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] Zhenya Li,[1] Gennady V. Merkulov,[1] Natalia V. Milshina,[1] Ashwinikumar K. Naik,[1] Rong Qi,[1] Bixiong Chris Shue,[1] Aihui Wang,[1] Jian Wang,[1] Xin Wang,[1] Xianghe Yan,[1] Jane Ye,[1] Shibu Yooseph,[1] Qi Zhao,[1] Liansheng Zheng,[1] Shiaoping C. Zhu,[1] Kendra Biddick,[1] Randall Bolanos,[1] Arthur L. Delcher,[1] Ian M. Dew,[1] Daniel Fasulo,[1] Michael J. Flanigan,[1] Daniel H. Huson,[1] Saul A. Kravitz,[1] Jason R. Miller,[1] Clark M. Mobarry,[1] Knut Reinert,[1] Karin A. Remington,[1] Qing Zhang,[1] Xiangqun H. Zheng,[1] Deborah R. Nusskern,[1] Zhongwu Lai,[1] Yiding Lei,[1] Wenyan 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] Kerry 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] Kelvin Li,[1] Xiangjun Liu,[1] Alexander A. Levitsky,[1] William H. Majoros,[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] Chinyere Ibegwam,[1] Mena Jalali,[1] David Johns,[1] Leslie Kline,[1] Daniel S. Ma,[1] Steven MacCawley,[1] Anand Magoon,[1] Felecia Mann,[1] David May,[1] Tina C. McIntosh,[1] Somil Mehta,[1] Linda Moy,[1] Mee C. Moy,[1] Brian J. Murphy,[1] Sean D. Murphy,[1] Keith A. Nelson,[1] Zubeda Nuri,[1] Kimberly A. Parker,[1] Alexandre C. Prudhomme,[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 Suh,[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. Windsor,[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 Venter[1]†

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.

The laboratory mouse is an invaluable model for helping us understand human biology and disease. The mouse genome sequence, in combination with the recently reported human genome sequence (*1*, *2*), offers the opportunity to rapidly improve our understanding of the relevance and importance of mouse models of human disease and the molecular bases for the similarities and differences between them and the corresponding human conditions (*3*). In addition, comparison of the complete DNA sequence of mice and humans will provide insights into the organization and evolution of the mammalian genome. To illustrate the utility of our whole-genome shotgun assembly of the mouse genome, we present the detailed architecture of one particular chromosome (Mmu 16) and compare it with the human genome. This comparison illustrates the power of comparative genomics based on nearly complete large-scale DNA sequence information.

The nearly complete sequence of a typical

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]GenetixXpress, 81 Bynya Road, Palm Beach, Sydney, 2108, Australia. [3]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel. [4]Department 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.

*To 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.

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 (Mbp), 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 Mbp 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 and radiation hybrid (RH) maps (*1*, *13a, 13b*). These maps contained a total of 545 unique sequence-tagged site (STS) markers on chromosome 16, of which 510 (93.6%) were found on scaffolds in our mouse assembly, 9 (1.3%) were found on scaffolds that are composed of conserved repeat sequences, 13 were found (2.4%) on unassembled fragments, and 13 (2.4%) were not present in our mouse sequence. Of the 510 STSs found on scaffolds, 498 were on scaffolds that ultimately mapped to chromosome 16 (Tables 1 and 2). The remaining 12 (2.2%) STSs were found on scaffolds where the preponderance of evidence placed them elsewhere in the mouse genome. We excluded the possibility that these scaffolds were chimeric—i.e., containing piece(s) of chromosome 16 and some other chromosome—by analyzing the clone coverage of these "mixed" scaffolds (*1*). 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 interscaffold 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).

The sequence of Mmu 16 has been deposited at DNA Data Bank of Japan/European

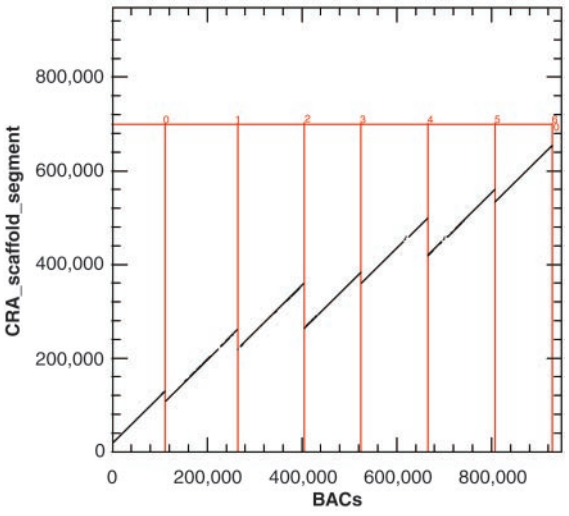


Fig. 1. Dot-plot analysis of a portion of Celera (CRA) scaffold GA_x5J8B7W4YGF and BAC sequences. Each dot represents a match of at least 98% sequence identity and is plotted at 200 bp per pixel. GenBank accession numbers from left to right: AC018559, AC007844, AC012397, AC009192, AC006447, AC006404, 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.

Number of scaffolds	20
Total scaffold size (kbp)	
Mapped (kbp)	92,612
Unmapped (kbp)	
Number of contigs	8,635
Total contig size (kbp)	87,009
N50 scaffold size (kbp)	10,976
N50 contig size (kbp)	16
Number of gaps	8,615
Gap size (kbp)	5,602

Table 2. Scaffold mapping of Mmu 16 by STS markers. Scaffolds were ordered and oriented along the chromosome by analysis of their STS content. At least two STS markers of consistent order were required to consider a scaffold "ordered and oriented." If the order or orientation of a scaffold was ambiguous, usually because of a small number of matching STSs, the scaffold was deemed "bounded and unoriented."

Category	Scaffolds	Contig length (kbp)	Scaffold length (kbp)	STS hits
Ordered and oriented	17	86,987	92,574	495
Ordered and unoriented	2	15	24	2
Bounded and unoriented	1	7	13	1
Total	20	87,009	92,612	498

Molecular Biology Laboratory/GenBank as a whole-genome shotgun project under project accession number AAAD00000000. The version described here is the first version, accession number AAAD01000000. Additional information is available at www.celera.com/mouse16.

Gene Annotation

The methods for computational sequence analysis and automated gene annotation for the mouse genome used the same basic tools described for annotation of the human genome (*1*, *16*). The autoannotation pipeline predicted 1055 genes with high to medium confidence (*17*) on Mmu 16. The gene predictions with weak confidence (unique *ab initio* gene predictions supported by only one type of homology evidence) are not included in this report or in the subsequent analysis. We manually examined the 1055 high- and medium-confidence gene predictions to facilitate comparison of the inventory of genes on Mmu 16 with the homologous genes in the human genome. On closer examination, taking into account genes that might have been split or merged in the autoannotation process, and after discounting pseudogenes and viral-related sequences (such as endogenous retroviruses), we reduced these 1055 predictions to 731 genes. The Refseq database (*18*) reports 130 genes that have been mapped to Mmu 16; of these, 129 correspond to genes we have predicted on chromosome 16, and one of the Refseq genes maps to Mmu 14 in our assembly. Genes on Mmu 16 cover, on average, 23,219 bases of chromosomal DNA, compared with 27,894 bases for genes in the human genome (*1*). This estimate is based on the average span covered by RefSeq transcripts, which represent the highest confidence set of gene predictions. A diagram of the gene content and other physical features of chromosome 16 is shown in Fig. 2.

Identification of Regions of Conserved Synteny

Despite the considerable evolutionary time since the ancestors of human and mouse lineages diverged, we expected that many segments in both genomes would be sufficiently conserved to be identified by similarity search between the two sequences while being substantially unique within each genome. Such segments would permit inference of orthologous regions (similarity by virtue of shared ancestry) between the two genomes. Orthologous mouse and human exons are frequently >80% sequence identical (*19*), and significant and unique matches between human and mouse have been noted in noncoding sequences, presumably as a result of selection pressures (structural RNAs, regulatory regions, and so forth) (*20*, *21*) and perhaps also as a result of chance preservation of the sequence of the common ancestor. Conserved 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 kbp. 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

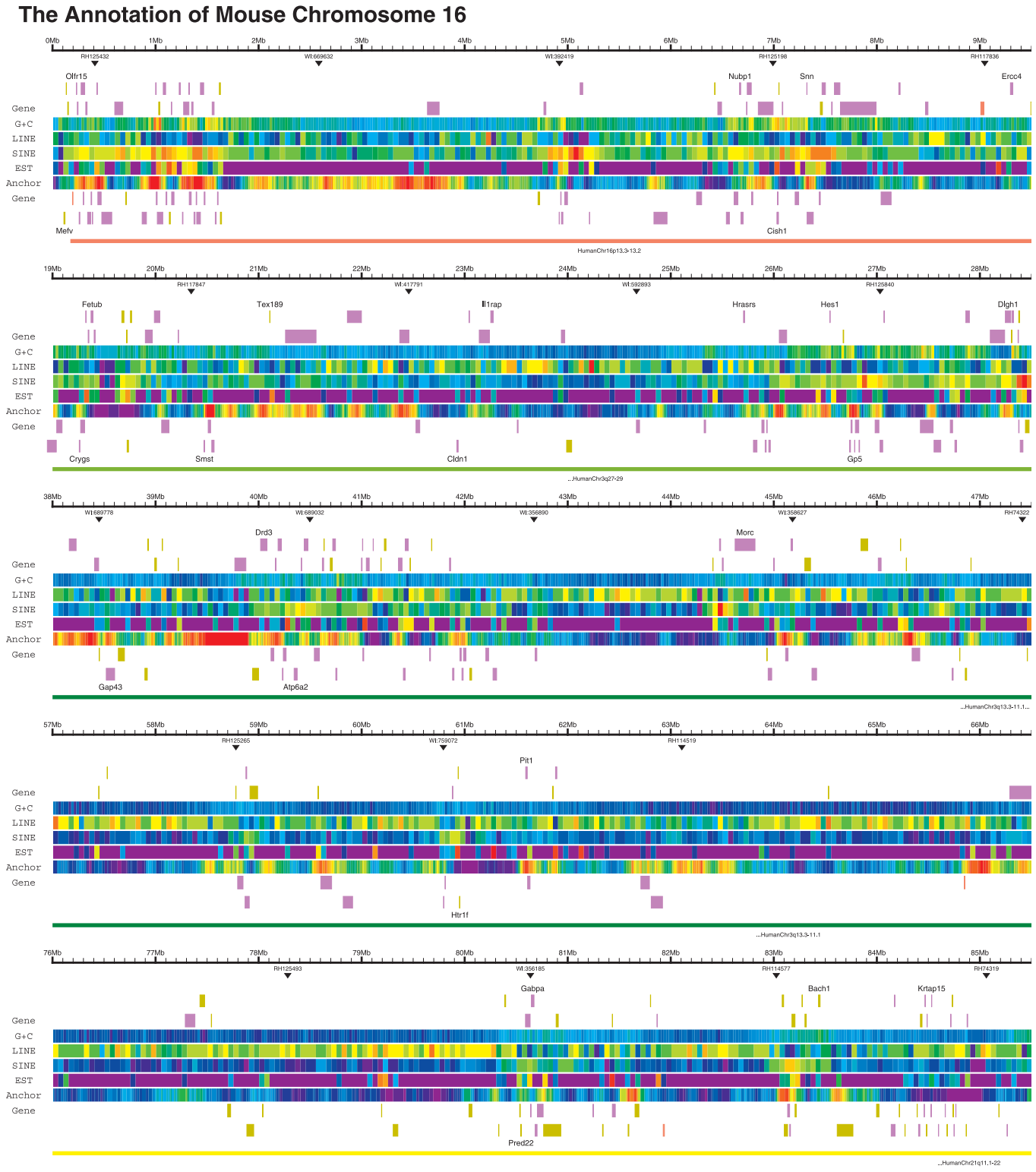
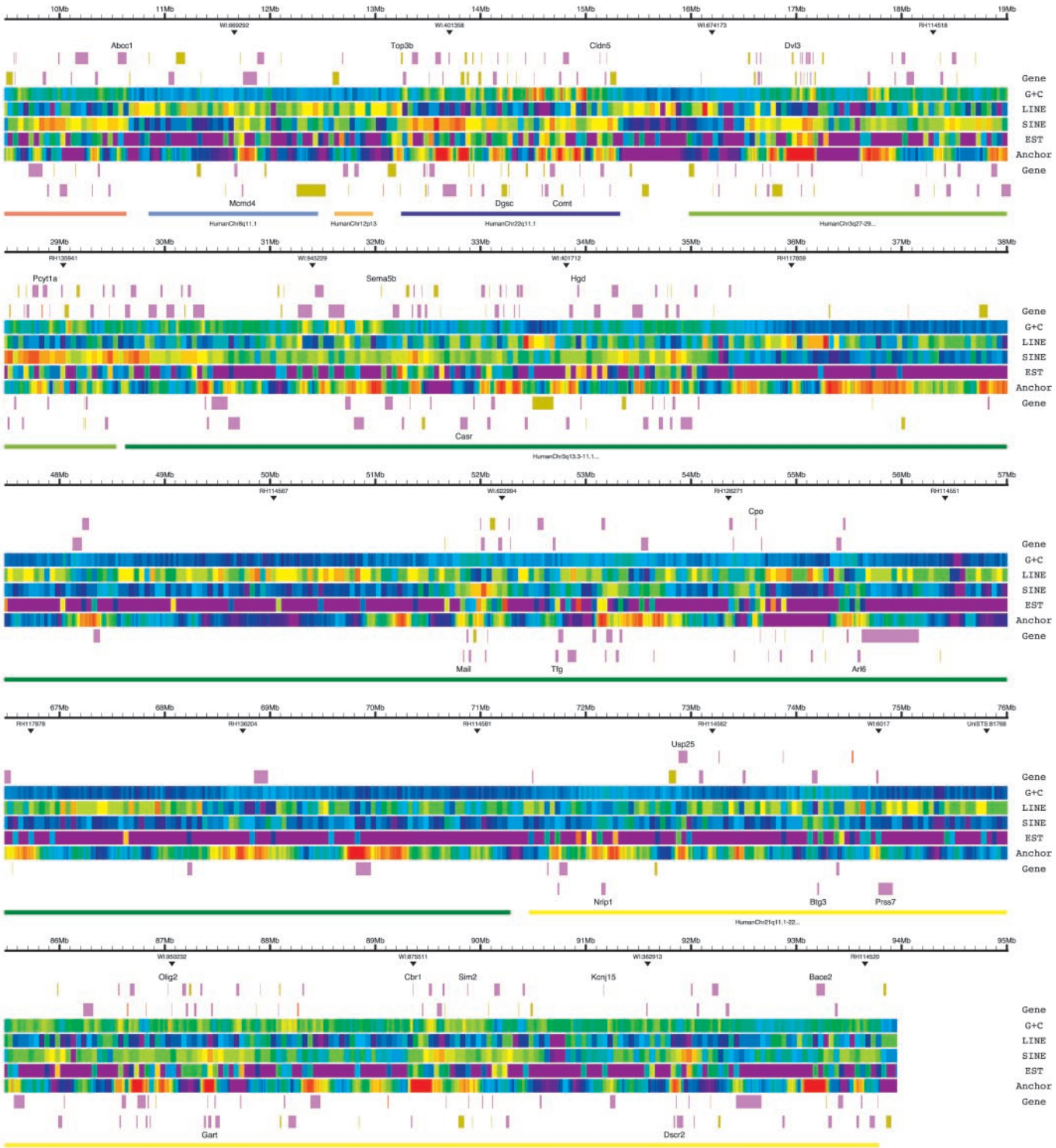
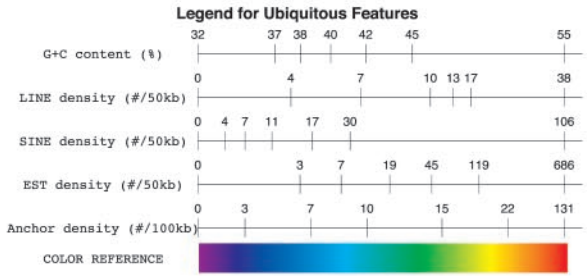
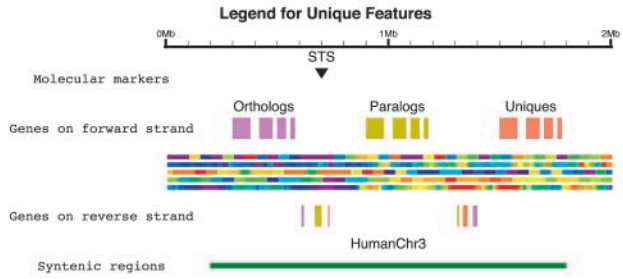


Fig. 2. Annotation of mouse chromosome 16. Celera's assembly and annotation is depicted in five tiers. Each tier contains nine tracks showing (top to bottom) molecular markers, genes mapped to the forward DNA strand, density plots for five ubiquitous features, genes mapped to the reverse DNA strand, 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 sextile (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*).


Abcc1
Top3b
Cldn5
Dvl3
Gene
G+C
LINE
SINE
EST
Anchor
Gene
Mcmd4
Dgsc
Comt
Pcyt1a
Sema5b
Hgd
Casr
Cpo
Mail
Tfg
Arl6
Usp25
Nrip1
Btg3
Prss7
Olig2
Cbr1
Sim2
Kcnj15
Bace2
Gart
Dscr2


Legend for Ubiquitous Features
G+C content (%)
LINE density (#/50kb)
SINE density (#/50kb)
EST density (#/50kb)
Anchor density (#/100kb)
COLOR REFERENCE


Legend for Unique Features
STS
Molecular markers
Orthologs
Paralogs
Uniques
Genes on forward strand
Genes on reverse strand
HumanChr3
Syntenic regions

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 euchromatic 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.

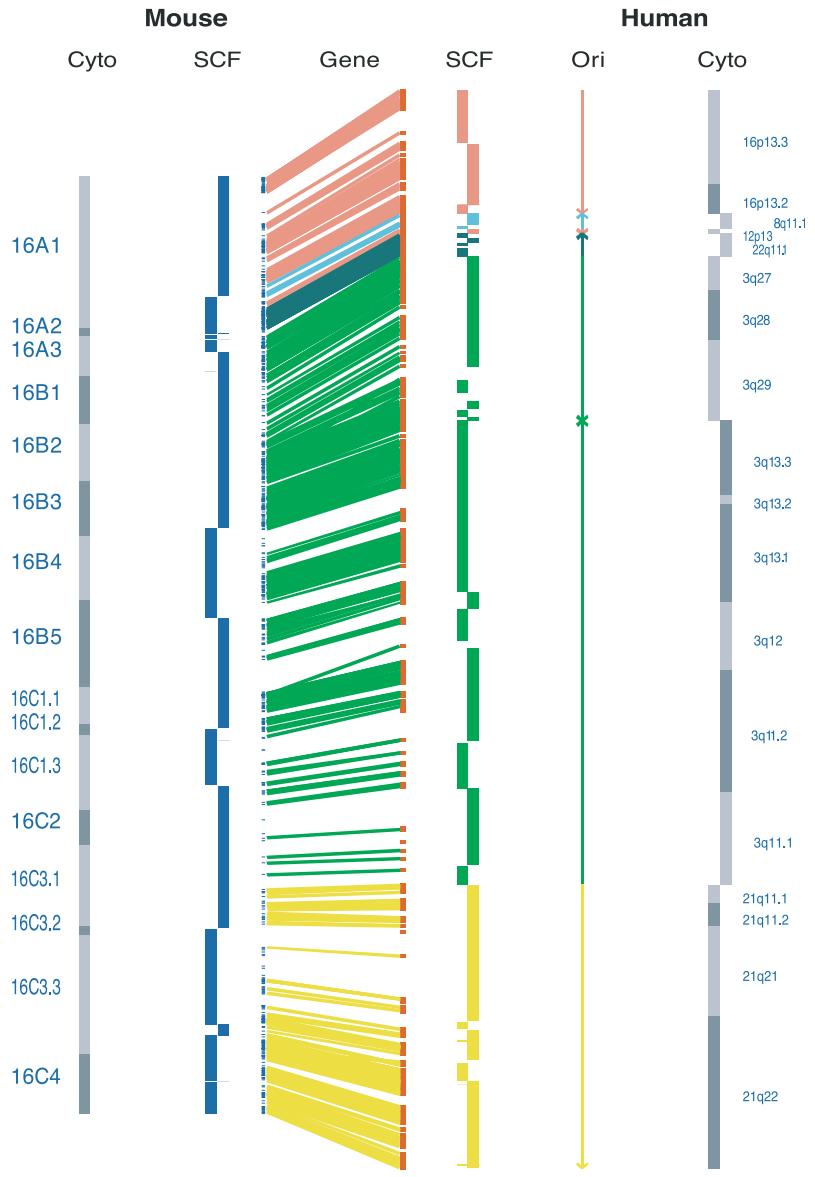


Fig. 3. Regions of conserved synteny between Mmu 16 and the human genome. The analysis was done at the protein level with the MUMmmer program; each line represents a pair of orthologous genes present in the mouse and human. Cyto, cytogenetic markers; SCF, scaffold distribution; and Ori, orientation of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

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 functionally 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 Mbp of the mouse chromosome and 28.42 Mbp 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 syntenic region is not uniform. For example, there is a region on Hsa 21 with only 7 genes in 7.8 Mbp from PRSS7 (an enterokinase) to APP (the amyloid A4 precursor protein implicated in Alzheimer's disease). The corresponding region in the mouse is slightly smaller, about 6 Mbp, 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 syntenic semidesert have remained largely intact since the divergence of lineages leading to humans and mice. In addition, about 400 conserved syntenic 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 Mbp in length in mouse and 1.4 Mbp in human; it contains the human genes *IFNAR2*, *IL10RB*, *IFNAR1*, *IFN*, *GART*, *RPS5L*, *ATP50*, *SON*, *ITSN*, *KCNE1*, and *DSCRI*, 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 *Tmprss2* (transmembrane protease, serine 2) (*39*). One mouse gene in this region, *Itgb21* (integrin-β2-like), does not have an ortholog in the syntenic region of Hsa 21 between *B3galt5* (β1,3-galactosyltransferase) and *Pcp4* (Purkinje cell protein 4). We confirmed the absence of a detectable ortholog of *Itgb21* and found an additional gene between *Itgb21* and *Pcp4*.

Description of individual syntenic regions: the region of Mmu 16 corresponding to a block of Hsa 16. The region of conserved synteny corresponds to ~10.32 megabases of the Mmu and 12.33 Mbp of Hsa 16p. There are 87 orthologs that extend from near the ortholog of the zinc finger gene *Znf174* to the gene for the multidrug resistance protein-*Abcc1*. Of these 87 orthologous pairs, only one is in a noncolinear position between the two genomes.

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 *EIF4a* 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 that 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 Mbp in mice and 16.46 Mbp 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 Mbp in mice and 46.49 Mbp in humans. The gene

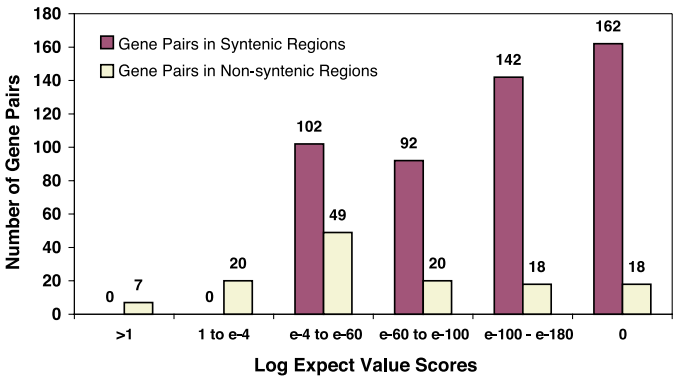


Fig. 4. Distribution of mouse-human orthologs and best hits mapping to syntenic and nonsyntenic regions based on sequence homology and expect value scores. A log expect value score of $<1 \times 10^{-180}$ is represented as 0.

Table 3. Features of regions of synteny shared between Mmu 16 and various regions of the human genome.

Human chromosome	Segment length of region (kbp) (mouse)	Segment length of region (kbp) (human)	Number of anchors	Number of inconsistent anchors (% inconsistent)	Orthologous genes
16	10,461	12,329	1,429	21 (1.5)	87
8	1,284	1,491	121	1 (0.8)	6
12	363	306	31	3 (9.7)	3
22	2,081	2,273	418	8 (1.9)	30
3q27-29	13,557	16,461	1,714	18 (1.0)	107
3q11.1-13.3	41,660	46,493	5,485	63 (1.1)	165
21	22,327	28,421	2,127	27 (1.3)	111

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

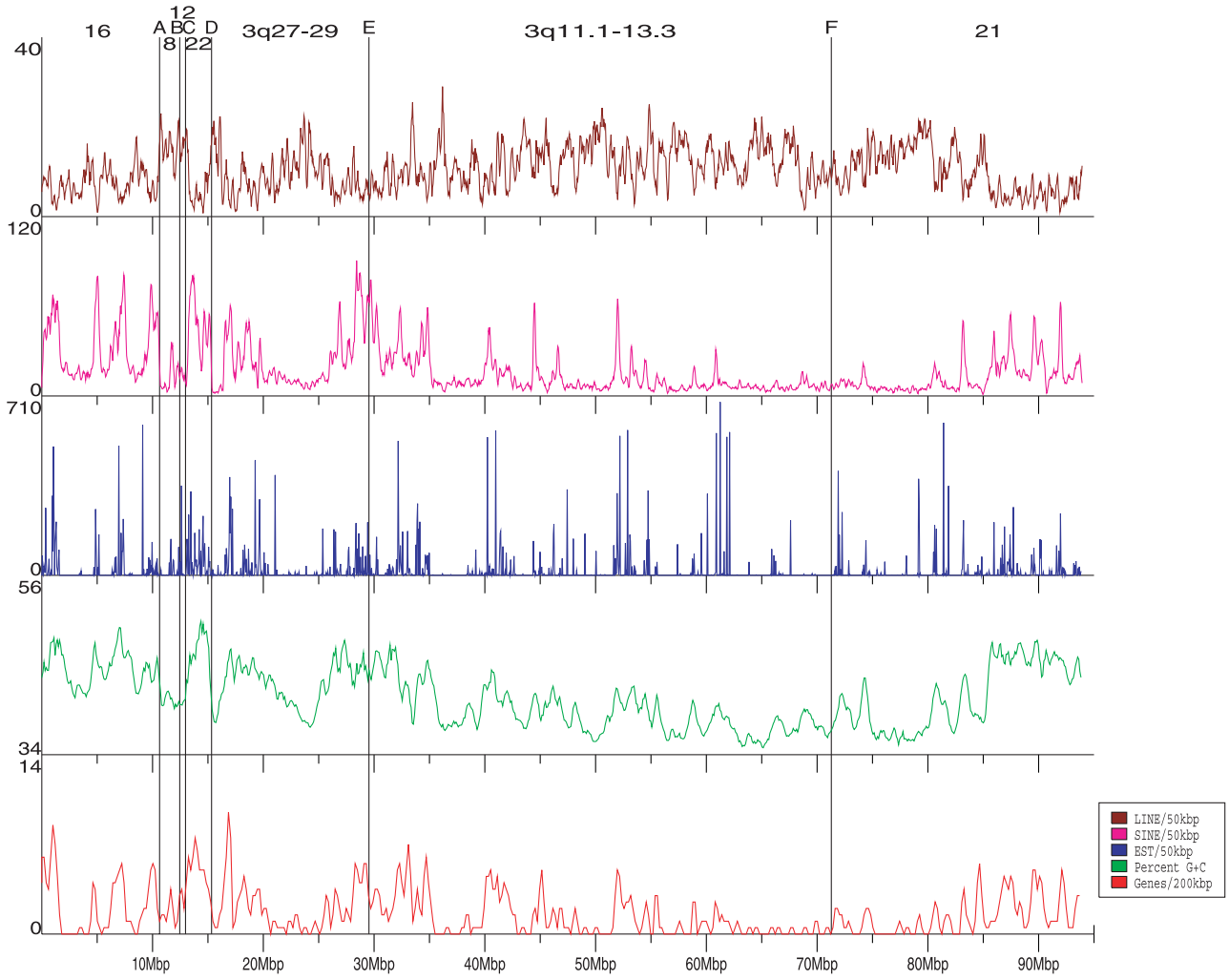


Fig. 5. Correlations of gene density (red), G+C content (green), EST density (blue), SINE density (purple), and LINE density (brown) between Mmu 16 with the regions of conserved synteny in the human genome. The syntenic boundaries are labeled A (Hsa16 and Hsa 8), B (Hsa 8 and Hsa 12), C (Hsa 12 and Hsa 22), D (Hsa 22 and Hsa 3q27–29), E (Hsa 3q27–29 and Hsa 3q11.1–13.3), and F (Hsa 3q11.1–13.3 and Hsa 21).

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 from *Cbr1* to *Tmprss2* (*39*), as well as an analysis in which Hsa 19 was compared with corresponding portions of the mouse genome (*42*).

Numerous authors have noted mosaic patterns in the organization of the mammalian genome (*43*, *44*). The distributions of many features such as gene density, G+C content, and density of repetitive elements are not uniform across the genome. Besides varying greatly among chromosome segments, the distributions of these features are not smooth but are characterized by rather sharp discontinuities. The reasons for these patterns are unclear. Mammalian evolution has been accompanied by rearrangements of the ancestral karyotype, leading to a wide divergence of chromosome number, from 3 pairs in the muntjac deer to 67 pairs in the black rhinoceros (*45*). There can be very extensive rearrangements even between closely related species such as two species of muntjac deer that have 3 and 23 pairs of chromosomes, respectively. The genomic rearrangements in these deer species have been accompanied by minimal effects on morphology, and they can be interbred to yield viable and healthy hybrids (*46*).

The remarkable conservation that we have observed in the regions of shared synteny between portions of Mmu 16 and the human genome suggest that these regions preserve the basic character of the regions as they existed in the early mammalian ancestor of both the murine and primate lineages. There is a large descriptive and empirical literature as well as considerable analytical, theoretical, and computational work that has been done on this problem, all of which is largely consistent with the "random breakage" model first proposed by Nadeau and Taylor (*47*) and is consistent with the most recent analysis (*2*).

Examining the various genomic features at the boundaries between syntenic regions (Fig. 4) provides an improved understanding of the mosaic patterns of the evolution of mammalian chromosomes. Notably, syntenic boundaries that do not show sharp transitions in these various features may provide evidence for conservation of the original (ancestral) pattern in the lineage that lacks such a transition. On Mmu 16, the boundary between the region syntenic to Hsa 3q11.1–13.3 and the region syntenic to Hsa 21 shows no sharp discontinuity for any of these features. Is there any evidence that the configuration seen in Mmu 16 represents an ancestral arrangement whereas the different arrangement in humans, where this region is split between Hsa 3 and Hsa 21, is derived? Chromosome 1 of the bovine genome shares 10 genes with Mmu 16, 5 of which are found on Hsa 3 and 5 of which are found on Hsa 21. This supports the Mmu 16 configuration as being ancestral. Given a monophyletic origin of mammals between 150 and 200 million years ago, it is remarkable that the signature of the ancestral genome structure is still preserved. As we extend our analysis of the mouse and other mammalian genomes, we may hope to reconstruct the ancestral mammalian karyotype, a problem that has been called "original synteny" (*48*).

The syntenic 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 genic 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 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

1. J. C. Venter *et al*., *Science* **291**, 1304 (2001).
2. E. S. Lander *et al*., *Nature* **409**, 860 (2001).
3. D. Gurwitz, A. Weizman, *Drug Discov. Today* **6**, 766 (2001).
4. M. D. Adams *et al*., *Science* **287**, 2185 (2000).
5. E. W. Myers *et al*., *Science* **287**, 2196 (2000).
6. G. M. Rubin, E. B. Lewis, *Science* **287**, 2216 (2000).
7. M. Hattori *et al*., *Nature* **405**, 311 (2000).

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 served as 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 [*a*; nonagouti (black) allele of the agouti gene

a; chromosome 2); brown (*b*; the brown allele of the tyrosine-related protein gene *Trp1*; chromosome 4); and albino (*c*; the albino allele of the tyrosinase gene *Tyr*; chromosome 7)] genes. DNA was prepared from mice that were obtained from the Jackson Laboratory (Bar Harbor, Maine). A/J mice are highly susceptible to cortisone-induced cleft palate; they show a high incidence of spontaneous and carcinogen-induced lung adenomas in multiparous females; they have a defect in macrophage function; and they fail to develop atherosclerotic lesions when fed an atherogenic diet. A/J mice are highly susceptible to lung and mammary cancer and have a moderate susceptibility to lymphomas. DBA was the first strain to be inbred and was established from various coat color stocks by C. C. Little in 1909. In 1929–1930, several DBA sublines were crossed and new strains were created, two of which persist today as DBA/1 and DBA/2 (*53b*). This strain is homozygous for the mutant alleles at the agouti and brown genes, as well as the dilute allele of the myosin Va gene *Myo5a* (chromosome 9). DBA/2J is one of two DBA sublines maintained by the Jackson Laboratory. DNA was prepared from mice obtained from the Jackson Laboratory. DBA/2J mice are resistant to atherosclerotic lesions when fed an atherogenic diet; they show a high frequency of age-dependent hearing loss and glaucomes; they are susceptible to audiogenic seizures; and they are extremely intolerant of alcohol and morphine. DBA/2J mice are highly susceptible to mammary cancer and have a moderate susceptibility to intestinal, liver, and lung cancer. The history of the 129×1/SvJ and 129S1/SvImJ strains is complex and has been recently characterized (*54–56*). The 129 family of strains originated with L. C. Dunn in 1928, and they are derived from the 101 strain. The founders were a combination of coat-color stocks from English fanciers and a chinchilla c^{ch} stock from Castle (*53b*). 129S1/SvImJ was formerly called 129/Sv-p+Tyr+Kitl+/J (Jackson Laboratory repository number JR02448), and it is homozygous for the white belly agouti allele (*Aw*/*Aw*) of the agouti gene. 129×1/SvJ was formerly called 129/SvJ, and it is homozygous for the white belly agouti (*Aw*) mutation as well as for the pink-eyed dilution mutation; it is segregating for the albino (*c*) and the albino-chinchilla (c^{ch}) alleles of the tyrosinase gene. DNA was prepared from 129×1 mice and 129S1 mice that were obtained from The Jackson Laboratory. 129 mice are susceptible to spontaneous testicular teratocarcinomas, and they are moderately susceptible to liver, lung, and mammary cancer as well as lymphomas. C57BL/6J mice are susceptible to diet-induced obesity, type II diabetes, and atherosclerosis. They have low bone density, late-onset heading loss, and a high incidence of microphthalmia. They are resistant to audiogenic seizures and are prone to hydrocephalus, and they show a strong preference for alcohol and morphine. C57BL/6J mice are highly susceptible to pituitary cancer and lymphomas and moderately susceptible to liver cancer.

8b. K. Lindbad-Toh *et al.*, *Genesis* **31**, 137 (2001).

9. H. Morse, in *Origins of Inbred Strains*, H. Morse, Ed. (Academic Press, New York, 1978), pp. 3–21.

10. H. Morse, in *The Mouse in Biomedical Research*, vol. 1, *History, Genetics and Wild Mice*, H. Foster, J. Fox, J. Small, Eds. (Academic Press, New York, 1981), pp. 1–16.

11. Plasmid DNA libraries of three size classes (2, 10, and 50 kbp) were constructed and sequenced as described (*1*). DNA was isolated from spleen (71%), liver (23%), and kidneys (6%) of the following mouse strains obtained from Jackson Laboratories: 129×1/SvJ, DBA/2A, A/J, 129S3/SvImJ. Only tissues from male mice were used. Tissues were homogenized in a solution containing 1.0% SDS, 1× SSC, and 0.25 mg of Pronase E (Sigma) per ml and incubated at 37°C for 30 min. The homogenized tissue was then extracted once with an equal volume of aqueous phenol and two times with chloroform-isoamyl alcohol (24:1). DNA was precipitated with ethanol and dissolved in 1 ml of TE buffer. Plasmid library construction and DNA sequencing were performed as described (*1*). A total of 27.4 million successful sequence reads were collected between May and December 2000, and the assembly presented here was completed in March 2001.

12. R. J. Mural *et al.*, unpublished data.

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 those STSs whose primers were not found by ePCR was searched against scaffolds by BLAST, at a 90% identity and 90% length match cutoff. The best matching primer or sequence location was used for subsequent mapping.

13b. W. J. Van Etten *et al.*, *Nature Genet.* **22**, 384 (1999).

14. BAC end sequences (BES from The Institute for Genomic Research) were mapped onto scaffolds by BLAST. Consistently oriented BES pairs between scaffolds are used to link the scaffolds to form a chain of scaffolds. Internal scaffolds recruited by this method and without mapping STS markers are considered bounded.

15. T. K. Footz *et al.*, *Genome Res.* **11**, 1053 (2001).

16. Scaffolds containing >10 kbp of sequence were analyzed for features of biological importance through a series of computational steps, and the results were stored in a relational database. For scaffolds >1 megabase, the sequence was cut into single megabase pieces before computational analysis. All sequences were masked for complex repeats with Repeatmasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) before gene finding or homology-based analysis. The computational pipeline required ~7 h of central processing unit (CPU) time per megabase, including repeat masking, or a total compute time of about 20,000 CPU hours. Protein searches were performed against the nonredundant protein database available at the National Center for Biotechnology Information (NCBI) and translations of the Celera Human Transcript data set. Nucleotide searches were performed against mouse, rat, and human Celera Gene Indices (assemblies of cDNA and EST sequences); mouse, rat, and human EST data sets parsed from the dbEST database (NCBI); mouse EST sequences from the Riken Mouse Encyclopedia sequence data archive (www.genome.rtc.riken.go.jp/); rat and human mRNA data sets parsed from the RefSeq experimental mRNA database (NCBI); the Celera Human Transcript data set; dog genomic DNA reads generated at Celera (1.5×); and pufferfish genomic DNA reads in the public domain, including *Tetraodon nigroviridis* and *Fugu rubripes* (NCBI). Initial searches were performed on repeat-masked sequence with BLAST 2.0 (*57*) optimized for the Compaq Alpha compute-server and an effective search space size of 3×10^9 for BLASTN searches and 1×10^9 for BLASTX searches. Additional processing of each query-subject pair was performed to improve the alignments. All protein BLAST results having an expectation score of $<1 \times 10^{-4}$, mouse nucleotide BLAST results having an expectation score of $<1 \times 10^{-8}$ with >94% identity, and human and rat nucleotide BLAST results having an expectation score of $<1 \times 10^{-4}$ with >80% identity were then examined on the basis of their high-scoring pair (HSP) coordinates on the scaffold to remove redundant hits, retaining hits that supported possible alternative splicing. For BLASTX searches, analysis was performed separately for selected model organisms (yeast, mouse, human, *Caenorhabditis elegans*, and *D. melanogaster*) so as not to exclude HSPs from these organisms that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the scaffold sequence were then realigned to the genomic sequence with Sim4 for ESTs and mRNAs, and with Genewise for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually give a better representation of intron-exon boundaries than standard BLAST analyses and thus facilitate further annotation (both machine and human). In addition to the homology-based analysis described above, three ab initio gene prediction programs were used (*58–61a*).

17. Corresponds to genes with two to four lines of evidence, as described in Table 8 of (*1*).

18. D. R. Maglott, K. S. Katz, H. Sicotte, K. D. Pruitt, *Nucleic Acids Res.* **28**, 126 (2000).

19. W. Makalowski, M. S. Boguski, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9407 (1998).

20. W. W. Wasserman, M. Palumbo, W. Thompson, J. W. Fickett, C. E. Lawrence, *Nature Genet.* **26**, 225 (2000).

21. K. A. Frazer *et al.*, *Genome Res.* **11**, 1651 (2001).

22. W. Krivan, W. W. Wasserman, *Genome Res.* **11**, 1559 (2001).

23. For the DNA level searches, an initial set of matching segments was obtained by comparing the human and mouse sequences using BLASTn, with standard settings and an 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 with any other match. Thus, all of the final anchors are apparently unique in both genomes. Moreover, most anchors probably reflect significant conservative pressures. Based on the frequency of synonymous substitutions, about 1 in 3 bases would show point substitutions in the absence of selection or insertion or deletion events (*61b*). Under a null hypothesis of point mutation only and fixation resulting from random drift, the expected interval between anchors can be estimated to be >50 kbp, based on a random walk Markov model with probability of a +1 step equal to two-thirds and a one-third chance of a −3 step, given that the minimum alignment score for any anchor was 38 (+1 and −3 are the default match and mismatch costs for BLASTn searches). In contrast, we observed anchors approximately every 9 kbp. Furthermore, given a minimum score of 38 and the same random walk model, <5% of matches would be expected to receive a score of at least 52, whereas 79% of observed anchors achieved this level. Some of the anchors may nonetheless be due to chance preservation of ancestral sequence. This does not mean that such an anchor is invalid (the human and mouse sequences are ultimately orthologous), but only that the similarity is not always a sign of conservative evolutionary pressures.

24. The sex chromosomes are underrepresented in a whole-genome shotgun assembly and hence are not appropriate substrates for this type of analysis.

25. 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 or the other. For instance, if a set of anchors occurs as ABCDEF in human and as ABD′C′EF in mouse, where D′ and C′ are in reverse orientation, then the CD/D′C′ is inverted. A run of consistent anchors is considered to be locally transposed if exchanging it with another run to one side 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.

26. J. Lund *et al.*, *Mamm. Genome* **10**, 438 (1999).

27. J. Lund *et al.*, *Genomics* **63**, 374 (2000).

28. V. Antona *et al.*, *Cytogenet. Cell Genet.* **83**, 90 (1998).

29. A. Lengeling *et al.*, *Genet. Res.* **66**, 175 (1995).

30. R. H. Reeves, E. Rue, J. Yu, F. T. Kao, *Genomics* **49**, 156 (1998).

31. Z. Deng *et al.*, *Genomics* **18**, 156 (1993).

32. A. L. Delcher *et al.*, *Nucleic Acids Res.* **27**, 2369 (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 hits in the target genome and had a BLAST expectation value of $<10^{-10}$. 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 include the maximal 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*).

35. For the purposes of this paper, we assert genes or

proteins to be orthologous only if they are mutual best matches by BLAST and map to regions of conserved synteny, as determined by DNA sequence–based anchors (see text), between the mouse and human genomes. In addition, we include in our analysis only the class of paralogs that are the result of local gene duplication in one or the other species. We are aware that this is a narrow view of these concepts but use it here for the sake of clarity.
36. J. C. Auffray, P. Fontanillas, J. Catalan, J. Britton-Davidian, *J. Hered.* **92**, 23 (2001).
37. R. H. Reeves *et al.*, *Nature Genet.* **11**, 177 (1995).
38. H. Sago *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6256 (1998).
39. M. T. Pletcher, T. Wiltshire, D. E. Cabin, M. Villanueva, R. H. Reeves, *Genomics* **74**, 45 (2001).
40. From the Jackson Laboratory maps, NCBI site maps, Oxford grids, and Annual Human Genome Project Report maps (www.informatics.jax.org/menus/map_menu.shtml, www.ncbi.nlm.nih.gov/Homology/, www.informatics.jax.org/searches/oxfordgrid_form.shtml, and www.ornl.gov/hgmis/research/mapping.html, respectively).
41. J. L. Doyle, U. DeSilva, W. Miller, E. D. Green, *Cytogenet. Cell Genet.* **90**, 285 (2000).
42. P. Dehal *et al.*, *Science* **293**, 104 (2001).
43. B. F. Koop, *Trends Genet.* **11**, 367 (1995).
44. S. Zoubak, O. Clay, G. Bernardi, *Gene* **174**, 95 (1996).
45. S. J. O'Brien *et al.*, *Science* **286**, 458 (1999).
46. B. John, G. L. Miklos, *Eukaryote Genome in Development and Evolution* (Allen & Unwin, Boston, 1988).
47. J. H. Nadeau, B. A. Taylor, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 814 (1984).
48. J. H. Nadeau, D. Sankoff, *Nature Genet.* **15**, 6 (1997).
49. A. M. Mallon *et al.*, *Genome Res.* **10**, 758 (2000).
50. R. J. Britten, E. H. Davidson, *Q. Rev. Biol.* **46**, 111 (1971).
51. P. Soriano, G. Macaya, G. Bernardi, *Eur. J. Biochem.* **115**, 235 (1981).
52. D. J. Griffiths, *Genome Biol.* **2**, 1017 (2001).
53a. A number of papers have been published by users of the Celera mouse data (*62–65*).
53b. M. F. Festing, in *Genetic Variants and Strains of the Laboratory Mouse*, M. Lyon *et al.*, Eds. (Oxford Univ. Press, Oxford, 1996), pp. 1537–1576.
54. E. M. Simpson *et al.*, *Nature Genet.* **16**, 19 (1997).
55. D. W. Threadgill *et al.*, *Mamm. Genome* **8**, 390 (1997).
56. M. F. Festing *et al.*, *Mamm. Genome* **10**, 836 (1999).
57. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
58. E. C. Uberbacher, Y. Xu, R. J. Mural, *Methods Enzymol.* **266**, 259 (1996).
59. C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997).
60. R. J. Mural, *Methods Enzymol.* **303**, 77 (1999).
61a. A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000).
61b. S. Kumar, S. Subramanian, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 803 (2002).
62. I. Rodriguez *et al.* *Nature Neurosci.* **5**, 134 (2002).
63. D. A. Sierra *et al.*, *Genomics* **79**, 177 (2002).
64. X. Zhang, S. Firestein, *Nature Neurosci.* **5**, 124 (2002).
65. J. M. Young *et al.*, *Hum. Mol. Genet.* **11**, 535 (2002).
66. J. F. Abril, R. Guigo, *Bioinformatics* **16**, 743 (2000).
67. We thank additional members of Celera's DNA sequencing, software, and IT groups for their valuable contributions to this work.

20 December 2001; accepted 29 April 2002

REPORTS

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 material 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 $\Omega(r,\theta)$ in the solar interior (*6, 7*) (Fig. 1) and has discovered the tachocline—a region of substantial radial shear ($d\Omega/dr$) 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 10^5 G (*8–10*). The classical α-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 decay of tilted sunspot pairs on the solar surface (*11, 12*). This mechanism can be incorporated in kinematic dynamo models by invoking an α-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 α-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 α-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 $d\Omega/dr$ 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 α-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

Department of Physics, Indian Institute of Science, Bangalore 560012, India. E-mail: dandy@physics.iisc.ernet.in, arnab@physics.iisc.ernet.in