A physical map of the mouse genome

Simon G. Gregory*, Mandeep Sekhon†, Jacqueline Schein‡, Shaying Zhao§, Kazutoyo Osoegawa‖, Carol E. Scott*, Richard S. Evans*, Paul W. Burridge*, Tony V. Cox*, Christopher A. Fox*, Richard D. Hutton*, Ian R. Mullenger*, Kimbly J. Phillips*, James Smith*, Jim Stalker*, Glen J. Threadgold*, Ewan Birney¶, Kristine Wylie†, Asif Chinwalla†, John Wallis†, LaDeana Hillier†, Jason Carter†, Tony Gaige†, Sara Jaeger†, Colin Kremitzki†, Dan Layman†, Jason Maas†, Rebecca McGrane†, Kelly Mead†, Rebecca Walker†, Steven Jones‡, Michael Smith‡, Jennifer Asano‡, Ian Bosdet‡, Susanna Chan‡, Suganthi Chittaranjan‡, Readman Chiu‡, Chris Fjell‡, Dan Fuhrmann#, Noreen Girn‡, Catharine Gray‡, Ran Guin‡, Letticia Hsiao‡, Martin Krzywinski‡, Reta Kutsche‡, Soo Sen Lee‡, Carrie Mathewson‡, Candice McLeavy‡, Steve Messervier‡, Steven Ness‡, Pawan Pandoh‡, Anna-Liisa Prabhu‡, Parvaneh Saeedi‡, Duane Smailus‡, Lorraine Spence‡, Jeff Stott‡, Sheryl Taylor‡, Wesley Terpstra‡, Miranda Tsai‡, Jill Vardy‡, Natasja Wye‡, George Yang‡, Sofiya Shatsman§, Bola Ayodeji§, Keita Geer§, Getahun Tsegaye§, Alla Shvartsbeyn§, Elizabeth Gebregeorgis§, Margaret Krol§, Daniel Russell§, Larry Overton§, Joel A. Malek§, Mike Holmes§, Michael Heaney§, Jyoti Shetty§, Tamara Feldblyum§, William C. Nierman§, Joseph J. Catanese‖, Tim Hubbard*, Robert H. Waterston†, Jane Rogers*, Pieter J. de Jong‖, Claire M. Fraser§, Marco Marra‡, John D. McPherson† & David R. Bentley*

** The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK*
† Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA
‡ Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 4E6, Canada
§ The Institute for Genomic Research, Rockville, Maryland 20850, USA
‖ Children's Hospital Oakland Research Institute, Oakland, California 94609, USA
¶ EMBL—European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK
 # Department of Electrical Engineering, Washington University, St Louis, Missouri 63130, USA

A physical map of a genome is an essential guide for navigation, allowing the location of any gene or other landmark in the chromosomal DNA. We have constructed a physical map of the mouse genome that contains 296 contigs of overlapping bacterial clones and 16,992 unique markers. The mouse contigs were aligned to the human genome sequence on the basis of 51,486 homology matches, thus enabling use of the conserved synteny (correspondence between chromosome blocks) of the two genomes to accelerate construction of the mouse map. The map provides a framework for assembly of whole-genome shotgun sequence data, and a tile path of clones for generation of the reference sequence. Definition of the human–mouse alignment at this level of resolution enables identification of a mouse clone that corresponds to almost any position in the human genome. The human sequence may be used to facilitate construction of other mammalian genome maps using the same strategy.

The recent revolution in large-scale analysis of genomes has already yielded near-complete DNA sequences of a diverse range of organisms including bacteria, a yeast, a worm, a fly and humans[1–7]. The sequence of the mouse genome will have a huge impact on biological research and human health. It will provide critical information and reagents for use in mouse experimental models. It will become possible to unravel the mechanisms of complex mammalian biological processes and human disease. The mouse genome sequence will provide the first opportunity to compare the complete sequence and organization of the human genome with that of another mammal. It will aid discovery and annotation of gene structures and other functionally important sequences in both genomes.

Assembly of each large genome sequence to date has been underpinned by production of a comprehensive map of overlapping large-insert bacterial clones (for example, cosmids[8] or bacterial artificial chromosome (BAC) clones[9])[10–13] for sequencing, and, in some cases, for integration with whole-genome shotgun sequence data (refs 5, 7; see also review in ref. 14). The clones provide substrates for finishing each section of the genome sequence separately, making it easier to eliminate errors and to achieve a final accuracy of more than 99.99% (refs 14–16). The study of other large genomes requires physical maps of a similar standard for sequencing and biological studies.

Comparisons between genomes reveal homologous sequences, for example within individual genes, that reflect their common evolutionary origin and subsequent conservation. Similarity between genomes is also evident at the level of long-range sequence organization. A 'conserved segment' (also called a 'conserved linkage') refers to a region where the order of multiple genes on a single chromosome segment (and hence the linkage between them) is the same in both species. A 'conserved synteny' refers to a region where the chromosomal location of multiple genes is conserved, but not necessarily their precise order[17–19]. In general, the degree of similarity at all levels is higher between species that diverged more recently from a common ancestor. The long-range organization of the mouse and human genomes is very similar, and may comprise about 180 conserved syntenic regions[6,17]. Detailed studies in some smaller regions suggest that precise gene order has been conserved, confirming the existence of conserved segments[20,21]. The similarity in sequence organization between these two genomes suggests that the human genome can be used as a framework to map the mouse genome[20,21], and possibly other mammalian genomes as well. Alignment of the different genome maps over their entire length would simultaneously define the syntenic relationship between them at a new level of resolution, and accelerate the process of constructing the mouse clone map.

Using the human sequence as a framework

We constructed a physical clone map of the mouse genome in two phases. In the first phase, we compared restriction digest patterns ('fingerprints') of 305,716 BAC clones[22]. We identified overlaps between clones on the basis of similarity between fingerprints and used this information to construct 7,587 contigs of overlapping clones. In parallel, we determined sequences at the ends of the mouse DNA fragments cloned in the BACs (that is, BAC end sequences, or BESs). We aligned the mouse BAC contigs to the human genome sequence by identifying matches between human sequence and mouse BESs. We extended and joined contigs where possible after re-examining the fingerprint data. Overlapping fin-

gerprints were required to support each join, while juxtaposition of the mouse clone contigs along the human genome greatly accelerated the manual editing process. This phase resulted in generation of a human–mouse homology clone map. In the second phase, we used a set of independently mapped mouse markers (available in existing genetic and radiation hybrid maps of the mouse) to position the BAC contigs in the mouse genome. After further manual contig editing was carried out, we generated a mouse clone map comprising 296 contigs.

An illustration of the human–mouse homology clone map produced in the first phase is shown in Fig. 1. Within a 1.6-megabase (Mb) region of human chromosome 6 (6q16.1), 11 of the 15 segments of human sequence match to 29 of the BESs within a mouse BAC contig. This contig is located on mouse chromosome 4 (Mmu4). The syntenic relationship between the two genomes in this area was previously indicated by the proximity of the two loci *CNR1* and *GABRR1* (human) and their corresponding homologues *Cnr1* and *Gabrr1* (mouse) (also shown in Fig. 1). In our analysis here, a further 29 homologous crosslinks between these two loci are added.

We found 51,486 homologous crosslinks between the two genomes during this analysis, which corresponds to one match per 54 kb on average (assuming a 2.8-Gb mouse genome; see Methods). These matches were derived from a complete set of 453,962 BESs. Some 73% of the BESs contained more than 100 base pairs (bp) of contiguous unique sequence; we excluded the rest from the analysis. In all, 1,326 of the BESs matched human chromosome 20 sequence. We found that 19% of them matched putative coding regions (163 matches to known genes, 80 to new genes, 8 to putative genes and 1 to a pseudogene). The rest were matches to other conserved sequences in introns (32%) and intergenic regions (49%). Human–mouse sequence homologies in non-coding regions have also been observed in other studies[20,23,24]. From this analysis it is clear that non-coding homologies contributed significantly to the alignment.

Of the clones in the human genome tile path, 88% are collinear with the mouse BAC map (Table 1). For individual human chromosomes, coverage by aligned mouse contigs exceeds 80% on all except chromosome 19 (61%) and the Y chromosome (0%). Lower coverage of these human chromosomes can be accounted for by several factors. First, we used only BES matches with unique locations in the human tile path for the alignment. BESs matching regions of human sequence containing dispersed low-copy repeats were filtered out of the analysis. Such regions are known on human chromosome 19 (notably including zinc-finger-containing genes among others[21,25]). Second, a significant fraction of chromosome 19 clones (49 in all) are included in the total in Table 1 but contain pericentromeric repetitive sequence, and any BES matches to them were excluded. A further 118 clones on chromosome 19 were covered by contigs aligned in our study, but we did not score them in Table 1 because they contained no high-scoring matches to BESs. This is reflected in the low overall density of BES matches on

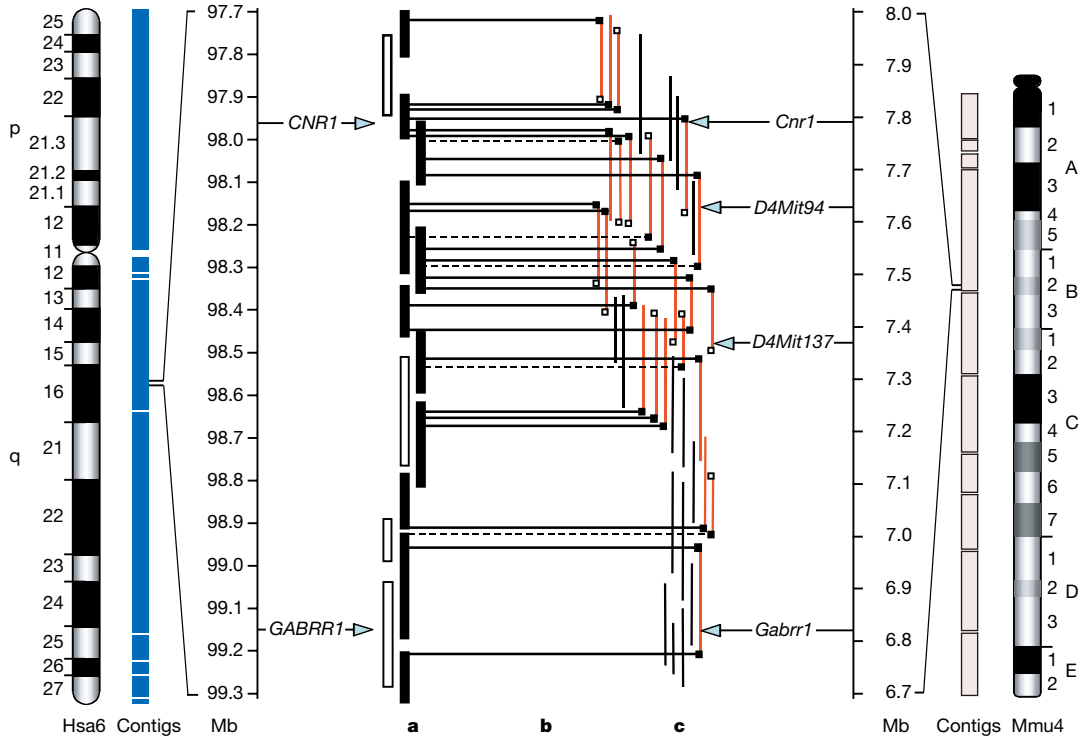


Figure 1 Construction of human–mouse homology clone map. Alignment between part of human chromosome 6 (Hsa6) and mouse chromosome 4 (Mmu4). Clone contigs are shown as blue and pale pink boxes, respectively. A 1.6-Mb interval is enlarged, showing part of Hsa6q16.1 aligned to a 1.3-Mb mouse BAC contig. The accompanying megabase scale for Hsa6 starts from one end of the chromosome (ptel, 0 Mb; qtel,172 Mb), whereas the megabase scale shown for the mouse contig starts at one end of that contig. **a**, Human accessions. Filled and clear boxes denote accessions with and without matches to BESs, respectively. **b**, Matches between human sequence and BESs. Black lines denote high-homology BLAST matches and dashed lines indicate medium-homology BLAST matches (see Methods for details). **c**, Mouse BAC clone contig. All clones having at least one BES that matches a human accession are shown in red (black and white squares at the ends of clones denote BESs with and without matches to human accessions, respectively). Black lines represent a subset of the other clones in the contig (the rest have been omitted for clarity, but can be viewed in Supplementary Information). Previous alignment of the two human loci *CNR1* and *GABRR1* with their mouse homologues can be viewed in the NCBI versus MGD alignment (http://www.ncbi.nlm.nih.gov/Homology). Another human locus (*RNGTT*) lies between *CNR1* and *GABRR1* in the human sequence, and its murine homologue (*Rngtt*) is present in the NCBI versus MGD alignment; however, *Rngtt* has not been placed in the mouse clone map yet. *D4Mit94* and *D4Mit137* are additional mouse markers that have been placed in the clone map by this study.

Table 1 **Summary statistics of human–mouse homology clone map**

Human chromosome number	Human sequence available (Mb)	Human–mouse sequence matches		Number of clones in human tile path	Clones in human tile path that are covered in mouse	
		Total	Per Mb		Number	Per cent
1	239	4,888	20.5	2,179	2,060	95
2	242	4,783	19.8	1,951	1,675	86
3	204	3,667	18.0	1,700	1,493	88
4	189	2,747	14.5	1,612	1,510	94
5	182	2,892	15.9	1,973	1,682	85
6	176	3,264	18.5	1,800	1,666	93
7	161	2,957	18.4	1,518	1,342	88
8	143	2,310	16.2	1,210	1,040	86
9	114	2,505	22.0	963	883	92
10	140	2,531	18.1	1,132	1,014	90
11	139	2,606	18.7	1,134	1,013	89
12	137	2,206	16.1	1,022	943	92
13	99	1,318	13.3	851	798	94
14	90	2,103	23.4	655	642	98
15	82	1,721	21.0	671	573	85
16	81	1,470	18.1	726	666	92
17	81	1,585	19.6	656	612	93
18	79	1,226	15.5	616	564	92
19	46	507	11.0	841	512	61
20	61	1,326	21.8	632	608	96
21	34	507	14.9	475	432	91
22	34	439	12.7	345*	279	81
X	149	1,925	12.9	1,555	1,276	82
Y	26	0	0.0	208	0	0
All	2,928	51,483	17.5	26,425	23,283	88

*For chromosome 22, the finished sequence was divided into abutting 100-kb segments for this analysis.

this chromosome (Table 1) and may be the result of chromosome-specific low-copy repeats, as observed previously[21,25]. We were not able to align any mouse contigs to the human Y chromosome. Two reasons may account for this. First, much of the human Y sequence is repetitive (with Y-specific, X–Y and autosome–Y homologies), and mouse BES matches to these regions would have been excluded. Second, only 39% of the BESs were derived from a male mouse library, and we estimate that only about 0.4% of the BESs were therefore derived from the mouse Y chromosome.

Of the total coverage of the mouse BAC map (in 211 contigs), 97% (2,658 Mb) is aligned to the human genome sequence. The other 3% (84 Mb, in 85 contigs) may not have been aligned at this stage for several reasons. First, the BESs may not find matches in the human sequence because the human sequence is incomplete. In a recent analysis, 3% of a curated set of 10,212 full-length human complementary DNAs (taken from the set in the 'reference sequence project', http://www.ncbi.nlm.nih.gov) were found to have no match in the human sequence (using BLAT[26] to identify matches with more than 95% identity between cDNA and the human genome sequence (NCBI Build 28); J.Kent and D. Haussler, personal communication). This fraction is sufficient to account for all of the unplaced mouse contigs. Second, most of the unplaced contigs are

Table 2 **Summary statistics of mouse physical clone map by chromosome**

Mouse chromosome number	Size estimate (Mb)		Total contigs			
	Old*	New	Number	Coverage (Mb)	Percentage of old size estimate	Percentage of new size estimate
1	216	196	10	211	98	108
2	209	190	13	186	89	98
3	180	163	8	173	96	106
4	177	160	10	145	82	90
5	170	154	24	156	92	101
6	166	151	16	157	95	104
7	156	141	14	143	92	101
8	149	135	12	140	94	104
9	144	131	15	127	88	97
10	145	131	9	135	93	103
11	142	129	6	107	75	83
12	146	132	8	126	86	95
13	131	119	11	132	101	111
14	134	121	8	130	97	107
15	122	111	7	108	89	98
16	114	103	7	101	88	97
17	116	105	6	95	82	91
18	116	105	6	96	83	91
19	82	74	4	65	80	88
X	187	170	30	132	71	78
Y	86	78	4	14	17	18
Subtotal			228	2,679	87	96
Unlocalized			68	60		
Total	3,088	2,800	296	2,739	89	98

*Taken from ref. 47.

short and contain few BESs. The production of additional mouse sequence in these contigs may allow identification of homology crosslinks, and hence alignment of the contigs to the corresponding region in the human genome. Third, there may be insufficient homology between the two genomes in some regions.

The mouse clone map

The mouse genome comprises 19 autosomes plus the two sex chromosomes, X and Y. In contrast to the human genome, all the mouse chromosomes are acrocentric (that is, the centromere is at or near one end). To construct the clone map for each mouse chromosome, we made use of the available genetic map (http://www-genome.wi.mit.edu/cgi-bin/mouse/index) of 6,559 gene-specific or simple sequence length polymorphism (SSLP) markers[27–29], and also a radiation hybrid (RH) map[30] of 13,558 markers, which included a subset of 2,446 of the markers previously ordered in the genetic map. This provided a total unique set of 18,379 markers. We placed 48% (8,789) of these markers within the mouse BAC contigs using the BESs or matches to available mouse genomic sequence, or by hybridization of markers to arrayed BAC clones[31–33]. We then added another 8,203 markers to the map by hybridization to the BACs. As a result, we were able to assign a unique position and orientation on a mouse chromosome to 203 contigs (containing 2,631 Mb, or 96% of the map coverage), and a chromosome assignment (but not a position) to 25 more contigs (containing a total of 47 Mb). A further 68 contigs (containing 60 Mb) have no chromosomal assignment. A summary of the mouse map by chromosome is shown in Table 2.

Most mouse BAC contigs contained multiple mouse markers (average 57 markers per contig). We compared the order of markers within the contigs to the order suggested previously by the genetic and RH data. Overall, there was excellent agreement between the order of markers in each map, as illustrated for example on mouse chromosome Mmu2 (see Fig. 2; and Supplementary Information for all chromosomes). There was also concordance of the marker order in the clone map with that of the European Union RH map (http://www.genoscope.cns.fr/externe/English/Projets/Projet_ZZZ/rhmap.html). The availability of integrated map information permitted assessment of the remaining 27 minor conflicts. We found two conflicts between the clone map and both the genetic and RH maps. The other 25 were conflicts solely with the RH map, and occur in regions where the genetic map offers limited or no resolution. For example, we inspected the RH data in the 80–90-Mb region of Mmu2 (Fig. 2) and found that the number of obligate breaks between the framework markers can be reduced if these markers are arranged according to the order in the clone map. We made the same observation in all eight cases examined in detail (see Supplementary Information). In general, it appears that the conflicts are caused by local inversions owing to a few (typically one or two) suboptimally located framework markers in the RH map. This in turn causes the order of all other markers previously placed in the RH map relative to that framework marker to be in conflict with their order in the clone map.

We carried out an additional check of contig assembly by comparing 'virtual' fingerprints calculated from finished mouse

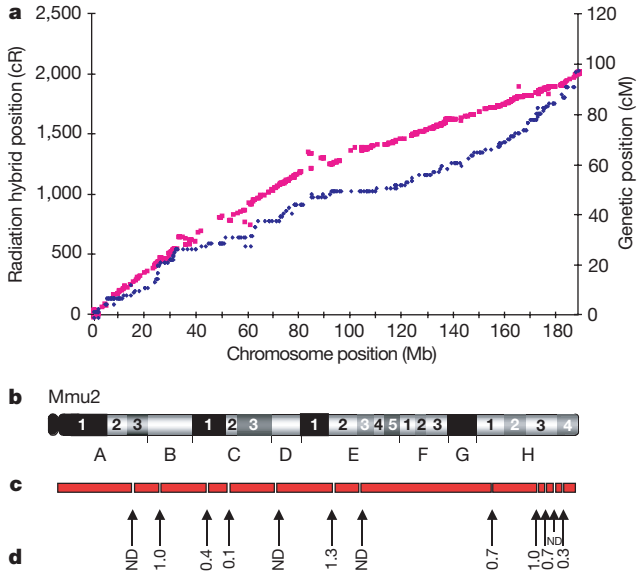


Figure 2 Comparison of radiation hybrid and genetic maps to the physical map of mouse chromosome 2. **a**, Radiation hybrid (pink) and genetic (dark blue) markers are plotted against their respective positions in the physical map of chromosome 2. **b**, Mouse chromosome 2 ideogram. **c**, **d**, Mouse BAC contigs (**c**) and gap sizes (**d**), estimated from bridging human sequence as described in the text. cR, centiRays; cM, centimorgans; ND, not determined.

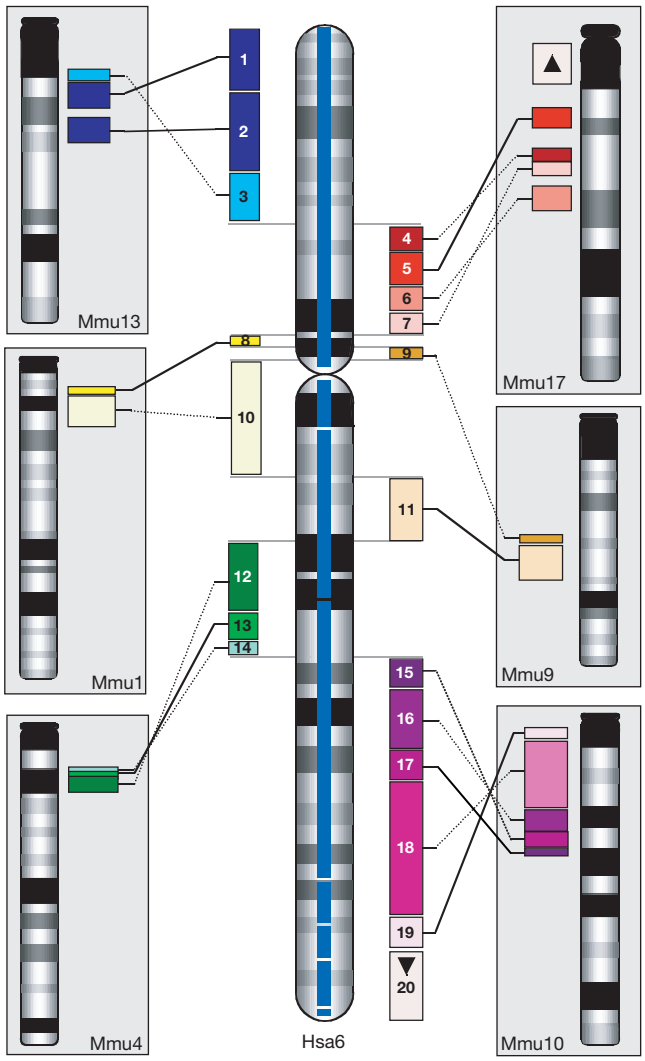


Figure 3 Conserved segments between human chromosome 6 (Hsa6) and the mouse genome. Blue bars in the centre of the chromosome ideogram represent the current extent of contig coverage in the human map. Numbered accession blocks correspond to Table 3 and denote conserved synteny with the mouse chromosomes. Dotted lines joining coloured blocks between the two species indicate conserved syntenic blocks inverted with respect to each other relative to the orientation of each chromosome as drawn.

Table 3 **Conserved segments between human chromosome 6 and mouse**

Segment	Human ptel–qtel number*	GenBank accession numbers	Human segment size (Mb)	Human–mouse homology hits	Hits per Mb	Mouse chromosome number	Mouse segment size (Mb)
6.1	3–97	AL365272–AL031785	9.0	220	24	13	9.9
6.2	99–207	AL031904–AL022726	10.3	303	29	13	9.0
6.3	209–302	AL008627–AL049543	9.5	146	15	13	9.9
6.4	310–374	AL022727–Z97183	5.6	89	16	17	3.4
6.5	376–432	AL161903–AL035690	5.1	116	23	17	4.4
6.6	434–453	AL136087–AL031778	1.8	37	21	17	1.9
6.7	458–537	AL139331–AL390247	6.5	178	27	17	6.6
6.8	560–582	AL360175–AL109918	2.7	70	26	1	3.1
6.9	582–613	AL109918–AL589796	3.1	57	18	9	2.9
6.10	620–771	AL031779–AL365232	12.4	204	16	1	13.2
6.11	773–900	AL603910–AL136082	11.1	257	23	9	10.9
6.12	918–998	AL096817–AL513186	7.0	174	25	4	7.5
6.13	1,005–1,031	AL607077–AL589740	2.9	45	16	4	2.7
6.14	1,033–1,044	Z84482–AL390959	1.1	24	22	4	1.7
6.15	1,048–1,120	AL080285–AL645528	6.8	120	18	10	8.3
6.16	1,121–1,227	AL355586–AL121953	9.2	157	17	10	10.4
6.17	1,229–1,297	AL354716–AL512283	6.2	141	23	10	7.0
6.18	1,301–1,579	AL591428–AL078581	26.1	616	24	10	27.7
6.19	1,584–1,637	AL355497–AL591419	3.9	65	17	10	4.1
6.20	1,639–1,797	AL591499–AL031259	15.0	246	16	17	11.4
Total			155.3	3,265	21		156.0

*Numbered arbitrarily from 1 (nearest the telomere of the short arm) to 1,800 (nearest the long-arm telomere).

sequence with the experimental fingerprints in the database. This was possible for 762 clones, each with more than 100 kb of finished sequence available. All virtual fingerprints were matched with high confidence to an experimentally derived fingerprint in the database, thus confirming the accuracy of the fingerprint data. Some 97% of them also matched to the expected position in contigs, confirming the assembled BAC map. The remaining 3% (20 clones) matched to other unique locations. These can be attributed to errors in tracking clone names during the project, and are generally resolved later by checking the overlaps between clones using sequence information as it becomes available.

Assuming the mouse genome is about 2.8 Gb in total (see Methods), coverage of the mouse genome in mapped BACs is virtually complete: 296 contigs of average size 9.3 Mb cover an estimated 2,739 Mb. The estimated coverage of each mouse chromosome by BAC contigs is listed in Table 2. Four contigs were placed on the mouse Y chromosome. The low coverage is partly due to the lack of previously mapped markers, and partly because only part of the fingerprinted BAC collection was derived from a male mouse. A number of the unplaced contigs contain clones derived solely from the male mouse BAC library, and may also map to the mouse Y chromosome when more data is available. The map continues to be updated, and the current version can be viewed at http://www.ncbi.nlm.nih.gov/genome/guide/mouse (choose 'MapViewer') or at http://www.ensembl.org/Mus_musculus/cytoview (use the 'Jump to chr.' box to select chromosome; if searching specific features, select 'cytoview' option to view the map). The archive of the map at the time of publication is available as Supplementary Information. The level of continuity of the mouse clone map is similar to that of the current human genome map, which comprises 279 contigs covering approximately 2,928 Mb (as of March 2002; International Human Genome Sequencing Consortium, I. Vastrik, personal communication).

There are 275 gaps in the mouse clone map. Many of these gaps occur where there is a break in synteny between the two genomes. No direct measurement of the existing gaps has yet been attempted (for example by fluorescent *in situ* hybridization to chromosomes or DNA fibres). However, 83 gaps in the mouse map are bridged by human contigs. The total length of human sequence across these gaps is 51.9 Mb. If we assume collinearity between the mouse and human genomes across these gaps, then the average size of each gap in the mouse map is 0.63 Mb. On mouse chromosome 2, for example, 9 of the 12 gaps between contigs are bridged by segments of the human sequence ranging from 0.3 to 1.3 Mb (see Fig. 2). More practically, the tracts of human sequence that bridge gaps in the mouse map may provide a source of probes (for example, selected from annotated exons in the human sequence) to screen for new mouse clones and thus assist in closing the gaps. Conversely, mouse clones that are collinear with gaps in the human tile path may be used to assist closure of the human map. Our study indicates that 132 of the 244 gaps in the human map (March 2002 release) are bridged by mouse contigs and might be closed in this way.

Human–mouse homology maps

Alignment of the human and mouse clone maps using 51,486 homology crosslinks allowed us to define the segments that are conserved in both genomes at a new level of resolution. In many cases we were able to locate the boundary between adjacent segments to within one or a few clones. For example, the sequence of human chromosome 6 contains 1,800 sections that overlap to form eight contigs. We numbered each section arbitrarily from 1 (nearest the telomere of the short arm) to 1,800 (nearest the long-arm telomere) for the purpose of this analysis (Table 3, 'ptel–qtel' numbers). We detected 220 high or medium stringency matches between a 9-Mb region of human 6p25.3 (segment 6.1, represented by accessions 3–97) and a region of about 9.9 Mb in a mouse BAC contig mapped to Mmu13A3 (Table 3 and Fig. 3). In addition, we found that the order of all 220 matches is consistent between the two maps, supporting the possibility that this is a conserved segment. The next segment (segment 6.2) starts in human accession 99 and extends as far as accession 207. This segment is defined by 303 matches of consistent order with a contig on Mmu13. Segments 6.1 and 6.2 are discontinuous in the mouse. The next segment (6.3) is aligned to another discontinuous region of Mmu13, whereas the fourth block (6.4) on human 6p is syntenic with a different mouse chromosome, Mmu17 (Table 3 and Fig. 3). In total, 20 conserved segments were identified along human chromosome 6, which subdivide the nine regions of conserved synteny (Fig. 3; see also Supplementary Information and updated views at http://www.ensembl.org/Homo_sapiens/syntenyview). The results of this analysis can also be viewed with respect to the mouse chromosomes. For example, mouse chromosome 11 contains five regions of conserved synteny with the human genome, which are subdivided into 21 conserved segments (Fig. 4; see also Supplementary Information and updated views at http://www.ensembl.org/Mus_musculus/syntenyview).

Overall, we defined 288 conserved segments contained within 167 regions of conserved synteny (Fig. 5). This compares with a previous

analysis using 2,997 markers, in which 183 conserved segments were detected[6]. The existence of multiple adjacent (but discontinuous) segments within the same syntenic region (for example segments 6.1–6.3 and 6.15–6.19 in Table 2 and Fig. 3) supports the hypothesis that a significant number of intrachromosomal inversions, in addition to interchromosomal rearrangements, contributed to formation of the present-day structures from the chromosomes of a common ancestor. Our results are consistent with previous observations for several chromosomal regions[34–37], including the recent high-resolution studies on human chromosomes 19 (refs 21, 38), 21 (ref. 39) and 7 (ref. 20). In the latter study, for example, comparative analysis using a high density of markers resulted in detection of 20 conserved segments within 16 regions of conserved synteny (compared with 25 and 15, respectively, in the present study). Small conserved segments will inevitably escape detection in the absence of sufficient homology crosslinks. For example, the Mmu2–Hsa7 segment reported in the previous study was estimated to be probably less than 100 kb (ref. 20), and was not detected in our study here. Anecdotal evidence from detailed comparisons of finished sequence between the mouse and human chromosomes indicates the presence of very small regions (as little as 1 kb) where conserved order is disrupted, either by recent duplications or insertion of unrelated sequence[23,34,39,40]. Ultimately, the alignment of finished sequences of the two genomes will be required to define the exact number and boundaries of every conserved segment.

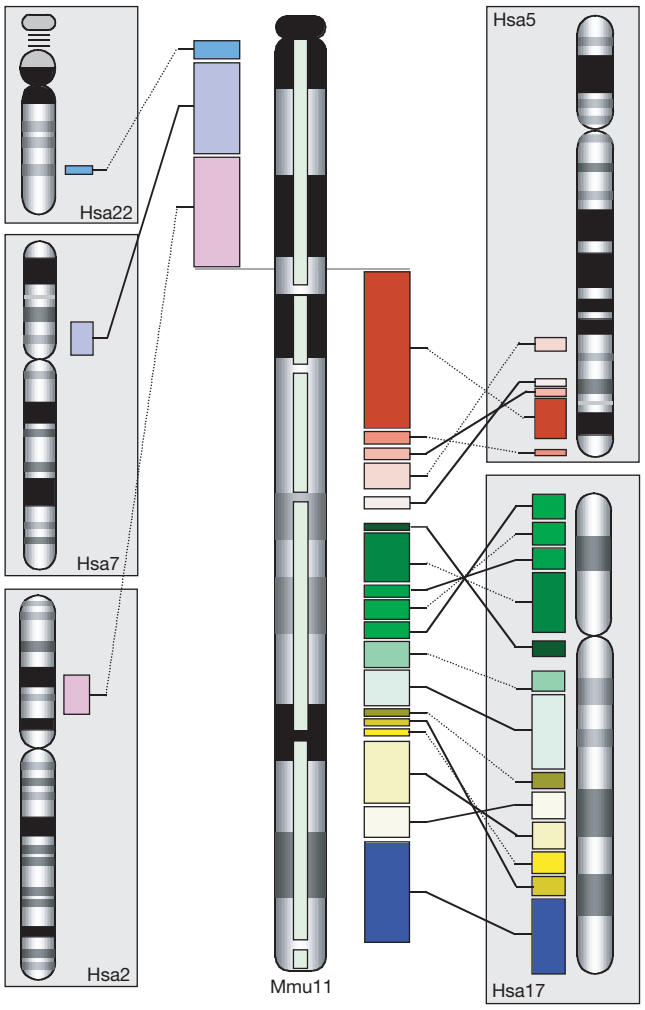


Figure 4 Conserved segments between mouse chromosome 11 and the human genome. The pale green bars in the centre of the chromosome ideogram represent the current extent of contig coverage in the mouse map. Dotted lines between the two species of same coloured blocks indicate conserved syntenic blocks inverted with respect to each other relative to the orientation of each chromosome as drawn.

Future work

Physical clone maps of complex genomes provide an essential framework to produce complete genome sequences. The clones also provide a source of well-characterized material for experimental studies. In the absence of large-scale sequencing programmes (and these may not be expected for a number of genomes in the foreseeable future), the availability of an accurate clone map will allow targeted sequencing of regions of specific interest. The alignment of maps of multiple genomes will also permit rapid investigation of orthologous gene sequences for structural and functional comparisons.

The first genomes to be mapped and sequenced in their entirety have been the result of enormous and expensive efforts. A hallmark of these studies has been the quality of the product, which has benefited from cross-validation between independently constructed maps for long-range information, and from the high accuracy of finished sequence. In our study, the availability of the human genome sequence has been exploited for faster characterization of the mouse genome. This approach will be applicable to many other mammals, and may also work for more distantly related vertebrates. Three factors contribute to the success of the mapping strategy: (1) good coverage of the genome in large contigs after the initial fingerprinting; (2) high-quality sequence cross-matches between most of these contigs and the reference sequence; and (3) sufficient independently mapped markers to order and orient all the contigs in the new map. Future projects would benefit from a similar depth of coverage in BACs (at least 15-fold) to ensure good representation of the genome in long contigs. Some savings might be made in the number of BESs required for the mapping alignment. One way to reduce the cost of the project would be to select clones from pre-existing contigs before sequencing the insert ends. However, this would reduce the number of sequence anchor points in the final map with consequent disadvantages (see below). For genomes more closely related than human and mouse, a higher proportion of the BESs would be expected to match the reference sequence (compared with 16% for the human–mouse study), which would allow reduction in the number of BESs generated without reduction of cross-matches.

The physical clone map of the mouse has provided 98% coverage of the genome in BAC contigs of average size more than 9 Mb. From the more than 300,000 BACs in the map, a complete tiling path of clones is being selected for production of the finished sequence. So

Figure 5 Homology maps of the mouse and human genomes. **a**, The mouse–human homology map. Bars in the centre of each mouse chromosome ideogram denote the ordered and oriented mouse contigs determined by this study. The coloured blocks adjacent to each mouse chromosome indicate discrete regions where marker order is conserved between the two genomes. Colours correspond to specific human chromosomes in **b**. Positions of human telomeres are located within mouse contigs by arrows in the colour of their corresponding chromosome: up arrow indicates the p telomere and down arrow indicates the q telomere. The X-chromosome relationship is represented by arrows drawn in the direction Xp to Xq from the human. **b**, The human–mouse homology map. Bars in the centre of human chromosome ideograms denote clone contigs in the human genome. Adjacent coloured blocks indicate discrete regions where marker order is conserved with the mouse. Colours correspond to specific mouse chromosomes as in **a**. An interactive display of these figures is available (http://www.sanger.ac.uk/Projects/M_musculus/publications/fpcmap-2002). Regularly updated synteny information is also available (http://www.ensembl.org/Mus_musculus/syntenyview and http://www.ensembl.org/Homo_sapiens/syntenyview). ▶

far, about 2.8% of the mouse genome sequence is available in this form (http://www.ncbi.nlm.nih.gov), and completion is scheduled for 2005. The mouse map contains more than 450,000 BESs (covering 220 Mb, or 7.8%, of the mouse genome), most of which contain sufficient unique sequence to anchor the current mouse whole genome sequence assemblies of 40 million reads (~6× coverage; http://trace.ensembl.org). This combined resource will provide a comprehensive working draft of the mouse genome in the public domain for immediate use. For example, the January 2002 release of this product included 20,652 annotated transcripts (http://www.ensembl.org/Mus_musculus/stats). The same sequence is also assisting the identification of genes on the basis of sequence conservation[41] where no human cDNA is available. Detailed comparison between human and mouse sequences will therefore provide a comprehensive catalogue of genes (and orthologous-pair relationships where they exist) for comparative functional and genetic studies. It will also be possible to identify the evolutionary events at the sequence level that have occurred since divergence of the two species. □

Methods

Map construction

Fingerprints of RPCI-23 and RPCI-24 mouse BAC clones[22] were generated by separation of *Hin*dIII restriction-enzyme digest fragments on 1.2% agarose gels[42]. Gel images were acquired with Image software (http://www.sanger.ac.uk/Software/Image) and restriction fragments identified using BANDLEADER software (D. Fuhrmann *et al.*, unpublished). Fingerprints were assembled within the program FPC[43] under high stringency conditions at a probability of 1×10^{-16} and a match tolerance of seven. The BESs[44] from clones within the fingerprint database were then matched to the tile path of human sequence clones. Human tile path accessions and mouse BESs were repeat masked and then aligned by BLASTN version 2.0a13MP-WashU, 10 June 1997 (http://blast.wustl.edu; W. Gish, personal communication)[45]. All matches longer than 100 bp, and with a significance value above 0.01, were clustered using the initial fingerprint assembly. A 'high score' blast match (that is, with a blast score >700) was required to anchor each mouse map contig to the human tile path, supported by at least one other BLAST match (Fig. 1). Only matches with unique locations in the human tile path were used. Matches between the mouse BESs and human accessions had an average score of 627.29 and an average identity of 72.3%. Contigs that were adjacent to each other on the basis of their alignment to the human sequence were inspected manually and joined if fingerprints of the end clones overlapped with probability scores of better than 1×10^{-12}. Markers were added to the map either by electronic PCR[31], or by hybridization using overgo probes[33]. Markers were incorporated into the physical map only if they were uniquely placed within the map. Additional joins were made between contigs that were adjacent according to consistent mouse marker data if supported additionally by fingerprint overlaps with probability scores of better than 1×10^{-10}.

Coverage estimates

The exact length of 125 mouse clones, for which finished sequence was available (totalling 25,349 kb) was divided by the number of *Hin*dIII fragments (5,597) observed as single bands in the fingerprints of these clones. This provided a conversion factor of 4.5 kb per observed fragment, which was used to calculate all contig sizes in the consensus map as listed in Table 2. An estimate of 2.8 Gb for the size of the mouse genome was obtained as follows. Assembled mouse whole-genome shotgun data (November release, covering a total of 2.41 Gb) contained 91% coverage of 41 Mb of finished mouse genomic sequence. On this basis, the mouse genome is estimated to cover ~2.8 Gb (2.65 Gb euchromatin plus heterochromatin). Length estimates for individual mouse chromosomes were taken from refs 46 and 47, adjusted using the estimate of whole genome size described above (see Table 2). This result suggests that the mouse genome is shorter than the human genome. Support for this was obtained by comparing contiguous and finished mouse and human genomic sequence in 13 regions that could be aligned. The combined total lengths were 6.29 Mb (mouse) and 7.78 Mb (human) (ratio 0.8), although values for individual segments varied between 0.73 and 1.06. These results are in line with previous observations[34,35,39,48–50].

Received 15 April; accepted 4 July 2002; doi:10.1038/nature00957.
Published online 4 August 2002.

1. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
2. Churcher, C. *et al.* The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IX. *Nature* **387,** 84–87 (1997).
3. The yeast genome directory. *Nature* **387** (Suppl.), 5 (1997).
4. The *C. elegans* Sequencing Consortium Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
5. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster. Science* **287,** 2185–2195 (2000).
6. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
7. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
8. Collins, J. & Hohn, B. Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Proc. Natl Acad. Sci. USA* **75,** 4242–4246 (1978).
9. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89,** 8794–8797 (1992).
10. Coulson, A., Sulston, J., Brenner, S. & Karn, J. Towards a physical map of the genome of the nematode *Caenorhabditis elegans. Proc. Natl Acad. Sci. USA* **83,** 7821–7825 (1986).
11. Olson, M. V. *et al.* Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl Acad. Sci. USA* **83,** 7826–7830 (1986).
12. Bentley, D. R. *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409,** 942–943 (2001).
13. McPherson, J. D. *et al.* A physical map of the human genome. *Nature* **409,** 934–941 (2001).
14. Green, E. D. Strategies for the systematic sequencing of complex genomes. *Nature Rev. Genet.* **2,** 573–583 (2001).
15. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
16. Waterston, R. & Sulston, J. E. The Human Genome Project: reaching the finish line. *Science* **282,** 53–54 (1998).
17. Nadeau, J. H. & Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA* **81,** 814–818 (1984).
18. DeBry, R. W. & Seldin, M. F. Human/mouse homology relationships. *Genomics* **33,** 337–351 (1996).
19. Nadeau, J. H. & Sankoff, D. Counting on comparative maps. *Trends Genet.* **14,** 495–501 (1998).
20. Thomas, J. W. *et al.* Comparative genome mapping in the sequence-based era: early experience with human chromosome 7. *Genome Res.* **10,** 624–633 (2000).
21. Kim, J. *et al.* Homology-driven assembly of a sequence-ready mouse BAC contig map spanning regions related to the 46-Mb gene-rich euchromatic segments of human chromosome 19. *Genomics* **74,** 129–141 (2001).
22. Osoegawa, K. *et al.* Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.* **10,** 116–128 (2000).
23. Hardison, R. C., Oeltjen, J. & Miller, W. Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **7,** 959–966 (1997).
24. Koop, B. F. *et al.* The human T-cell receptor TCRAC/TCRDC (Cα/Cδ) region: organization, sequence, and evolution of 97.6 kb of DNA. *Genomics* **19,** 478–493 (1994).
25. Ashworth, L. K. *et al.* An integrated metric physical map of human chromosome 19. *Nature Genet.* **11,** 422–427 (1995).
26. Kent, W. J. The BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
27. Copeland, N. G. *et al.* A genetic linkage map of the mouse: current applications and future prospects. *Science* **262,** 57–66 (1993).
28. Dietrich, W. F. *et al.* A comprehensive genetic map of the mouse genome. *Nature* **380,** 149–152 (1996).
29. Cai, W. W. *et al.* An SSLP marker-anchored BAC framework map of the mouse genome. *Nature Genet.* **29,** 133–134 (2001).
30. Hudson, T. J. *et al.* A radiation hybrid map of mouse genes. *Nature Genet.* **29,** 201–205 (2001).
31. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* **7,** 541–550 (1997).
32. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
33. Ross, M. T., LaBrie, S., McPherson, J. P. & Stanton, V. P. in *Current Protocols in Human Genetics* (eds Dracopoli, N. C. *et al.*) 5.6.1–5.6.5 (Wiley, New York, 1999).
34. Puttagunta, R. *et al.* Comparative maps of human 19p13.3 and mouse chromosome 10 allow identification of sequences at evolutionary breakpoints. *Genome Res.* **10,** 1369–1380 (2000).
35. Carver, E. A. & Stubbs, L. Zooming in on the human–mouse comparative map: genome conservation re-examined on a high-resolution scale. *Genome Res.* **7,** 1123–1137 (1997).
36. Watkins-Chow, D. E. *et al.* Genetic mapping of 21 genes on mouse chromosome 11 reveals disruptions in linkage conservation with human chromosome 5. *Genomics* **40,** 114–122 (1997).
37. Stubbs, L. *et al.* Detailed comparative map of human chromosome 19q and related regions of the mouse genome. *Genomics* **35,** 499–508 (1996).
38. Dehal, P. *et al.* Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293,** 104–111 (2001).
39. Pletcher, M. T., Wiltshire, T., Cabin, D. E., Villanueva, M. & Reeves, R. H. Use of comparative physical and sequence mapping to annotate mouse chromosome 16 and human chromosome 21. *Genomics* **74,** 45–54 (2001).
40. Oeltjen, J. C. *et al.* Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* **7,** 315–329 (1997).
41. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10,** 950–958 (2000).
42. Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7,** 1072–1084 (1997).
43. Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4. 7. *Genome Res.* **10,** 1772–1787 (2000).
44. Zhao, S. *et al.* Mouse BAC ends quality assessment and sequence analyses. *Genome Res.* **11,** 1736–1745 (2001).
45. Altschul, S. F. & Gish, W. Local alignment statistics. *Methods Enzymol.* **266,** 460–480 (1996).
46. Nusbaum, C. *et al.* A YAC-based physical map of the mouse genome. *Nature Genet.* **22,** 388–393 (1999).
47. Evans, E. P. in *Genetic Variants of the Laboratory Mouse* (eds Lyon, M. F., Rastan, S. & Brown, S. D. M.) 1446 (Oxford Univ. Press, New York, 1996).
48. Pletcher, M. T. *et al.* Chromosome evolution: the junction of mammalian chromosomes in the formation of mouse chromosome 10. *Genome Res.* **10,** 1463–1467 (2000).
49. Martindale, D. W. *et al.* Comparative genomic sequence analysis of the Williams syndrome region (LIMK1-RFC2) of human chromosome 7q11.23. *Mamm. Genome* **11,** 890–898 (2000).
50. DeSilva, U. *et al.* Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. *Genome Res.* **12,** 3–15 (2002).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature). Updated views of the map are available from the authors' websites (http://www.ensembl.org/Mus_musculus/cytoview and http://www.ncbi.nlm.nih.gov/genome/guide/mouse), as is an archive version for this publication, plus comparisons and discussion of genetic, RH and clone maps, and an interactive version of the synteny displays of Fig. 5 (http://www.sanger.ac.uk/Projects/M_musculus/publications/fpcmap-2002).

Acknowledgements

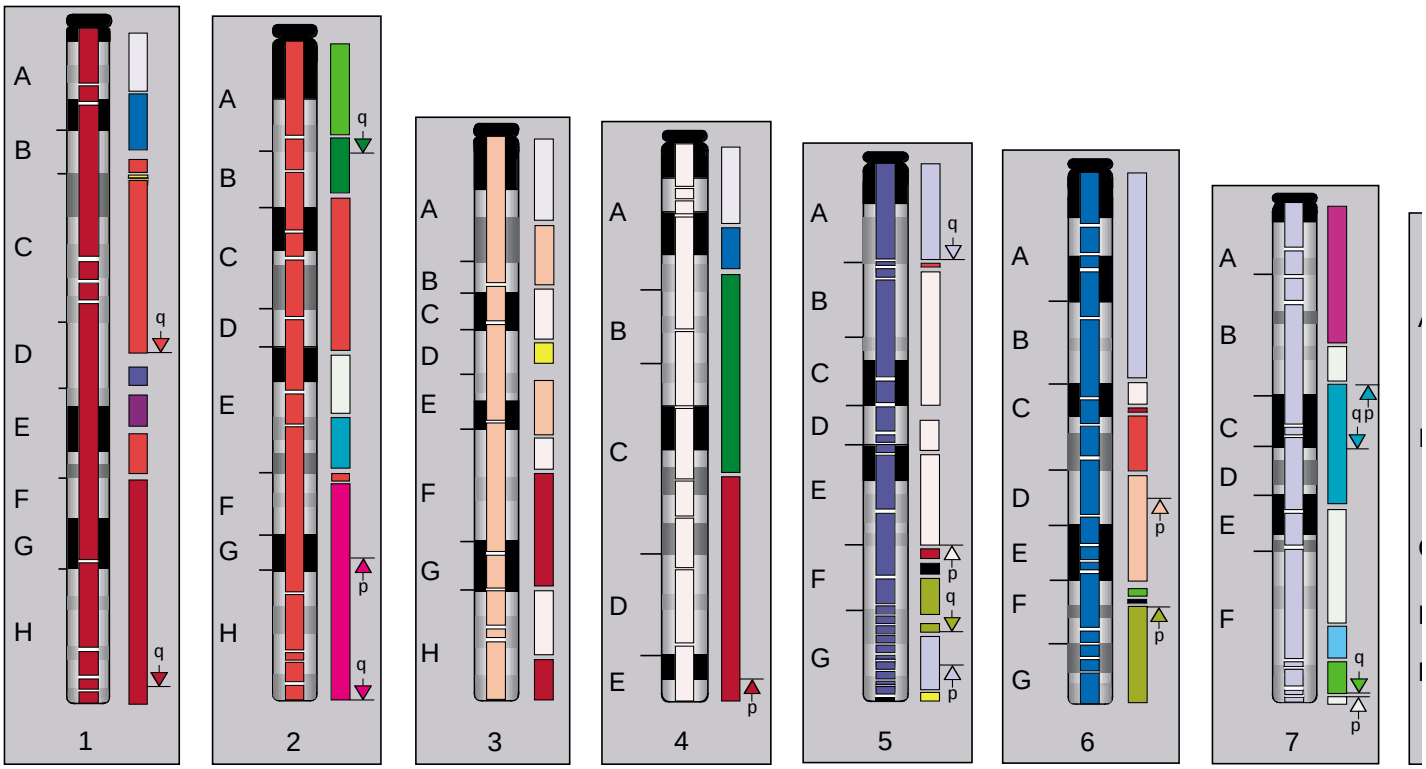
The authors acknowledge the support of the Wellcome Trust, the National Institutes of Health and the US Department of Energy. We are grateful to the web team at the Sanger Institute for assistance with developing map displays, to P. Deloukas for RH map analysis, and E. Arnold-Berkowits, S. Lo, J. Gill and all present and past members of the Institute for Genomic Research BAC end sequencing team for the sequencing work.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.R.B. (e-mail: drb@sanger.ac.uk).

a Mouse–human homology map



b Human–mouse homology map