

A YAC-based physical map of the mouse genome

Chad Nusbaum^{1*}, Donna K. Slonim^{1*}, Katrina L. Harris¹, Bruce W. Birren¹, Robert G. Steen¹, Lincoln D. Stein¹, Joyce Miller¹, William F. Dietrich¹, Robert Nahf¹, Victoria Wang¹, Olga Merport¹, Andrew B. Castle¹, Zeeshan Husain¹, Gail Farino¹, Delphine Gray¹, Mechele O. Anderson¹, Richard Devine¹, Lloyd T. Horton Jr¹, Wenjuan Ye¹, Xiaoyun Wu¹, Vardouhie Kouyoumjian¹, Irina S. Zemsteva¹, Yi Wu¹, Alville J. Collymore¹, Dorothy F. Courtney¹, James Tam¹, Matthew Cadman², Andrew R. Haynes², Christine Heuston², Tracy Marsland², Anne Southwell², Pamela Trickett², Mark A. Strivens², Mark T. Ross³, Wojciech Makalowski⁴, Yixun Xu^{4,5}, Mark S. Boguski⁴, Nigel P. Carter³, Paul Denny², Steve D.M. Brown², Thomas J. Hudson^{1,6} & Eric S. Lander^{1,7}

*These authors contributed equally to this work.

A physical map of the mouse genome is an essential tool for both positional cloning and genomic sequencing in this key model system for biomedical research. Indeed, the construction of a mouse physical map with markers spaced at an average interval of 300 kb is one of the stated goals of the Human Genome Project¹. Here we report the results of a project at the Whitehead Institute/MIT Center for Genome Research to construct such a physical map of the mouse. We built the map by screening sequenced-tagged sites (STSs) against a large-insert yeast artificial chromosome (YAC) library and then integrating the STS-content information with a dense genetic map. The integrated map shows the location of 9,787 loci, providing landmarks with an average spacing of approximately 300 kb and affording YAC coverage of approximately 92% of the mouse genome. We also report the results of a project at the MRC UK Mouse Genome Centre targeted at chromosome X. The project produced a YAC-based map containing 619 loci (with 121 loci in

common with the Whitehead map and 498 additional loci), providing especially dense coverage of this sex chromosome. The YAC-based physical map directly facilitates positional cloning of mouse mutations by providing ready access to most of the genome. More generally, use of this map in addition to a newly constructed radiation hybrid (RH) map² provides a comprehensive framework for mouse genomic studies.

We used the same basic strategy and methods to construct the YAC-based map of the mouse genome as in our previous work assembling a physical map of the human genome³. The STSs used for screening came from several sources: simple-sequence length polymorphisms (SSLPs) taken from a dense genetic map that we had previously constructed^{4,5}; STSs developed from cDNA sequences from mouse genes; and STSs developed from random mouse genomic sequence. The genetic markers are especially valuable because they provide 'top-down' information about location.

Table 1 • Distribution of loci and contigs on genome-wide map

Chromosome	Physical length (Mb) ¹⁵	Total loci	Average spacing (kb)	Loci on genetic map	Loci on STS-content map	Doubly linked contigs	Singly linked contigs
1	216	802	269	511	635	97	71
2	208.5	771	270	507	600	101	70
3	179.7	549	327	343	398	58	49
4	176.7	536	330	350	400	71	59
5	170.4	583	292	402	422	24	54
6	165.9	602	276	368	471	69	48
7	155.7	545	286	357	416	72	53
8	149.1	493	302	350	376	61	45
9	143.7	526	273	336	398	64	44
10	144.9	428	339	293	320	61	48
11	141.6	593	239	350	518	69	40
12	146.4	432	339	278	367	60	38
13	131.4	483	272	303	425	58	36
14	133.8	400	335	259	323	50	41
15	121.5	405	300	264	332	47	38
16	114	371	307	215	318	49	33
17	115.5	365	316	255	198	37	31
18	116.4	356	327	231	270	42	30
19	81.9	207	396	134	158	29	23
X	186.9	340	550	230	241	51	43
Total	3,000	9,787	317	6,336	7,586	1,170	894

¹Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. ²MRC UK Mouse Genome Centre and Mammalian Genetics Unit, Harwell, OX11 0RD, UK. ³Sanger Centre Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ⁴National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. ⁵Department of Biochemistry and Molecular Biology, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ⁶Montreal General Hospital Research Institute, McGill University, Montreal H3G 1A4, Canada. ⁷Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. Correspondence should be addressed to E.S.L. (e-mail: lander@genome.wi.mit.edu) or T.J.H. (e-mail: thudson@genome.wi.mit.edu).

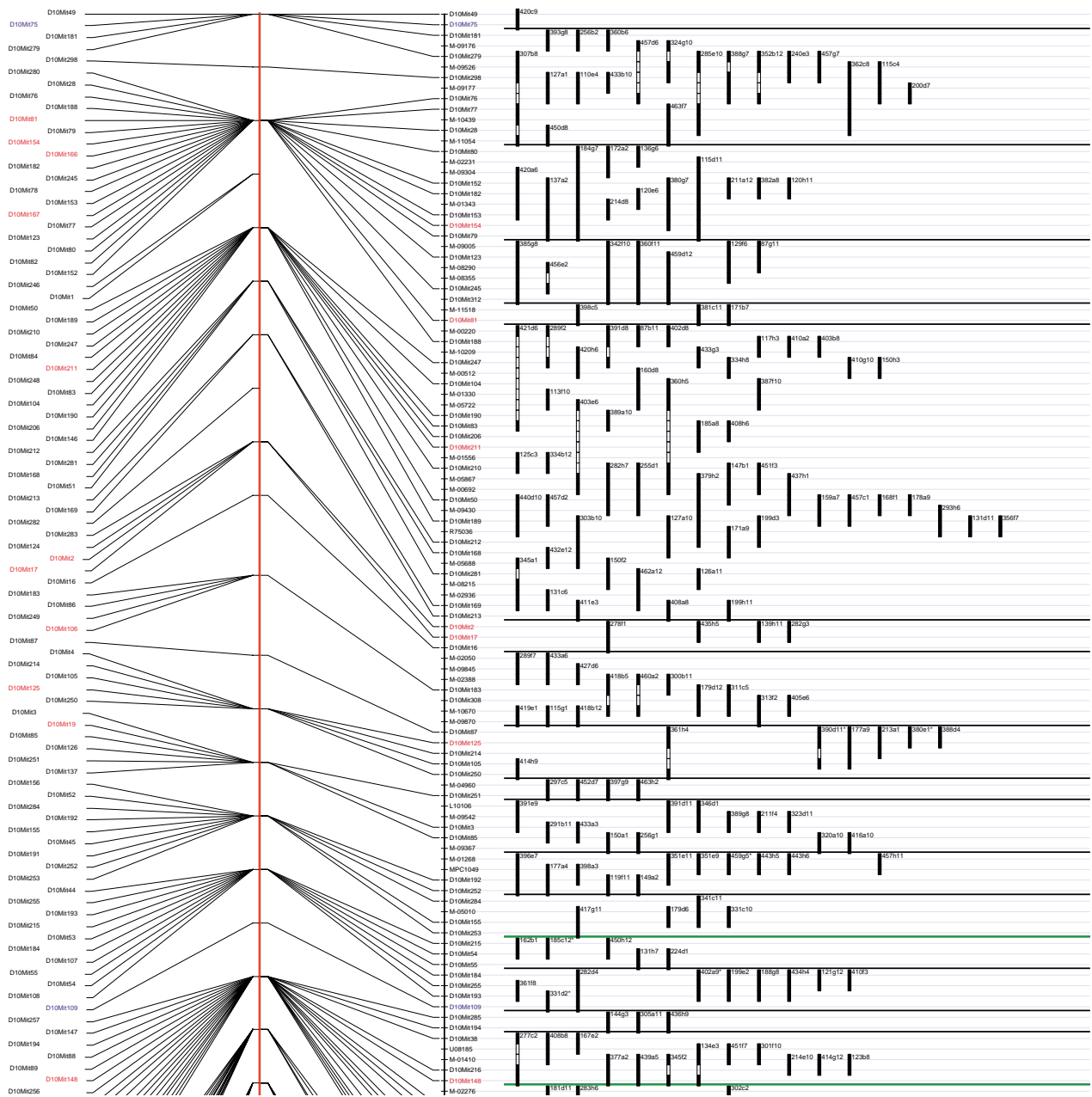


Fig. 1 Integrated genetic and physical map of mouse chromosome 10 (continued on the next two pages). The two long vertical axes represent the two maps, with the genetic map⁴ on the left and the physical map on the right. The associated columns of marker names represent the order of the markers on the two maps. On the genetic map, the relative spacing reflects the genetic distances between the markers based on the Whitehead/MIT cross⁴. Markers on the physical map are displayed as equally spaced because intermarker distances are not known. Lines connecting the two map axes connect markers that appear on both maps. Names in red denote dominant markers whose genetic map positions are known with slightly lower resolution. Names in blue denote STSs doubly linked to more than one chromosome but assigned to chromosome 10 by the genetic map. Bars displayed to the right of the maps represent YAC clones. Filled bars represent a positive YAC hit with the designated STS. Open bars represent negative assays for markers thought to be contained within a YAC. YAC names are displayed at the top right of each bar. YACs detected by only a single marker are not displayed. Gaps between YAC contigs are displayed as horizontal lines; green horizontal lines indicate that the contig above has been correctly oriented by genetic data from the EUCIB project⁵. Searchable versions of these maps are available (<http://www-genome.wi.mit.edu/cgi-bin/mouse/index>).

The STSs were screened against a subset of the MIT YAC library⁶ consisting of 21,120 clones. These clones have an average insert size of 820 kb; thus the subset provides an estimated 5.8-fold coverage of the genome. We identified on average 5.9 YAC addresses per STS, consistent with the expected coverage. The screening of the YAC library involved PCR of combinatorial pools using a novel five-dimensional pooling scheme designed to be robust against false positives and false negatives. The project involved more than 17 million PCR reactions performed using a

large-scale automation system referred to as the Genomatron³. Positive YAC addresses were identified for 9,864 STSs.

We then undertook the construction of a physical map. Considerable care was required because YAC libraries include a substantial number of chimaeric clones, which contain inserts from more than one genomic region; it is estimated that 35% of the MIT mouse YAC library clones may be chimaeric⁶. Accordingly, two STSs are not considered nearby in the genome on the basis of linkage by a single YAC clone. Rather, double linkage (by two

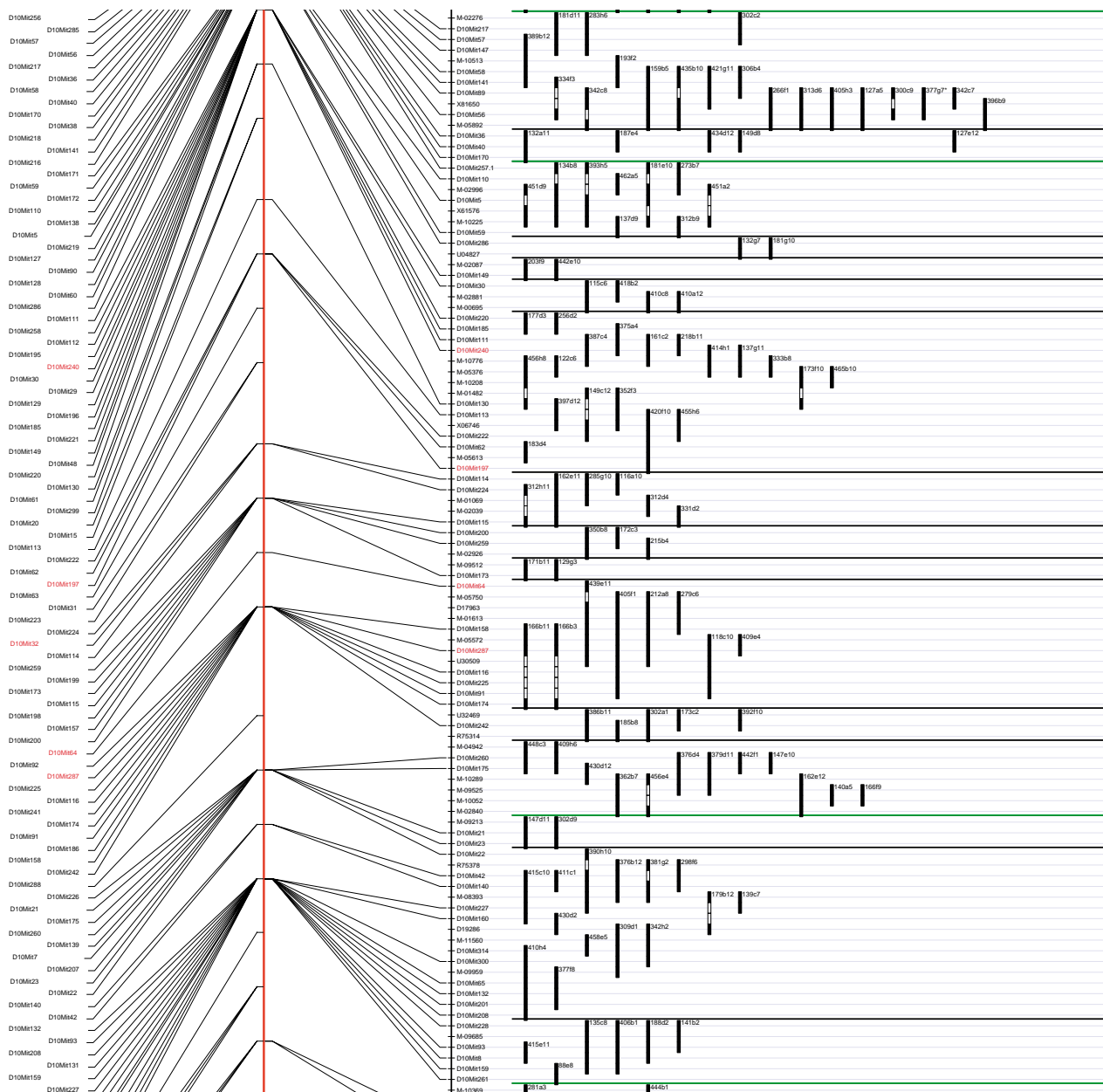


Fig. 1 Integrated genetic and physical map of mouse chromosome 10 (continued from previous page).

independent YAC clones) is required.

We first assigned 8,203 STSs to individual chromosomes based on double linkage to genetic markers. We then constructed a map of each chromosome by integrating the YAC-linkage information with the known genetic map positions of the genetic markers using the *co2* software package³. Doubly linked contigs were identified and then single-linkage information was used to join doubly linked contigs known to lie nearby. The maps were closely inspected to identify apparent conflicts in order between the genetic and physical maps. We found 57 instances in which a genetic marker was triply linked to loci in a different 'bin' on the genetic map. The linkage was to an adjacent bin in 53% of cases and to markers at most two bins away in 68% of the cases. Such local conflicts are likely due to either chromosome-specific repeats or to the statistical uncertainty inherent in the two mapping methods. Only nine instances were found of strong linkage to markers separated by more than five bins on

the genetic map. These cases are likely to be due to laboratory errors such as mislabelled primer tubes. Finally, we oriented the contigs with regard to the genetic map using data from both the MIT and EUCIB crosses^{4,5}. The ordering of markers is consistent with the genetic map, but fine-structure order may be imperfect owing to imperfections in YAC-STs data.

The map contains 9,787 loci distributed across the 19 autosomes and the X chromosome. We have integrated the final STS-content map containing 7,586 loci with a 6,432-locus genetic map by virtue of 4,229 loci in common. The 4,229 genetically mapped markers provide a secure framework for the STS-content map. The remaining 3,355 loci on the STS-content map consist primarily of random genomic STSs together with 700 STS derived from mouse genes. The STS-content map contains 1,170 doubly linked anchored contigs, with neighbours joined by single links into 894 singly linked anchored contigs. Not yet included in the STS-content map are 180 doubly linked contigs

last 500 random STSs having positive YAC addresses. (The YAC library itself is known to contain most of the genome on the basis of test screening⁶.) Approximately 76% of these STSs fell into contigs, with 87% being anchored and the rest unanchored. Most of the remaining 24% of the STSs were singly but not doubly linked to other loci in the map. Of these, 68% were reliably incorporated into the YAC map on the basis of single linkage confirming that the link was to a marker in a nearby position on the RH map² (thereby eliminating the risk of linkage through a chimaeric YAC). In all, 92% of the STSs were reliably incorporated into YAC contigs on the basis of double linkage or single linkage together with RH position.

The overall distribution of STSs is fairly uniform across the genome, with the exception of chromosome X, which is sparser owing to the known under-representation of SSLP genetic markers on this chromosome⁴. Given the importance of chromosome X for genomic studies, a separate project was undertaken at the MRC Mouse Genome Centre to construct a denser map.

We sequenced plasmids containing DNA from flow sorting of chromosome X. We obtained 1,043 novel sequences and derived successful STS assays for 702 of these loci. Genetic mapping of a subset of these STSs demonstrated that more than 94% were in fact derived from chromosome X. We screened the 702 STSs, together with 230 SSLP markers, against the MIT YAC library and recovered 4,367 coordinates corresponding to 791 markers. The resulting data were assembled into an STS-content map on the basis of double and single linkage, together with information about the genetic order of the SSLP markers. The problem of linkage through chimaeric YACs is minimized because most markers are known to lie on a single chromosome (whereas most chimaeric linkages will, simply by chance, be inter-chromosomal).

The resulting map contains 619 markers, of which 168 are SSLP genetic markers (<http://www.mgc.har.mrc.ac.uk/xmap/xmap.html>). The loci are organized into 60 doubly linked and 79 singly linked contigs. The MRC map substantially extends the Whitehead map of chromosome X: it contains 498 new loci including 44 SSLP genetic markers, and shares 121 SSLPs in common. Together, the two maps provide 834 mapped STSs on chromosome X, corresponding to an average marker spacing of 224 kb. An integrated map image that joins the two physical maps by the 121 shared genetic markers is available on the Whitehead and the MRC UK web sites.

Construction of a dense YAC-STS map of the mouse genome fulfils one of the stated goals of the Human Genome Project. Furthermore, the STS-content maps are currently being integrated with newly-developed RH maps² to provide a comprehensive genomic resource parallel to that developed for the human genome^{3,7,8}. The availability of such large collections of mapped STSs, together with associated YAC clones, is essential to a wide variety of endeavours. The YAC clones facilitate rapid positional cloning of mouse mutations based on genetic location. The two maps together permit efficient localization of genes^{7,8} for both candidate gene and comparative mapping purposes. Finally, the STSs provide a scaffold for sequencing of the genome: they can be used either to identify specific clones to be sequenced or to anchor genomic sequence obtained from random clones or genome-wide shotgun sequencing. Thus, this map helps set the stage for the completion of the mouse genome sequence and the beginning of a new era in comparative genomics.

Methods

STS generation (Whitehead). STSs were taken from three sources. First, simple-sequence length polymorphism markers from the MIT genetic map

were used, and, in cases in which the existing PCR assay performed poorly in screening of the YAC library, new assays were designed to amplify a target sequence flanking but not including the simple sequence repeat. STSs defined by such assays are denoted by the original locus name followed by a period and a number (for example, D10Mit33.2), indicating that the STS was derived from sequence near the SSLP but the specific primers are different. Second, gene sequences were taken from GenBank. Third, we generated random genomic sequences from 10,000 clones from a small-insert pUC library prepared from sheared, size-selected genomic DNA from a C57BL/6 J female and sequenced them with dye-primer chemistry on an ABI377 sequencer. Sequences were analysed to remove cloning vector and known repetitive sequences using the Whitehead/MIT STS pipeline software³. PCR primers were selected using PRIMER (M.J. Daly, S. Lincoln and E.S.L.) with a calculated T_m of 58 °C. All STSs were characterized before screening to confirm that they amplify mouse DNA under a uniform set of PCR conditions.

Pooling scheme (Whitehead). The YAC library was divided into 22 blocks of ten 96-well plates (960 clones) each and pooled using a five-dimensional pooling scheme into 55 pools per block. The pooling scheme was based on parallel planes in the three-dimensional affine geometry over the finite field GF(11). Specifically, the 960 clones were mapped into the $11^3=1,331$ points of GF(11)³ and the pools corresponded to 5 sets of 11 parallel planes. The planes were chosen to be orthogonal to the vectors (1,0,0) (0,1,0) (0,0,1) (0,1,2) and (1,2,1). These directions have the property that any three are linearly independent. Accordingly, in blocks containing a single positive YAC, any three of the five dimensions are thus sufficient to recover the address. We also chose the combinations so that a single false-positive hit in any dimension is usually consistent with only a single legitimate address and so that two YAC hits within a block can typically be resolved without ambiguity. Total yeast DNA was prepared for each of the 1,210 sub-pools. Including various controls, each STS was tested in 1,536 PCR reactions.

PCR screening of the YAC library (Whitehead). High-throughput PCR screening was performed as described³.

Chromosomal assignment of STSs (Whitehead). STSs having more than 14 hits in the YAC library were eliminated from analysis because such outliers may represent STSs detecting low-copy repeat sequences. Genetic markers were assigned to the chromosomes to which they had been previously mapped. A small number of genetic markers showing double linkage to more than one chromosome were eliminated from future rounds of the chromosomal assignment process, because the data were likely to be due to repeats or laboratory errors. Of the remaining markers, we assigned those loci showing double linkage to markers on a single chromosome to that chromosome and we eliminated any loci doubly linked to more than one chromosome. This step was repeated 3 times and resulted in chromosomal assignment of 8,203 of the STSs.

Map construction (Whitehead). The *co2* software package (developed at Whitehead and publicly available at <http://www-genome.wi.mit.edu/ftp/distribution/software/co2>) was designed to integrate map information from multiple sources. It searches marker orders to maximize a scoring function. The scoring function awards a high score for YACs hitting a pair of adjacent markers, assesses large penalties for violating the genetic map order and awards smaller penalties for introducing gaps or breaks in clones. The costs are optimized to approximate the log likelihood of the given order, so that the chosen marker order is consistent with as much of the data as possible.

STS generation (MRC). Mouse chromosomes were prepared for flow sorting essentially as described⁹, except that dextran sulphate (20 µg/ml) was added to the culture medium to enhance the stimulation of the splenocytes. Aliquots of 250,000 X chromosomes were sorted using an Epics Elite ESP flow cytometer (Beckman-Coulter) in sheath buffer (100 mM sodium chloride, 10 mM Tris base, 1 mM EDTA, pH 7.2) into tubes coated with yeast tRNA (12.5 µg per tube). To each tube (containing ~450 µl of chromosome suspension) we added 0.25 M EDTA/10% sodium lauryl sarcosine (50 µl) and Proteinase K (5 µl, 20 mg/ml), then gently mixed the contents and incubated overnight at 42 °C. Proteinase K was inactivated for 40

min at RT by the addition of phenyl methyl sulphonyl fluoride (PMSF) to 40 µg/ml. The DNA was purified and, after digestion with *Hind*III, used to construct a small insert library in pBluescript II–SK+ (Stratagene) using standard methods. Plasmid DNA was prepared on a Biomek 2000 (Beckman-Coulter) using an alkaline lysis method (<http://www.genome.ou.edu/proto.html>) and clones containing inserts >100 bp were selected for sequencing. Sequencing reactions were performed using BigDye Terminators (PE Applied Biosystems) and run on a PE377 Prism sequencer. Sequences were downloaded to the HOSEpipe management system¹⁰ for STS design. STSs derived from flow-sorted purified chromosome X material were named according to microtitre plate number and well address, prefixed by 'px-'. Primer sequences and other STS data are publicly available (<http://www.mgc.har.mrc.ac.uk/xmap/xmap.html>).

YAC library DNA pooling and PCR screening (MRC). The YAC library⁶ was replicated in 96-well plates and the plates divided into 46 stacks, each comprising 8 plates. Yeast cultures were combined according to a three-dimensional pooling scheme to provide primary pools (representing each stack) and secondary pools (representing individual plates, row pools and column pools from each stack). Yeast DNA was prepared by lithium dodecyl sulphate lysis of cells embedded in agarose¹¹. We screened YAC library DNA pools by PCR with STSs using AmpliTaq Gold DNA polymerase (Perkin-Elmer). PCR templates and reactions were prepared on a Biomek 2000 and run on PTC-225 thermal cyclers (MJ Research). We eliminated STSs giving >20 primary pool positives (out of a possible 46) because they were likely to represent repetitive elements. Secondary level PCR screens yielded specific plate, row and column coordinates. The library screening process provides unambiguous coordinate data (where secondary screening of an individual stack for a single STS provides a single library address) as well as ambiguous (or 'inferred') results where STS screening produces multiple potential library addresses from a single stack. All PCR reactions generated from library screening, including negative and C57BL/6 genomic controls, were electrophoresed on 52-well, 8 comb, Nusieve agarose (Flowgen) gels.

Genetic mapping (MRC). A panel of DNAs from 44 progeny from the EUCIB (refs 5,12) cross, selected for breakpoints along the length of the mouse X chromosome, was used for the genetic mapping of random flow sorted X-chromosome STSs. PCR products were analysed on 8% non-

denaturing polyacrylamide gels and single-strand conformational polymorphisms (SSCP) detected using a silver-staining method¹³. STSs that showed significant linkage (lod>3) to anchor markers on chromosome X in the EUCIB database (MBx, <http://www.hgmp.mrc.ac.uk/MBx/MBx-Homepage.html>) were assigned 'DXMgc' numbers. Loci were ordered by minimizing observed recombinants among the available haplotypes.

Map construction (MRC). The X chromosome map was assembled by reading files produced by HOSEpipe into the SAM software¹⁴. Subsequently, the STS content map was built in three stages. Only unambiguous YAC coordinate data were used in the first two stages. Initially, we incorporated markers that were genetically mapped to the X chromosome into the developing STS content map along with STSs that were doubly linked by unambiguous YAC coordinates to these genetic anchors. Order as defined by the EUCIB genetic map was primary. Markers that were on either WI-CGR or MGD (<http://www.informatics.jax.org/bin/ccr/index>) genetic maps, but not EUCIB, were positioned by comparison with flanking markers shared with the EUCIB map. We subsequently incorporated markers that were singly linked to other markers already on the map. In the final stage, markers that detected inferred coordinates were then introduced into the map where they appeared to improve contiguity. These inferred coordinates were tested by PCR amplification from YAC clones from the appropriate wells. If the coordinates were confirmed, the marker was included in the final map.

Acknowledgements

We thank D. Henriques, S. Jackson, Y.Y. Lau, S. Greenaway, P. Middlehurst, P. Weston, P. Avner, I. Poras, C. Mundy, B. Gorick, H. Blair, Y. Boyd, J. Crabtree, B. Roe, L. Rogers and J. King for their help in the X chromosome project. T.J.H. is a recipient of a Clinician-Scientist award from the Medical Research Council of Canada. This work was supported in part by grants from the National Institute for Human Genome Research (to E.S.L.), the Medical Research Council, UK, and the Wellcome Trust.

Received 16 April; accepted 28 June 1999.

- Collins, F. & Galas D. A new five-year plan for the U.S. Human Genome Project. *Science* **262**, 43–46 (1993).
- Van Etten, W.J. et al. A radiation hybrid map of the mouse genome. *Nature Genet.* **22**, 384–387 (1999).
- Hudson, T.J. et al. An STS-based map of the human genome. *Science* **270**, 1945–1954 (1995).
- Dietrich, W.F. et al. A comprehensive genetic map of the mouse genome. *Nature* **380**, 149–152 (1996).
- Rhodes, M. et al. A high-resolution microsatellite map of the mouse genome. *Genome Res.* **8**, 531–542 (1998).
- Haldi, M.L. et al. A comprehensive large-insert yeast artificial chromosome library for physical mapping of the mouse genome. *Mamm. Genome* **7**, 767–769 (1996).
- Schuler, G.D. et al. A gene map of the human genome. *Science* **274**, 540–546 (1996).
- Deloukas, P. et al. A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
- Rabbitts, P. et al. Chromosome specific paints from a high resolution flow karyotype of the mouse. *Nature Genet.* **9**, 369–375 (1995).
- Strivens, M.A., Middlehurst, P., Brown, S.D. & Denny, P. HOSEpipe—a WWW-hosted data management and analysis system for STS content mapping projects. *Mamm. Genome* **8**, 467–471 (1997).
- Anand, R. et al. A yeast artificial chromosome contig encompassing the cystic fibrosis locus. *Genomics* **9**, 124–130 (1991).
- Breen, M. et al. Towards high resolution maps of the mouse and human genomes—a facility for ordering markers to 0.1 cM resolution. European Backcross Collaborative Group. *Hum. Mol. Genet.* **3**, 621–627 (1994).
- Denny, P. & Brown, S. Genome mapping. in *Mouse Genetics & Transgenics: A Practical Approach* (eds Jackson, I. & Abbott, C.) (Oxford University Press, Oxford, 1999).
- Soderlund, C. & Dunham, I. SAM: a system for iteratively building marker maps. *Comput. Appl. Biosci.* **11**, 645–655 (1995).
- Evans, E.P. in *Genetic Variants and Strains of the Laboratory Mouse* (eds Lyon, M.F., Rastan, S. & Brown, S.D.M.) 1446 (Oxford University Press, New York, 1996).