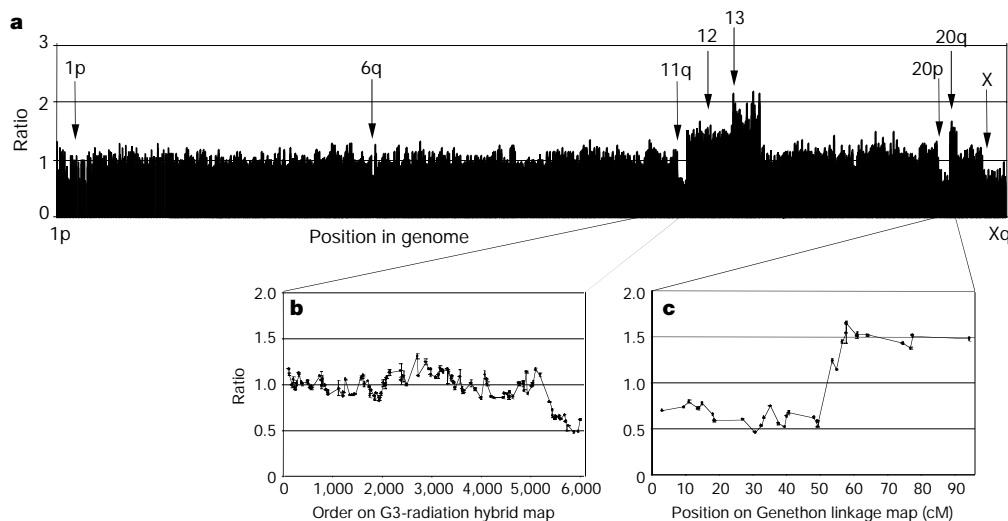


Figure 3 Copy-number analysis of myeloblastic leukaemia ML-2 cell line using CGH and a genome-wide array of around 2,000 BAC clones. The ML-2 cell line has acquired chromosomal abnormalities in addition to those present in the original tumour during long-term culture. CGH maps regions of abnormal copy number by comparing the relative efficiency with which test (Cy3-labelled ML-2 DNA) and reference (Cy5-labelled normal female DNA) hybridize to clones on the array. The array excludes clones that hybridize to multiple sites in the genome. **a**, Fluorescence ratios of Cy3 to Cy5 fluorescence for each

BAC normalized to the median ratio for all 2,000 clones on the array, ordered from 1pter to Xqter. Arrows, chromosomal regions showing significant copy number variations. The lower ratio on the X indicates expected ratio for mismatched sex of test and reference DNAs. Fluorescence ratios of clones on chromosomes 11 (**b**) and 20 (**c**) are shown with clones ordered according to position of their STSs on the G3 radiation hybrid or Genethon linkage maps, respectively.

combination with increasingly detailed knowledge of genes and other functional motifs in the human sequence will transform the process of identifying genes that are altered in cancer and other diseases. Ultimately, this resource will contribute to a better understanding of the organization of the cell nucleus, the compacting of DNA into mitotic chromosomes, and the basis of the chromosomal banding patterns that have been so valuable in uncovering the aetiology of human diseases. □

Methods

GenBank was screened for draft, finished or end sequences derived from clones in this collection. BACs were screened for STS content by a combination of hybridization and polymerase chain reaction (see refs 8, 25 and Supplementary Information for details). Sequence tags were located on the draft sequence by a combination of methods (see Supplementary Information and refs 26, 27). Sequence at these locations was compiled with the results of a genome-wide BLAST analysis (ref. 2 and J. A. Bailey and E. E. Eichler, in preparation) to identify paralogous regions of the genome (regions in the draft sequence containing ~ 20 kb of sequence that match sequence of the FISH-mapped clone or that of a surrogate clone from the assembly at ~ 90% identity in non-repeat-masked bases over each 1-kb segment), and these locations were translated into estimated band positions using a dynamic programming algorithm (T. S. Furey *et al.*, in preparation; and see Supplementary Information).

Details of FISH procedures are provided elsewhere^{4,28}. Only locations of unique or low-copy portions of the clone are identified, because high-copy interspersed repetitive sequences were suppressed by addition of unlabelled *Cot1* DNA. Replicate analyses indicate that the precision of FISH assignments to metaphase bands is roughly 5–10 Mb (1–1.5 band). A subset of 442 clones was ordered at very high (~2–3-Mb) resolution¹¹. FISH analyses were performed using DNA from the bacterial stock used for STS typing. Data that failed to replicate (for example, replicate FISH analyses of the same clone or different clones assigned the same marker) have been removed. Hybridization to arrays was carried out as described previously²⁹ and by Snijders *et al.* (in preparation).

Received 7 November 1999; accepted 20 December 2000.

1. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Caspersson, T. *et al.* Chemical differentiation along metaphase chromosomes. *Exp. Cell Res.* **49**, 219–222 (1968).
4. Trask, B. J. in *Genome Analysis: A Laboratory Manual* Vol. 4, 303–413 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1999).
5. Collins, F. S. Positional cloning moves from perdictional to traditional. *Nature Genet.* **9**, 347–350 (1995).
6. Mitelman, F. *Catalog of Chromosome Aberrations in Cancer* (Wiley, New York, 1998).
7. Eichler, E. E. Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res.* **8**, 758–762 (1998).

8. Cheung, V. G. *et al.* A resource of mapped human bacterial artificial chromosome clones. *Genome Res.* **9**, 989–993 (1999).
9. Korenberg, J. R. *et al.* Human genome anatomy: BACs integrating the genetic and cytogenetic maps for bridging genome and biomedicine. *Genome Res.* **9**, 994–1001 (1999).
10. Leversha, M. A., Dunham, I. & Carter, N. P. A molecular cytogenetic clone resource for chromosome 22. *Chromosome Res.* **7**, 571–573 (1999).
11. Kirsch, I. R. *et al.* A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nature Genet.* **24**, 339–340 (2000).
12. Kirsch, I. R. & Ried, T. Integration of cytogenetic data with genome maps and available probes: Present status and future promise. *Semin. Hematol.* **37**, 420–428 (2000).
13. Osoegawa, K. *et al.* Bacterial artificial chromosome library for sequencing the human genome. *Genome Res.* (in the press).
14. Olivier, M. *et al.* A high resolution radiation hybrid map of the human genome draft sequence. *Science* (in the press).
15. Yu, A. *et al.* Comparison of human genetic and sequence-hosed physical maps. *Nature* **409**, 951–953 (2001).
16. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
17. Trask, B. J. *et al.* Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum. Mol. Genet.* **7**, 2007–2020 (1998).
18. Saccone, S. *et al.* Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci. USA* **90**, 11929–11933 (1993).
19. Mitelman, F. (Ed.) *ISCN (1995): An International System for Human Cytogenetics Nomenclature* (S. Karger, Basel, 1995).
20. Strehl, S. *et al.* Characterization of two novel protocadherins (PCDH8 and PCDH9) localized on human chromosome 13 and mouse chromosome 14. *Genomics* **53**, 81–89 (1998).
21. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20**, 207–211 (1998).
22. Solinas-Toldo, S. *et al.* Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosom. Cancer* **20**, 399–407 (1997).
23. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821 (1992).
24. Ohyashiki, K., Ohyashiki, J. H. & Sandberg, A. V. Cytogenetic characterization of putative human myeloblastic leukemia cell lines (ML-1, -2, and -3): origin of the cells. *Cancer Res.* **46**, 3642–3647 (1986).
25. Morley, M. GenMapDB: A database of mapped human BAC clones. *Nucleic Acids Res.* **29**, 144–147 (2001).
26. Schuler, G. D. Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol.* **16**, 456–459 (1998).
27. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).
28. Korenberg, J. R., Chen, X. -N. Human cDNA mapping using a high-resolution R-banding technique and fluorescence in situ hybridization. *Cytogenet. Cell Genet.* **69**, 196–200 (1995).
29. Albertson, D. G. *et al.* Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genet.* **25**, 144–146 (2000).
30. Trask, B. J., van den Engh, G., Mayall, B. & Gray, J. W. Chromosome heteromorphism quantified by high resolution bivariate flow karyotyping. *Am. J. Hum. Genet.* **45**, 738–752 (1989).

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

Acknowledgements

We thank M. Arcaro, M. Bakis, J. Burdick, J. Chang, H.-C. Chen, S. Chiu, Y. Fan, C. Harris, L. Haley, R. Hosseini, J. Kent, M. A. Leversha, J. Martin, L.-T. Nguyen, P. Quinn, Y. H. Ramsey, T. Reppert, L. J. Rogers, J. Shreve, J. Stalica, M. Wang, T. Weber, A. M. Yavor, J. Young, K. Zatloukal, and members of the TIGR BAC Ends Team for assistance. This work was supported by grants from NIH (NCI, NHGRI, NIDCD and NICHD), US DOE, NSF, HHMI, PPG, Merck Genome Research Institute, Vysis, Inc., and start-up funds provided by Obstetrics and Gynecology at Brigham and Women's Hospital.

Correspondence should be addressed to B.J.T.
(e-mail: btrask@fhcrc.org).

The BAC Resource Consortium

V. G. Cheung^{1*}, N. Nowak^{2*}, W. Jang³, I. R. Kirsch⁴, S. Zhao⁵, X.-N. Chen⁶, T. S. Furey⁷, U.-J. Kim^{8,†}, W.-L. Kuo⁹, M. Olivier¹⁰, J. Conroy², A. Kasprzyk¹¹, H. Massa¹², R. Yonescu⁴, S. Sait², C. Thoreen^{13,†}, A. Snijders⁹, E. Lemyre¹⁴, J. A. Bailey¹⁵, A. Bruzel¹, W. D. Burrill¹¹, S. M. Clegg¹¹, S. Collins¹³, P. Dhani¹¹, C. Friedman¹², C. S. Han¹⁶, S. Herrick¹⁴, J. Lee⁸, A. H. Ligon¹⁴, S. Lowry¹⁷, M. Morley¹, S. Narasimhan¹, K. Osoegawa^{2,18}, Z. Peng¹⁷, I. Plajzer-Frick¹⁷, B. J. Quade¹⁴, D. Scott¹⁷, K. Sirotkin³, A. A. Thorpe¹¹, J. W. Gray⁹, J. Hudson¹⁹, D. Pinkel⁹, T. Ried⁴, L. Rowen²⁰, G. L. Shen-Ong^{4,†}, R. L. Strausberg⁴, E. Birney¹¹, D. F. Callen²¹, J.-F. Cheng¹⁷, D. R. Cox¹⁰, N. A. Doggett¹⁶, N. P. Carter¹¹, E. E. Eichler¹⁵, D. Haussler²², J. R. Korenberg⁶, C. C. Morton¹⁴, D. Albertson⁹, G. Schuler³, P. J. de Jong^{2,18} & B. J. Trask¹²

* These authors contributed equally to this work.

¹, Department of Pediatrics, University of Pennsylvania, The Children's Hospital of Philadelphia, 3516 Civic Center Boulevard, ARC 516, Philadelphia, Pennsylvania 19104, USA; ², Roswell Park Cancer Institute, Elm and Carleton Street, Buffalo, New York 14263, USA; ³, National Center for Biotechnology Information, National Library of Medicine, Building 38A/Room 8N805, Bethesda, Maryland 20894, USA; ⁴, National Cancer Institute, NIH, Building 10/Room 12N214, Bethesda, Maryland 20889-5105, USA; ⁵, The Institute for

Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; ⁶, Departments of Pediatrics and Human Genetics, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Los Angeles, California 90048, USA; ⁷, Computer Science Department, University of California Santa Cruz, 1156 High Street, Santa Cruz, California 95064-1077, USA; ⁸, Department of Biology, California Institute of Technology, Mail Code 147-75, Pasadena, California 91125, USA; ⁹, University of California San Francisco Cancer Center, Box 0808, San Francisco, California 94143-0808, USA; ¹⁰, Stanford University, Genome Lab, Mail Code 5120, Stanford, California 94305-5120, USA; ¹¹, Sanger Center, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; ¹², Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North C3-168, P.O. Box 19024, Seattle, Washington 98109-1024, USA; ¹³, Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, Washington 98195-7730, USA; ¹⁴, Departments of Obstetrics and Gynecology and Pathology, Brigham and Women's Hospital, Amory Lab Building 3rd floor, Boston, Massachusetts 02115, USA; ¹⁵, Department of Human Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA; ¹⁶, Joint Genome Institute-Los Alamos National Laboratory, MS M888 B-N1, P.O. Box 1663, Los Alamos, New Mexico 87545, USA; ¹⁷, Joint Genome Institute-Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 84-171, Berkeley, California 94720, USA; ¹⁸, Children's Hospital Oakland Research Institute, 747 52nd Street, Oakland, California 94609, USA; ¹⁹, Research Genetics, 2130 Memorial Parkway, Huntsville, Alabama 35801, USA; ²⁰, Institute for Systems Biology, 4225 Roosevelt Way NE, Suite 200, Seattle, Washington 98105-6099, USA; ²¹, Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, 72 King William Road, North Adelaide, South Australia 5006, Australia; ²², Howard Hughes Medical Institute, Computer Science Department, University of California Santa Cruz, 1156 High Street, Santa Cruz, California 95064-1077, USA
† Present addresses: PanGenomics, 6401 Foothill Boulevard, Tujunga, California 91024, USA (U.-J.K.); Harvard Medical School, 240 Longwood Avenue, Cell Biology, Cambridge, Massachusetts 02115, USA (C.T.); Gene Logic, Inc., 708 Quince Orchard Road, Gaithersburg, Maryland 20878, USA (G.L.S.-O.).