A Comparison of Whole-Genome Shotgun-Derived Mouse Chromosome 16 and the Human Genome

Richard J. Mural,1*, Mark D. Adams,1 Eugene W. Myers,1 Hamilton O. Smith,1 George L. Gabor Miklos,2 Ron Wides,3 Aaron Halpern,1 Peter W. Li,1 Granger G. Sutton,1 Joe Nadeau,4 Steven L. Salzberg,5 Robert A. Holt,1 Chinnappa D. Kodira,1 Fu Lu,1 Lin Chen,1 Zuoming Deng,1 Carlos C. Evangelista,1 Weiniu Gan,1 Thomas J. Heiman,1 Jiayin Li,1 Zheny Li,1 Gennady V. Merkulov,1 Natalia V. Milshina,1 Ashwinikumar K. Naik,1 Rong Qi,1 Bixiong Chris Shue,1 Ai-Hui Wang,1 Jian Wang,1 Xin Wang,1 Xianghe Yan,1 Jane Ye,1 Shibu Yooseph,1 Qi Zhao,1 Liangsheng Zheng,1 Shaoping C. Zhu,1 Kendra Biddick,1 Randall Bolanos,1 Arthur L. Delcher,1 Ian M. Dew,1 Michael J. Flanigan,1 Daniel H. Huson,1 Saul A. Kravitz,1 Jason R. Miller,1 Clark M. Moberly,1 Knut Reinert,1 Karin A. Remington,1 Qing Zhang,1 Xiangjun H. Zheng,1 Deborah R. Nusskern,1 Zhongwu Lai,1 Yiding Lei,1 Wenyuan Zhong,1 Alison Yao,1 Ping Guan,1 Rui-Ru Ji,1 Zhiping Gu,1 Zhen-Yuan Wang,1 Fei Zhong,1 Chunlin Xiao,1 Chia-Chien Chiang,1 Mark Yandell,1 Jennifer R. Wortman,1 Peter G. Amanatides,1 Suzanne L. Hladun,1 Eric C. Pratts,1 Jeffery E. Johnson,1 Kristina L. Dodson,1 Kenny J. Woodford,1 Cheryl A. Evans,1 Barry Gropman,1 Douglas B. Rusch,1 Eli Venter,1 Mei Wang,1 Thomas J. Smith,1 Jarrett T. Houck,1 Donald E. Tompkins,1 Charles Haynes,1 Debbie Jacob,1 Soo H. Chin,1 David R. Allen,1 Carl E. Dahlke,1 Robert Sanders,1 Kevin Li,1 Xiangjun Liu,1 Alexander A. Levitsky,1 William H. Majors,1 Quan Chen,1 Ashley C. Xia,1 John R. Lopez,1 Michael T. Donnelly,1 Matthew H. Newman,1 Anna Glodek,1 Cheryl L. Kraft,1 Marc Nodell,1 Feroze Ali,1 Hui-Jin An,1 Danita Baldwin-Pitts,1 Karen Y. Beeson,1 Shuang Cai,1 Mark Carnes,1 Amy Carver,1 Parris M. Caulk,1 Angela Center,1 Yen-Hui Chen,1 Ming-Lai Cheng,1 My D. Coyne,1 Michelle Crowder,1 Steven Danaher,1 Lionel B. Davenport,1 Raymond Desilets,1 Susanne M. Dietz,1 Lisa Doup,1 Patrick Dullaghan,1 Steven Ferriera,1 Carl R. Fosler,1 Harold C. Gire,1 Andres Gluecksmann,1 Jeannine D. Gocayne,1 Jonathan Gray,1 Brit Hart,1 Jason Haynes,1 Jeffery Hoover,1 Tim Howland,1 Chinarye Ibegwam,1 Mena Jalali,1 David May,1 Lisa J. Johnson,1 Leslie Kline,1 Daniel S. Ma,1 Steven MacCawley,1 Anand Magono,1 Felecia Manton,1 David May,1 Tina C. McIntosh,1 Somil Mehta,1 Linda Moy,1 Mei C. Moy,1 Brian J. Murphy,1 Sean D. Murphy,1 Keith A. Nelson,1 Zubeda Nuri,1 Kimberly A. Parker,1 Alexandre C. Pudhomme,1 Vinita N. Puri,1 Hina Qureshi,1 John C. Raley,1 Matthew S. Reardon,1 Megan A. Regier,1 Yu-Hui C. Rogers,1 Deanna L. Romblad,1 Jakob Schutz,1 John L. Scott,1 Richard Scott,1 Cynthia D. Sitter,1 Michella Smallwood,1 Arlan C. Sprague,1 Erin Stewart,1 Renee V. Strong,1 Ellen Su,1 Karena Sylvester,1 Reginald Thomas,1 Ni Ni Tint,1 Christopher Tsonis,1 Gary Wang,1 George Wang,1 Monica S. Williams,1 Sherita M. Williams,1 Sandra M. Windsors,1 Keriellen Wolfe,1 Mitchell M. Wu,1 Jayshree Zaveri,1 Kabir Chaturvedi,1 Andrei E. Gabrielian,1 Zhaoxi Ke,1 Jingtao Sun,1 Gangadharan Subramanian,1 J. Craig Venter1†

1Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. 2GenetixXpress, 81 Bynya Road, Palm Beach, Sydney, 2108, Australia. 3Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel. 4Department of Genetics, Case Western Reserve University School of Medicine, Center for Computational Genomics, Case Western Reserve University, and Center for Human Genetics, University Hospitals of Cleveland, Cleveland, OH 44106, USA. 5To whom correspondence should be addressed. E-mail: richard.mural@celera.com
†Present address: TIGR Center for the Advancement of Genomics, 1901 Research Boulevard, Suite 600, Rockville, MD 20850, USA.

The nearly complete sequence of a typical mammalian genome is demonstrated through analysis of the sequence of mouse chromosome 16 (Mmu 16), which was obtained as part of a whole-genome shotgun assembly of the mouse genome. The mouse genome is about 10% smaller than the human genome, owing to a lower repetitive DNA content. Comparison of the structure and protein-coding potential of Mmu 16 with that of the homologous segments of the human genome identifies regions of conserved synteny with human chromosomes (Hsa) 3, 8, 12, 16, 21, and 22. Gene content and order are highly conserved between Mmu 16 and the syntenic blocks of the human genome. Of the 731 predicted genes on Mmu 16, 509 align with orthologs on the corresponding portions of the human genome, 44 are likely paralogous to these genes, and 164 genes have homologs elsewhere in the human genome; there are 14 genes for which we could find no human counterpart.

The high degree of similarity between the mouse and human genomes is demonstrated through analysis of the sequence of mouse chromosome 16 (Mmu 16), which was obtained as part of a whole-genome shotgun assembly of the mouse genome. The mouse genome is about 10% smaller than the human genome, owing to a lower repetitive DNA content. Comparison of the structure and protein-coding potential of Mmu 16 with that of the homologous segments of the human genome identifies regions of conserved synteny with human chromosomes (Hsa) 3, 8, 12, 16, 21, and 22. Gene content and order are highly conserved between Mmu 16 and the syntenic blocks of the human genome. Of the 731 predicted genes on Mmu 16, 509 align with orthologs on the corresponding portions of the human genome, 44 are likely paralogous to these genes, and 164 genes have homologs elsewhere in the human genome; there are 14 genes for which we could find no human counterpart.
mouse chromosome (Mmu 16), presented here, was determined as part of the whole-genome shotgun sequencing and assembly of the mouse genome, a strategy that was used successfully for the *Drosophila* (4–6) and human genomes (1). Mmu 16 was chosen from this assembly for analysis because it shares a large region of synteny, about 25 megabase pairs (Mb), with human chromosome 21 (Hsa 21), which has been extensively characterized (7). This assembly was carried out with DNA sequence derived from four strains of laboratory mice [A/J, DBA/2J, 129X1/SvJ, and 129S1/SvImJ (8a)] that were chosen partly because they complement the C57BL/6J strain, which is being sequenced by a separate mouse genome sequencing effort (8b). In addition, these four strains belong to distinct lineages of the laboratory mouse (9, 10), and they differ in numerous traits of biological and medical interest.

**Chromosome 16 Sequence**

Procedures for DNA extraction, library construction, and DNA sequencing are modified from those described in (1, 11). The resulting data set consisted of 27.4 million sequencing reads sufficient to cover the genome 5.3 times. These sequences came from both ends of stringently size-selected 2-, 10-, and 50-kbp clones derived from randomly sheared mouse genomic DNA. We estimate that the combined lengths of these clones cover the genome 44 times. Over 80% of all sequencing reads could be associated as pairs coming from the ends of any given clone.

This data set, generated solely at Celera, was then analyzed with the whole-genome assembler previously used to produce the sequence of the *Drosophila* and human genomes (1, 4). To evaluate the chromosome 16 statistics, we point out that this whole-genome assembly resulted in 19,788 scaffolds (contigs that are ordered and oriented with information from paired reads) spanning 2446 Mb of the mouse genome. In this assembly, 50% of the bases (the so-called N50 statistic) are in scaffolds of at least 4.476 Mbp, and the N50 contig size is 14,559 bp (12).

We mapped whole-genome assembly scaffolds to the chromosomes by pairing the locations of known markers on the scaffold to the locations of the same markers on public genetic sequence. The results indicated that all the suspect regions have strong clone and fragment coverage, suggesting that assembly was unlikely to be the source of the unexpected placement (12). We then used the bin assignments of STS markers from Whitehead genetic and RH maps to order and orient the scaffolds. One additional scaffold of 105 kbp was recruited to chromosome 16 because it contained end sequences from two bacterial artificial chromosome (BAC) clones, which were mapped to Mmu 16 (14).

Twenty scaffolds (92 Mbp) mapped to chromosome 16, 14 of which are >1 Mbp in size, with the largest being 15.65 Mbp (Table 2). These scaffolds contain 8635 contigs that cover 87,008,971 bp; the longest contig is 117,734 bp. The smallest six scaffolds cover only 434,340 bp, so that more than 99% of the bases in the mapped scaffolds are in scaffolds of >1 Mbp in length. Of the 19,788 total scaffolds, 2260 (95% of total scaffold length) could be assigned to chromosome locations. Therefore, based on the unmapped portion of the assembled mouse genome (53.9 Mbp, excluding the Y chromosome) (12), we estimate that an additional 2 to 3 Mbp of DNA maps to Mmu 16.

The sequence that is missing in the gaps between contigs and scaffolds is largely composed of (i) short regions that lacked any read coverage due to random sampling and (ii) repeats that could not be entirely filled with mate pairs. Based on our experience with the human genome, we expect that some large (>20 kbp), nearly identical duplicated regions of the genome might be underrepresented in the scaffolds and contribute to the number of interscaffolding gaps.

Perhaps the best measure of assembly accuracy is comparison with independently sequenced clones from the same genome. Seven BACs from the Cat eye syndrome region on chromosome 16 (15) have been sequenced, and the structure and sequence of the Celera scaffold corresponding to this region are in good agreement with the structure and sequence of each BAC clone (Fig. 1).

Fig. 1. Dot-plot analysis of a portion of Celera (CRA) scaffold GA_x5J8B7WYGF and BAC sequences. Each dot represents a match of at least 98% sequence identity and is plotted at 200 bp per pixel. GenBank accession numbers from left to right: AC018559, AC007844, AC012397, AC009192, AC006447, AC006404, AC006945. Small discontinuities or “chatter” in the plot reflect gaps between contigs in the scaffold assembly (which are expected) and do not indicate problems with contig order and/or orientation.

### Table 1. Scaffold statistics for mouse chromosome 16.

<table>
<thead>
<tr>
<th>Category</th>
<th>Total Scaffold Count (kbp)</th>
<th>Total Contig Count (kbp)</th>
<th>N50 Scaffold Size (kbp)</th>
<th>N50 Contig Size (kbp)</th>
<th>Number of Gaps</th>
<th>Gap Size (kbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffold</td>
<td>19,788</td>
<td>2260</td>
<td>10,976</td>
<td>16</td>
<td>8,615</td>
<td>5,602</td>
</tr>
<tr>
<td>Bounded and orient</td>
<td>12</td>
<td>13a, 13b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>86,987</td>
<td>92,574</td>
<td>495</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Fig. 1. Dot-plot analysis of a portion of Celera (CRA) scaffold GA_x5J8B7WYGF and BAC sequences. Each dot represents a match of at least 98% sequence identity and is plotted at 200 bp per pixel. GenBank accession numbers from left to right: AC018559, AC007844, AC012397, AC009192, AC006447, AC006404, AC006945. Small discontinuities or “chatter” in the plot reflect gaps between contigs in the scaffold assembly (which are expected) and do not indicate problems with contig order and/or orientation.*

*Table 2. Scaffold statistics for mouse chromosome 16.*
sequences in upstream putatively regulatory sequences have been noted between different species in the genus Caenorhabditis and between Drosophila melanogaster and D. virilis. Identification of such sequences provides a valuable complement to sets of orthologous protein sequences for several purposes—e.g., phylogenetic footprinting in regulatory regions (22), identification of novel gene candidates where no transcripts had previously been predicted, and finer-grained inference of conserved synteny and mapping of syntenic breakpoints. Consequently, paired segments of human and mouse sequence that are putatively orthologous were identified by comparison of the predicted proteins from each genome as well as by comparison of the genomic DNA sequence.

Synteny based on DNA comparisons. We use the term “syntenic anchors” to describe conserved locations in the two genomes that are identified by significant DNA sequence similarity and that constitute a bidirectionally unique match (two segments are designated as syntenic anchors if their alignment is the only significant match either segment shows to the other genome) (23). A total of 11,822 syntenic anchors mapped to chromosome 16 with a mean length of 198 bp and a mean identity of 88.1%; 11,496 chromosome 16 anchors matched human scaffolds that had been mapped to a specific chromosome, with an additional 426 anchors mapping to a human scaffold whose location in the genome is unknown (Table 3). This corresponds to an anchor about every 8 kb. As expected, anchors are not uniformly spaced. Indeed, the largest “gap” between two anchors is 3,461, 469 bp for human autosomes (24), and 20% of the (human) genome is such that the distance from one anchor to the next is at least 100,000 bp. In the mouse genome, the largest gap between anchors is 2,347,210 bp; for Mmu 16, it is 707,282 bp (12).

We found that not only was there a substantially nonrandom distribution of anchors relative to genes, but also that anchors are not simply conserved exons; only 34% of all anchors overlapped annotated mouse exons, and 56% were found within the boundaries of annotated mouse genes. The remainder (44%) were in intergenic regions. Interestingly, the sizes of anchors did not differ among those found in genes and those found in intergenic regions. Moreover, as described below, the density of anchors appears to be much less affected by gene density than expected, which makes anchors an important complement to protein-based markers of conserved synteny.

Syntenic anchors between mouse and human are remarkably consistent. Over 50% of syntenic anchors on Mmu 16 are in runs of at least 128 in a row with the same order and orientation in each of the genomes and with no additional (intervening) anchors. In addition, 50% of the (human) genome is in such uninterrupted, ordered, and oriented blocks, defined solely by anchors, that are at least 994 kbp (12) long. These values are somewhat lower than would be expected given perfect, complete assemblies, because errors or unmapped sequences in one genome lead to breakpoints in the correspondence between the two genomes. Many of these breakpoints involve more than local inversion or transposition. We observed a number of small local rearrangements, inversions averaging 7.2 kbp and transpositions averaging 2.5 kbp, involving short runs of anchors (25). Of 193 breakpoints identified on Mmu 16 mapped scaffolds that were not coincident with either mouse or human scaffold boundaries, 63 were due to 32 local inversions and 10 were due to five local transpositions. By excluding breakpoints that resulted from single inconsistent anchors, we found 48 breakpoints on Mmu 16 not at scaffold boundaries, of which 22 are caused by 11 inversions and 2 are caused by a single transposition.

The separation between adjacent anchors was consistently shorter in mouse than in human for pairs of anchors contained within the same scaffolds on both assemblies (mean 7167 bp for human, 6188 bp for mouse) and for pairs contained within the same contigs (2429 bp for human, 2034 bp for mouse), which is consistent with a smaller genome size for mouse than for human, as discussed more extensively below. Mmu 16 shares regions of conserved synteny with six human chromosomes: 3, 8, 12, 16, 21, and 22 (Fig. 3) (26–31). The pattern of anchor distribution, in terms of lengths of runs of consistent anchors, is similar to that described above. Table 3 presents the total number of anchors between the regions of conserved synteny involving Mmu 16 and the corresponding portions of human chromosomes. Inconsistent anchors—i.e., anchors that are not consistently placed relative to their neighbors on both genomes—are also enumerated, and they account for only about 1.14% of all anchors between the regions of shared synteny.

In general, the length of regions of conserved synteny tend to be longer on the human chromosome than on the corresponding mouse chromosome (Fig. 3). The sizes (in kbp) of each of these chromosome segments in humans, relative to their mouse counterparts, are shown in Table 3. In every case, the human genomic block is larger than that of the corresponding segment in the mouse genome, even though it harbors a similar genetic content. In total, the mapped portion of Mmu 16 consists of 92 Mbp of DNA, whereas the sum of the corresponding human blocks is 108 Mbp. In the regions of conserved synteny with Mmu 16, for example, short interspersed nuclear elements (SINEs) account for 31.6% of the bases in the human genomic segments, whereas SINEs account for only 21.7% of the bases on Mmu 16. For the same regions, long interspersed nuclear elements (LINEs) account for 16.4% of bases in
Fig. 2. Annotation of mouse chromosome 16. Celera’s assembly and annotation is depicted in five tiers. Each tier contains nine tracks showing (top to bottom) molecular markers, genes mapped to the forward DNA strand, density plots for five ubiquitous features, genes mapped to the reverse DNA strand, and human synteny. Molecular markers are from UniSTS and dbSTS (www.ncbi.nlm.nih.gov/genome/sts/) and RHdb (http://corba.ebi.ac.uk/RHdb/). For clarity, only selected markers are shown. Each gene is depicted by a single rectangle spanning the transcribed nucleotides. As shown in the legend, color differentiates genes with putative human orthologs (syntenic homologs), putative human paralogs (nonsyntenic homologs), and genes with no human homolog (see text). For clarity, only selected gene names are shown, and adjacent genes are separated vertically. The five density plots show G+C content and LINE, SINE, EST, and anchor density. EST, LINE, and SINE densities were computed in nonoverlapping 50-kb windows; anchor density was computed in 100-kb windows overlapping by 87.5 kb; G+C content was computed in nonoverlapping 25-kb windows with interpolation of content across gaps. The results are represented with distinct nonlinear scales, illustrated in the legend; scale values show the minimum, maximum, and centile (analogous to quartile) data values. Mouse ESTs are from dbEST (www.ncbi.nlm.nih.gov/dbEST/). Anchors are homologous segments of unique sequence (see text). Synteny to human chromosomes is based on large runs of consistent anchors (see text). The figure was generated with gff2ps (66).
human and only 12.3% of bases in mouse. Genomewide SINEs plus LINEs account for about 46% of bases in the human and only about 36% of bases in the mouse genome. This difference faithfully mirrors the genomewide disparity in the sizes of the mouse and human euchromatinic regions.

**Synteny based on protein comparisons.** Another method used to identify regions of conserved synteny between the human genome and Mmu 16 was based on protein comparisons and originally was devised to identify human intragenomic duplications (1). This method identifies clusters of predicted proteins and their homologs that have conserved relative order on two different chromosomes. Identification of similar proteins is determined by the suffix-tree comparison method MUMmer (32) or, alternatively, by matches that have mutual best BLAST scores. Briefly, when at least three proteins (33) within a small interval along a chromosome can be aligned with three similar proteins along a target chromosome, this correlation of clusters is the basis for assertion of a “conserved synteny.” As applied to find syntenic stretches between the mouse and human genomes, this method found only one human chromosomal stretch for each Mmu block (34). These syntenic blocks consisted of dozens to hundreds of identically ordered genes on the mouse and human chromosomes (Fig. 3). Because similar syntenic blocks were delineated regardless of whether the criteria for protein matches were MUMmer identities between proteins or best BLASTP scores in the target genome, results for the two analyses were merged.

**Regions of conserved synteny**

![Fig. 3. Regions of conserved synteny between Mmu 16 and the human genome. The analysis was done at the protein level with the MUMmer program; each line represents a pair of orthologous genes present in the mouse and human. Cyto, cytogenetic markers; SCF, scaffold distribution; and Ori, orientation of scaffolds.](image-url)

At the criterion used (34), 99% of chromosome 16 could be mapped to single, unique homologous human chromosome segments.

For Mmu 16, the 731 gene predictions ordered along the mouse chromosome matched genes from seven distinct syntenic blocks on six human chromosomes. These blocks, in the order of their alignment along Mmu 16 from telomere to telomere, are Hsa 16, 8, 12, 22, 3 (two separate blocks aligning to two different regions of Hsa 3), and 21 (Figs. 3 and 4; Table 3). The findings agree with previous partial descriptions of mouse-human conserved synteny, with the exception of a region on Hsa 12 that has not been previously described, and augment those findings with greatly increased resolution. Ninety-eight percent (717/731) of the high-confidence gene predictions have homologs in the human genome, and 76% (556/731) have homologs in the corresponding syntenic block in human. Ninety-two percent (509/556) of these homologs are likely to be orthologs (35) of the human genes, and the remaining 8% (44/556) represent unequal local expansions of these genes, that is, paralogs that have arisen since the mouse and human lineages diverged. As discussed below, the 509 pairs of homologous genes that map to regions of conserved synteny between the mouse and human genomes are very likely to be orthologs; this assertion is weaker for the 164 pairs of homologous genes that map elsewhere. Most of these pairings probably do not represent orthologous relationships, because the percentage of identical residues for the orthologs as asserted above is 86%, whereas these pairings have, on average, 69% identical residues and because these pairings are not reciprocal best matches between the human and mouse proteins. Figure 4 illustrates the differences in the distributions of expectation values for these two sets of homologous genes.

For 14 mouse genes, we could find no related human genes in either our assembly of the human genome (1) or in any other databases (see supplementary table 1 on Science Online at www.sciencemag.org/cgi/content/full/296/5573/1661/DC1). These mouse genes could either be specific to the mouse genome, or they could have functionality active, but extremely diverged (i.e., unrecognizable), counterparts in the appropriate region of conserved synteny, or elsewhere, in the human genome. In particular cases, we found evidence for remnants of small open reading frames in syntenic regions of the human genome that correspond to remnants of orthologous mouse genes. These sequences are noncoding and almost certainly decayed footprints (pseudogenes) of orthologous mouse-human pairs. Finally, we found an additional 33 predicted “genes” that are related to retroviruses and another 65 that are pseudogenes. The last two categories are not included in the 731 predicted protein coding genes assigned to Mmu 16.

**Description of individual syntenic regions:**
the region of Mmu 16 corresponding to a block of Hsa 21. The region of conserved synteny between Mmu 16 and Hsa 21 corresponds to 22.37 Mb of the mouse chromosome and 28.42 Mb of the human chromosome. This region contains major determinants for Down syndrome, and trisomy of this region through a Robertsonian translocation (36) and segmental trisomy (37, 38) have provided mouse models for Down syndrome. We have been able to assign 129 orthologous gene pairs to this region, which extends from near the human STCH gene (microsomal stress adenosine triphosphatase core) to the ZNF295 gene (a Kelch-related transcription factor). The order of the 111 mouse genes in our assembly of Mmu 16 corresponds exactly to the order of published human genes in this region (7) except for one small, previously known inversion (39) near the Bace2 gene (position 93.2 Mb in Fig. 2).

Gene density within this synteny region is not uniform. For example, there is a region on Hsa 21 with only 7 genes in 7.8 Mbp from PRSS7 (an enterokinese) to APP (the amyloid A4 precursor protein implicated in Alzheimer’s disease). The corresponding region in the mouse is slightly smaller, about 6 Mb, and has the same gene content with no addition or loss of transcription units. The sizes of the orthologous proteins are similar, indicating that there have been no major gains or losses of protein domains with orthologs. The size and coding capacity of this synteny semidesert have remained largely intact since the divergence of lineages leading to humans and mice. In addition, about 400 conserved syntentic anchors are evenly spread throughout this desert. These conserved regions do not correspond to any of the repetitive elements that occur throughout this semidesert, and the reasons for their conservation remain obscure.

To contrast a gene-poor region with a more populated one, we examined the 17-gene region on Hsa 21 extending from the human interferon α/β receptor (IFNAR2) to DSCR1 (the human Down syndrome candidate region protein). This region is ~1.08 Mb in length in mouse and 1.4 Mb in human; it contains the human genes IFNAR2, IL10RB, IFNAR1, IFN, GART, RPSIL1, ATP50, SON, IITSN, KCNE1, and DSCR1, most of which are implicated in predispositions to various human diseases and one of which, the interleukin-10 receptor β (IL10RB), is a known drug target. This portion of Mmu 16 has the same complement of genes in exactly the same order.

We also compared our automated annotation with a previously published annotation of the 4.5-Mbp region from Cbr (carbonyl reductase) with Trimpsr2 (transmembrane protease, serine 2) (39). One mouse gene in this region, Ighb21 (integrin-β2-like), does not have an ortholog in the syntenic region of Hsa 21 between B3galts (β1,3-galactosyltransferase) and Pcp4 (Purkinje cell protein 4). We confirmed the absence of a detectable ortholog of Ighb21 and found an additional gene between Ighb21 and Pcp4.

Description of individual syntenic regions: the region of Mmu 16 corresponding to a block of Hsa 8. This region of conserved synteny covers about 1.29 Mbp of Mmu 16, which corresponds to a 1.39-Mbp region of Hsa 8. There are six orthologous gene pairs in this region.

Description of individual syntenic regions: the region of Mmu 16 corresponding to a block of Hsa 12. This particular region of synteny between the two genomes was previously unidentified. Based on its content of orthologous genes and anchors, we find that it constitutes a 0.363-Mbp region of Mmu 16, and a 0.470-Mbp segment of Hsa 12. The human segment consists of only three genes, which have orthologs in the mouse.

Description of individual syntenic regions: the region of Mmu 16 corresponding to a block of Hsa 22. The region of conserved synteny between Mmu 16 and Hsa 22 corresponds to 2.08 Mbp of the mouse chromosome and 2.27 Mbp of the human chromosome. The gene content is 30 orthologous loci that are in conserved order except for a block of eight genes, which is inverted in mouse relative to human. The entire region of conserved synteny extends from near the Top3b (DNA topoisomerase III β1) gene to near the HIRA (histone cell cycle regulation defective, Saccharomyces cerevisiae) homolog A gene. Previous reports (40) showed a segment of Mmu 16 distal to the portion corresponding to Hsa 22 as corresponding to a portion of Hsa 18. We find no syntenic region or even single genes related to those on Hsa 18. Indeed, the single reliable datum that had been used to establish this conserved synteny, the Elf-4a gene, was found in both recent publications of the human genome sequence (1, 2) to map to Hsa 3q27.2. This places it within the syntenic block between Hsa 3q27–29 and Mmu 16, which follows the block of Mmu 16 syntenic to Hsa 22, and not within Hsa 18.

Description of individual syntenic regions: the region of Mmu 16 corresponding to a block of Hsa 3q27–29. The region of conserved synteny between Mmu 16 and Hsa 3q27–q29 corresponds to ~13.56 Mb in mice and 16.46 Mb in humans. This region has 107 orthologous pairs of genes that have a conserved order between mouse and human.

Description of individual syntenic regions: the region of Mmu 16 corresponding to a block of Hsa 3q11.1–13.3. The region of conserved synteny between Mmu 16 and Hsa 3q11.1–13.3 corresponds to ~41.66 Mb in mice and 46.49 Mb in humans. The gene

![Fig. 4. Distribution of mouse-human orthologs and best hits mapping to syntenic and nonsyntenic regions based on sequence homology and expect value scores. A log expect value score of <1 × 10−180 is represented as 0.](www.sciencemag.org)
order for mouse and human appears to be conserved across this region. Much of this large chromosome segment is gene-poor in both the mouse and human genomes. This segment contains 165 orthologous pairs of genes, or about 4.15 genes (with orthologs in the human genome) per Mbp on Mmu 16 versus an average of 8.25 genes per Mbp in the other region of conserved synteny with Hsa 3 described above. The gene density is even lower for this segment of Hsa 3 with 3.72 genes per Mbp contrasting with 6.8 genes per Mbp for the 3q27–29 interval.

The Mosaic Nature of Mammalian Chromosomes

The mosaic patterns in the organization of various mammalian genomes must, in part, reflect the rearrangements of these genomes over their evolutionary history. The relation between various structural features of Mmu 16, including gene density, G+C content expressed sequence tag (EST) density, SINE density, and LINE density, and how these features relate to the boundaries of regions of conserved synteny with the human genome is shown in Fig. 5. Three of the six syntenic boundaries (those between the regions of conserved synteny between Mmu 16 and Hsa 16 and Hsa 8, Hsa 12 and Hsa 22, and Hsa 22 and Hsa 3q27) show marked discontinuity for several of these features—namely, G+C content, SINE density, and LINE density (Fig. 4). Not only are there sharp discontinuities at these boundaries, but the density of the feature is often more uniform, and divergent from the flanking regions, in the syntenic block than over the local chromosomal neighborhood. This is probably best illustrated by the region of shared synteny between Mmu 16 and Hsa 22. G+C content on Mmu 16 immediately (2 Mbp) before the portion syntenic to Hsa 22 is about 40% and abruptly changes to nearly 50%, which is maintained across the entire 2 Mbp of the Hsa 22 syntenic region, and this drops to below 40% immediately following the Hsa 22 region (Fig. 5). This pattern holds for SINE and LINE density in this region. Here, the mosaic nature of this portion of Mmu 16 relative to human would seem to be explained by the breaking and joining of chromosomal regions, as they existed in an early ancestor, with very disparate properties. The implications of these discontinuities, and their absence at other boundaries (Fig. 5), for the mosaic pattern of mammalian chromosome evolution is discussed below.

Discussion

What are the salient features that emerge from comparison of an initial examination of a mouse chromosome with the regions of conserved synteny in the human genome? The first is that large regions of this chromosome have been remarkably conserved during the more than 100 million years that have elapsed since the lineages leading to humans and mice diverged. One-third of Mmu 16, ~32.8 Mbp, that has conserved synteny with Hsa 16 and Hsa 21, has preserved gene content and gene order, with only two exceptions. The remaining 60 Mbp...
correspond to five other regions of the human genome in six conserved syntenic segments. Another measure of the remarkable conservation between the corresponding regions comes from the analysis of syntenic anchors. The order and orientation of these anchors are highly conserved; of the 11,496 sequence pairs that were identified between Mmu 16 and the corresponding regions of the human genome, only about 1% were not strictly conserved in order and orientation relative to their neighbors.

Of the 731 proteins coding genes that we have identified on Mmu 16, 509 have orthologs in the human genome based not only on their sequence similarity but also on conservation of their map position in regions of shared synteny, as determined by the local consistency of syntenic anchors. Although the assertion of orthology can be problematic (35), this combination of features is currently the strongest evidence available for such an assertion, which is important because it allows one to predict conservation of function with greater confidence, especially in those gene families in which inferences about functional relations between family members in the two species are difficult to interpret owing to large-scale expansion.

We found 14 putative genes on Mmu 16 for which we could find no counterpart in humans. Thus, the percentage of genes that are unique to the mouse lineage, based on what we have found from chromosome 16 (14 of 731 genes), is likely to be ~2%. A similar figure, 2.9% (21/725), is found for genes that are unique to humans, based on the number of genes that are present in humans and absent in mice in the regions of conserved synteny with Mmu 16.

The differences in gene number in any given region of conserved synteny are mainly attributable to differences in the number of paralogous genes in one species compared with the size of the locally expanded family in the other species, although the differential use of exons sometimes produces orthologous proteins that differ by a protein domain and hence have functional consequences (41). At a more detailed level, our results are consistent with, and extend, a number of smaller scale efforts that have compared parts of the mouse and human genomes, such as the 4.5-Mbp region of mouse 16 compared with the 4.5-Mbp region of mouse 16 compared with portions of the mouse and human genomes, only about 4% were not strictly conserved in order and orientation relative to their neighbors.

The synteny regions in the human genome are about 10% larger than those in the mouse, and this figure is the result of a larger proportion of repetitive sequences (SINE and LINE elements) in the human genome. Certain chromosomal regions such as the gene-poor desert in both Mmu 16 and Hsa 21 have largely retained their size since the evolutionary separation of the two lineages. Whether this is due to functional constraints or simply to the slow rate of DNA loss or addition is not known. Previous studies of small-scale sequenced chromosomal regions of similar genetic content between the human and mouse genomes have revealed that, in general, a given human segment is larger than the corresponding mouse segment (15, 42, 49). Reassociation data originally suggested the existence of a fraction of a human dispersed repetitive DNA that was lacking from the mouse genome, (50, 51). A major contributor to this difference is the substantially larger fraction of SINE elements in the human genome as compared with that of the mouse, an observation that others have made for smaller sets of comparative data (42).

The insertion of retroviral sequences is one of the more dynamic processes in genome evolution. This phenomenon has been widely studied (52) and accounts for many of the differences between the mouse and human genomes. The distribution of LINE elements, the general class of retrotransposons, in mammals is radically different between mouse and human (12). Comparative analysis yields the following summary results: 33 retroviral-related sequences were found on Mmu 16. For the corresponding syntenic regions of the human genome, the number was 19. A full review of this topic is beyond the scope of this study.

Only with genome assemblies that are robust and have extensive long-range contiguity can the sorts of analyses reported here be obtained. Examples such as those highlighted here illustrate the power and importance of comparative genomics to clarify the relationships between the genomes of various organisms and to understand where they are sufficiently conserved that they specify similar biology. The availability of sequenced mouse and human genomes holds the promise that we shall be able to more rapidly identify the genes and associated regulatory elements that are critical to any biological phenomenon and to filter and validate those that are relevant to intervention in disease. Although the work presented here represents only a small beginning in analysis of the mouse genome, it is clear that the tools are in place for comprehensive studies of the evolutionary and functional relationship between the mouse and human genomes (53a).

References and Notes
8a. The strains used in this study originate in whole or in part from mice obtained from A. Lathrop, who maintained a mouse farm in Granby, MA (53b). Her mice were used for progenitors for many of the inbred strains that are in use today. The founders of the A/J strain arose from a cross of the Cold Spring Harbor albino stock and the Bagg albino stock in 1921 by Strong (53b). This strain is homozygous for mutant alleles at the agouti [aj; nonagouti (blac)] allele of the agouti gene.
13a. Whitehead genetic and RH STS markers were searched by ePCR and BLAST. A modified version of ePCR was used, allowing one gap and/or one mismatch 5’ of the 7th bp from the 3’ end. The full sequence of the STSs and their primers were not found by ePCR was checked against scaffolds by BLAST, at a 90% identity and 90% length match cutoff. The best matching primer or sequence location was used. The expectation value cutoff of 1 × 10^-10. The matches were further filtered by requiring that at least 70 bp of the alignment, or the length of the alignment if <70 bp, did not overlap in either the human or mouse sequence. Because of the high similarity of the final anchors are uniquely apparent in both genomes. Moreover,most anchors probably reflect significant conser-
vative substitutions. But only the similarity is not always a sign of conservative evolutionary pressures.

25. The sex chromosomes are underrepresented in a whole-genome shotgun assembly and hence are not appropriate substrates for BLASTX.

26. A run of consistent anchors is considered to be inverted if, by inverting that run of anchors on one genome, it could be merged with another run to one side of the other. For instance, if a set of anchors occurs as ABCDE in human and as D’C’E’F’ in mouse, where D’ and C’ are in reverse orientation, then the CD’O’C” is inverted. A run of consistent anchors is considered to be locally transposed if exchanging it with another run to one side of the other. For instance, the first anchors in one genome would allow the two runs to be merged. For instance, if a set of anchors occurs as ABCD in human but as CDAB in mouse, then exchanging CD and AB in the mouse would result in all four anchors belonging to the same run.

31. Z. Deng et al., Genomics 18, 156 (1999).


33. Three was picked empirically.

34. MUMmer was run with a minimum word match length of 11 amino acids. At least three distinct protein matches were required, within a 15 “protein call” span, in order to be recognized as a candidate syntenic block. The entire mouse proteome was compared with the human proteome by MUMmer in 3 h of computation. In addition, all predicted human proteins were compared with predicted mouse proteins with BLASTP. Proteins were considered orthologous only if they were among the top five BLASTP target genes and had a BLASTP expectation value of <1 × 10^-10. Proteins scored as orthologs were then clustered into syntenic blocks in the same manner as for the MUMmer matches. Results derived from the two comparisons were merged in order to increase the total number of orthologous proteins in the syntenic blocks. By using more sensitive MUMmer parameters, we identified additional syntenic blocks, corresponding to ancient human intragenomic duplications reported in (1).
Explaining the Latitudinal Distribution of Sunspots with Deep Meridional Flow

Dibyendu Nandy and Arnab Rai Choudhuri

Sunspots, dark magnetic regions occurring at low latitudes on the Sun's surface, are tracers of the magnetic field generated by the dynamo mechanism. Recent solar dynamo models, which use the helioseismically determined solar rotation, indicate that sunspots should form at high latitudes, contrary to observations. We present a dynamo model with the correct latitudinal distribution of sunspots and demonstrate that this requires a meridional flow of material that penetrates deeper than hitherto believed, into the stable layers below the convection zone. Such a deep meridional flow may have important implications for turbulent convection and elemental abundance in the Sun and similar stars.

The 11-year cycle of sunspots was discovered about 150 years ago (1); the magnetic nature of this solar cycle became apparent about 100 years ago (2), and hydromagnetic dynamo theory to explain the origin of this cycle was developed about 50 years ago (3). Still, to this date, we do not have a solar dynamo model that explains different aspects of the solar cycle in detail. An understanding of the complex plasma motions and the conditions that exist within the turbulent solar convection zone (SCZ) is essential for building any model of the solar dynamo (4, 5). Even a decade ago, our knowledge about the solar interior was too limited to develop a realistic dynamo model. The following developments within the past few years have drastically changed the scenario. Helioseismology has mapped the angular velocity distribution ($\Omega(x,0)$) in the solar interior (6, 7) (Fig. 1) and has discovered the tachocline—a region of substantial radial shear ($\partial \Omega/\partial r$) in the rotation at the base of the SCZ. There the strong toroidal magnetic field is produced (as a result of the stretching of poloidal field lines) and then rises radially (as a result of magnetic buoyancy) to form sunspots. Simulations of this buoyant rise have established that the strength of the “sunspot-forming” toroidal field at the base of the SCZ must be on the order of $10^7$ G (8–10). The classical a-effect (3, 5), which involves the twisting of the toroidal field by helical turbulence to produce the poloidal field, cannot affect such a strong field. Therefore, the most likely mechanism for the generation of the poloidal field (the weak “diffuse” field observed on the solar surface outside of sunspots) is the shearing of tilted sunspot pairs on the solar surface (11, 12). This mechanism can be incorporated in kinematic dynamo models by invoking an a-effect concentrated in a thin layer near the solar surface (13–16). The meridional circulation of the Sun carries the weak poloidal field generated at the solar surface (by this a-effect) first poleward and then downward to the tachocline, where it can be stretched to produce the toroidal field (13–19).

Although a meridional flow of material toward the poles is observed in the outer 15% of the Sun (20, 21), nothing much is known about the counterflow toward the equator, except that it must exist to conserve mass. This flow is driven by turbulent stresses in the SCZ, and standard theories of convection suggest a flow that is contained mainly within the SCZ. When kinematic models of the solar dynamo include such a flow (black contours with arrows in Fig. 1), a helioseismically determined rotation profile, and an a-effect concentrated near the solar surface, such models tend to produce strong magnetic fields at higher latitudes (14, 15, 19). A buoyancy algorithm that we have constructed (16, 22), when included in such a model, results in sunspots at polar latitudes (Fig. 2A). The radial shear in the differential rotation $\partial \Omega/\partial r$ within the tachocline, which has opposite signs at higher and lower latitudes (Fig. 1), has larger amplitude at high latitudes. This results in the toroidal field being produced predominantly at high latitudes in the tachocline (from the poloidal field that has been dragged down by the meridional downflow near the poles), giving rise to sunspots there. Even in solar interface dynamo models, a solar-like rotation tends to produce non–solar- like solutions (23).

These results are insensitive to the strength of the a-effect, nor do they change qualitatively if we change the value of the turbulent diffusivity in the SCZ. The situation does not improve if one tries to evade the negative radial shear in the tachocline at high latitudes with a counterflow that is contained within the SCZ at high latitudes and enters the tachocline at low latitudes only. It turns out that the latitudinal shear within the SCZ at high latitudes is strong...