Comparative Genomic Sequence Analysis of the Human Chromosome 21 Down Syndrome Critical Region

Atsushi Toyoda,[1] Hideki Noguchi,[1] Todd D. Taylor,[1] Takehiko Ito,[2] Mathew T. Pletcher,[3] Yoshiyuki Sakaki,[1,4] Roger H. Reeves,[3] and Masahira Hattori[1,5]

[1]*Human Genome Research Group, Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, Japan;* [2]*Mitsubishi Research Institute, 2-3-6, Otemachi, Chiyoda-ku, Tokyo, Japan;* [3]*Department of Physiology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA;* [4]*Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1, Shirokanedai, Minato-ku, Tokyo, Japan*

Comprehensive knowledge of the gene content of human chromosome 21 (HSA21) is essential for understanding the etiology of Down syndrome (DS). Here we report the largest comparison of finished mouse and human sequence to date for a 1.35-Mb region of mouse chromosome 16 (MMU16) that corresponds to human chromosome 21q22.2. This includes a portion of the commonly described "DS critical region," thought to contain a gene or genes whose dosage imbalance contributes to a number of phenotypes associated with DS. We used comparative sequence analysis to construct a DNA feature map of this region that includes all known genes, plus 144 conserved sequences ≥100 bp long that show ≥80% identity between mouse and human but do not match known exons. Twenty of these have matches to expressed sequence tag and cDNA databases, indicating that they may be transcribed sequences from chromosome 21. Eight putative CpG islands are found at conserved positions. Models for two human genes, *DSCR4* and *DSCR8*, are not supported by conserved sequence, and close examination indicates that low-level transcripts from these loci are unlikely to encode proteins. Gene prediction programs give different results when used to analyze the well-conserved regions between mouse and human sequences. Our findings have implications for evolution and for modeling the genetic basis of DS in mice.

[Sequence data described in this paper have been submitted to the DDBJ/GenBank under accession nos. AP003148 through AP003158, and AB066227. Supplemental material is available at http://www.genome.org.]

Down syndrome (DS) is caused by trisomy of human chromosome 21 (HSA21) and occurs in approximately one of 700 newborns (Hassold et al. 1996). DS shows various complex phenotypes, including developmental abnormalities, deficiencies of the immune system, characteristic facial features, mental retardation, and congenital heart disease (Epstein et al. 1991). The range and variability of clinical traits indicate that multiple genes are involved in the pathogenesis of DS.

Identification of a smallest region of overlap in individuals who are trisomic for only part of HSA21 (segmental trisomy 21) and share the same subset of DS features has been used as a basis for defining DS critical regions (DSCRs; Delabar et al. 1993). These regions are posited to contain a gene or genes, dosage imbalance for which contributes to the subset of DS features assigned to the DSCR. Some investigators have indicated that a region as small as 1.6 to 2.5 Mb could contain all of the genes with a dosage imbalance that produces most of the features of DS (Ohira et al. 1996; Dahmane et al. 1998). Thus, analysis of the DSCR is important in identifying genes involved in the pathogenesis of DS. One caveat to the DSCR hypothesis is that there are no known individuals with DS who are trisomic only for this minimal segment. Current research does not exclude the possibility of the involvement of other genes mapped outside the DSCR in a number of DS phenotypes (Korenberg et al. 1994; Sumarsono et al. 1996; Barlow et al. 2001), nor is there evidence currently to link dosage imbalance of a single gene with a specific feature (Reeves et al. 2001).

The minimal human DSCR contains several known genes, including *SIM2*, *HLCS*, *DSCR5*, *TTC3*, *DSCR3*, *DYRK1*, *KCNJ6*, *DSCR4*, *KCNJ15*, and *ERG*. All of these were identified in the initial gene catalog of the 33.5-Mb HSA21 sequence (Hattori et al. 2000). Subsequently, two new genes have been annotated, *DSCR6* (Shibuya et al. 2000) and *DSCR8* (accession no. AF321193). *DSCR8* was formerly designated *DCR1-24.0* (Dahmane et al. 1998). In addition, a number of provisional genes have been described based on several classes of evidence (Hattori et al. 2000). Comparative sequence analysis provides a powerful tool to extend and refine these annotations of mammalian genome sequence.

A sequence-ready, P_1-derived artificial chromosome (PAC)–based physical map spanning 4.5 Mb of distal region of mouse chromosome 16 (MMU16) has been constructed (Pletcher et al. 2001). We report here the finished sequence of

[5]Corresponding author.
E-MAIL hattori@gsc.riken.go.jp; FAX 81-45-503-9170.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.153702.

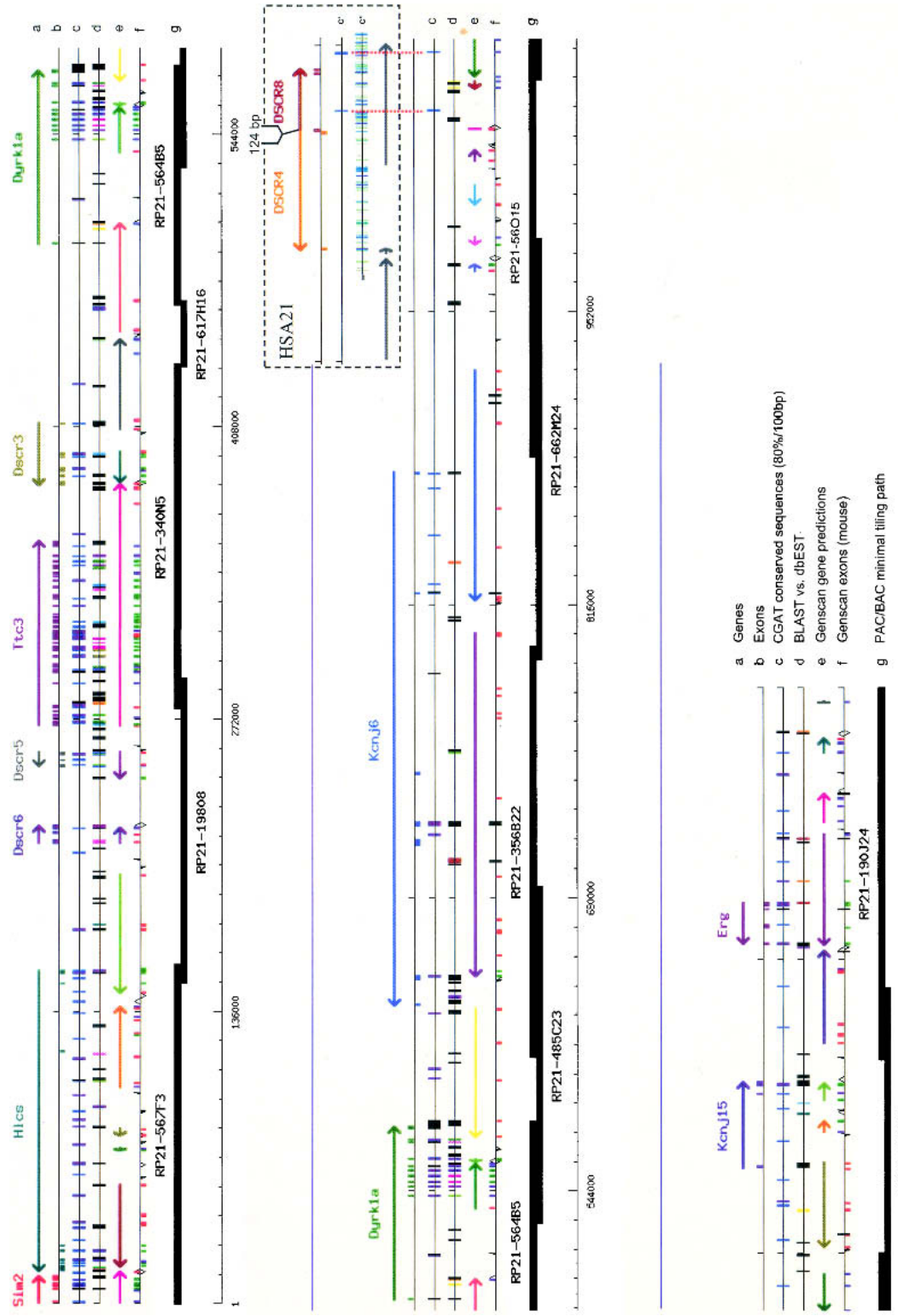

Figure 1 Mouse Down syndrome critical region (DSCR). The CGAT output for 1.3-Mb finished mouse sequence contains 10 known genes (line *a*) for which exons are indicated (line *b*). Conserved sequences detected with the CGAT (>80% identity/100-bp window) are indicated (line *c*), with 80% identity/100 bp shown in light blue and progressively greater identity in darker colors. There is a good, although not comprehensive, correspondence of these sequences with matches from expressed sequence tag and nonredundant databases (line *d*). Mouse GENSCAN-predicted exons and gene models are also shown (lines *e*,*f*). The positions of the *DSCR4* and *DSCR8* genes and their exons in the corresponding human sequence are shown in the "HSA21" box with a stringent (80%/100-bp window, *c*′) and nonstringent (65%/50 bp, *c*′′) in comparison to mouse. The P_1-derived artificial chromosomes (PAC) clones forming the minimal tiling path used for sequencing are presented to scale (line *g*).

11 clones forming a contig of 1.35 Mb, which includes the mouse counterpart of the minimal DSCR segment discussed above. This is the longest comparison of finished human to finished mouse sequence undertaken to date. Comparative sequence analysis shows a high level of conservation overall, including 144 significantly homologous regions in the DSCR which are likely to represent genes or regulatory regions that may contribute to the anomalies of development that characterize DS.

RESULTS

Genomic Sequencing of Mouse PACs

A minimal tiling path of 11 PACs derived from a deep contig covering the region of MMU16 that corresponds to the human DSCR provided the substrate for sequencing (Fig. 1; Pletcher et al. 2001). The clones were isolated from the RPCI-21 female 129/Sv mouse library (Osoegawa et al. 2000). Shotgun sequencing of the PAC clones was performed to provide 8× coverage of draft sequence. In addition, we constructed plasmid clone libraries from appropriate restriction fragments with an insert size from 1 to 10 kb and sequenced both ends of these clones to provide 2× additional coverage. After assembly of all the sequence data using `Phred/Phrap` (Ewing et al. 1998), gap-filling and resequencing of ambiguities in the assembled data were performed by nested deletion (Hattori et al. 1997), primer-walk, polymerase chain reaction (PCR)–coupled primer-walk, and direct sequencing of appropriate PAC clones. We obtained 1,350,235 bp of contiguous finished sequence data. The accuracy was estimated to be >99.99 % through the entire sequence, and each clone was finished according to the rules for human genome sequencing (http://genome.wustl.edu/Overview/finrulesname.php).

Gene Content in Human and Mouse

The end points of the mouse sequence were located in the *Sim2* locus at the proximal side and in the *Erg* locus at the distal side. We searched the mouse genomic sequence against human cDNA sequences in public databases and identified 10 of the 12 known human genes, *SIM2*, *HLCS*, *DSCR6*, *DSCR5*, *TTC3*, *DSCR3*, *DYRK1A*, *KCNJ6*, *KCNJ15*, and *ERG* (Supplemental Table 1, available online at http://www.genome.org). Mouse orthologs of nine of these genes were also found in these databases. The mouse *Hlcs* gene has been mapped, but a sequence of the mouse *Hlcs* transcript was not available. We therefore isolated a cDNA of *Hlcs* from a mouse testis cDNA library by reverse transcription–PCR, determined the sequence, and performed alignment with genomic sequence to determine the exon-intron junctions (accession no. AB066227; see Methods). The mouse *Hlcs* cDNA sequence was 4311 bp and had 78.2% nucleotide identity with the human gene. It encoded a predicted protein of 722 amino acids, with 76.6% amino acid identity to human *HLCS*.

All of the mouse genes shared canonical GT-AG splice junction sequences and the same exon splicing boundaries as their human orthologs. Sequence alignment between mouse cDNA sequences and the finished mouse genomic sequences showed nucleotide differences at various positions. Most of these differences were insertions, deletions, or base substitutions in the 5′ and 3′ untranslated regions (UTRs) of each gene. These sequence differences may represent strain variation between 129/Sv strain mice, which was the substrate for genomic sequencing in this study, and the strains from which the cDNAs originated. Several base substitutions predicted amino acid changes and thus could contribute to functional differences between strains.

Given the well-established conservation between MMU16 and HSA21, we were surprised to find that two human genes, *DSCR4* and *DSCR8*, did not match the mouse genomic sequence using relatively stringent search criteria of 80% match across 100 bp. *DSCR4* and *DSCR8* are adjacent to each other in head-to-head orientation on opposite strands, separated by only 124 bp on HSA21 (Fig. 1). No mouse expressed sequence tags (ESTs) matched the human ESTs or predicted proteins of these genes. No computer-predicted gene candidates occurred at corresponding positions in the mouse genomic sequence. The Celera 5.5× mouse genomic sequence contained no closely related nucleotide or predicted protein sequences. To rule out the possibility that this region was deleted from the mouse PAC clone chosen for sequencing, we sequenced two other bacterial artificial chromosome (BAC) clones that covered this region to 5× coverage, confirming the original sequence assembly. Finally, PCR on mouse genomic DNA produced the fragments predicted from the sequence, showing that the multiple clones were representative of the genomic DNA.

These results indicate that orthologs for *DSCR4* and *DSCR8* might be absent from mice. However, a low-stringency sequence comparison (65% identity/50-bp window) between mouse and human genomic sequence found homology across this region, including the last exons of *DSCR4* and *DSCR8* (Fig. 1, inset). The existence of human genes at this site is based on multiple independent isolations of spliced ESTs, detection of low-level transcripts in Northern blots of placental RNA, and our results of PCR on placental cDNA (data not shown). However, it seems unlikely that these are protein-coding transcripts; the predicted human genes contain repetitive elements as part of their very small putative open reading frames (ORFs), indicating that the gene models may not be correct.

The possibility remains that *DSCR4* and *DSCR8* represent rapidly diverging mammalian genes with a function that is not dictated by protein coding sequences. There are additional candidate genes of this type in the region. The HSA21 transcript known variously as DCR1-17.0, EUROIMAGE 43103, and *PRED69* is found just distal to the 5′ end of *KCNJ6*. It is represented only by unspliced human ESTs and therefore is considered to be less convincing than if it were spliced. No match to mouse genomic sequence is seen under conditions of 80% identity over 100 bp, but a 65%/50-bp survey detects homology across the area (data not shown).

Alignment of publicly available cDNA and EST data with genomic sequence allowed us to define intra- and intergenic regions (Table 1). A "gene region" corresponds to the transcribed sequence, defined by the most 5′ and 3′ ends of available cDNA sequences. Gene regions occupied 1167 kb (65%) of the human sequence and 755 kb (56%) of the mouse sequence. When repetitive sequences are removed, 746 kb (71%) of the human and 614 kb (63%) of the mouse represent gene regions. Discounting the problematic *DSCR4* and *DSCR8* gene models, unique sequence comprising the intragenic regions is essentially the same in both species.

Landscape of Genomic Structure

The GC-contents of mouse and human were 45.35% and 42.4%, respectively, in good agreement with results for the whole genome in each case (Hattori et al. 2000; Mallon et al.

Table 1. Summary of Gene and Integenic Regions in the Human DSCR and the Corresponding Mouse Region

	Gene size (kb)[a]				Intergenic region size (kb)[a]			
Gene	human		mouse		human		mouse	
SIM2/Sim2	19.6	(14.4)	12.8	(11.9)				
					1.3	(1.3)	2.7	(2.5)
HLCS/Hlcs	211.0	(124.7)	139.4	(113.8)				
					44.4	(18.4)	59.5	(31.4)
DSCR6/Dscr6	13.1	(7.8)	8.1	(7.0)				
					45.7	(16.4)	27.5	(18.0)
DSCR5/Dcrc	7.7	(4.2)	6.3	(4.7)				
					12.7	(4.7)	13.2	(10.2)
TTC3/Ttc3	117.3	(73.5)	85.3	(71.3)				
					20.3	(8.6)	26.5	(17.0)
DSCR3/Dcra	44.1	(29.7)	28.4	(20.2)				
					100.0	(35.7)	84.1	(62.0)
DYRK1A/Dyrk1a	147.6	(93.1)	80.3	(65.4)				
					109.3	(54.4)	68.9	(50.1)
KCNJ6/Kcnj6	291.9	(184.6)	234.3	(182.5)				
					137.6	(71.7)		
DSCR4	67.1	(37.5)	—	—				
					0.1	0	251.0	(181.2)
DSCR8	35.0	(20.0)	—	—				
					73.3	(38.9)		
KCNJ15/Kcnj15	70.9	(52.3)	39.4	(35.1)				
					81.0	(48.9)	61.9	(49.5)
ERG/Erg[b]	141.8	(103.8)	120.9	(102.0)				
Total Size	1167.1	(745.5)	755.2	(613.9)	625.7	(299.0)	595.3)	(421.9)

[a]Numbers in parentheses indicate size of unique sequences that was calculated by subtracting all the repetitive sequences from entire sequences.
[b]Only parts of the gene regions are included.

2000; Lander et al. 2001). CpG islands were predicted by searching for sequence segments that are >250 bp with a GC content of >50% and an expected/observed CpG count of >0.6. We identified 13 possible CpG islands in human and eight in the mouse (Table 2). All eight mouse sequences had counterparts at the corresponding positions on HSA21. Six of the eight conserved CpG islands were located just upstream of or surrounding the first exons of *Hlcs/HLCS*, *Dscr6/DSCR6*, *Dscrc/DSCR5*, *Dcra/DSCR3*, *DYRK1A/Dyrk1a*, and *Kcnj6/KCNJ6* genes, indicating an association with regulation of expression of these genes. The two remaining conserved CpG islands were located at the 3′ UTRs of *Sim2/SIM2* gene and the intergenic region between *Hlcs/HLCS* and *Dscr6/DSCR6*; another predicted CpG sequence occurs between *HLCS* and *DSCR6* on HSA21 but not on MMU16. The remaining CpG-rich human sequences were located around 5′ and 3′ UTRs of *DSCR3*, in the intergenic region between *DYRK1A* and *KCNJ6*, and in the third intron of *KCNJ6* (accession no. D873279). It will be of interest to see if the nonconserved CpG sequences lead to a different pattern of transcription in humans than is seen in mouse.

Repetitive elements were identified using `RepeatMasker` (Table 3). The fractions of short interspersed elements (SINEs) in the human and mouse sequences were 13% and 8.5%, respectively, whereas the fractions of long interspersed elements (LINEs) were 16% and 5.3%. The higher ratio of SINE to LINE sequences in mouse contributes to the higher GC content. Repetitive elements belonging to viral long terminal repeats (LTRs) were found in both sequences to a similar degree, but elements such as MER1 were present at threefold higher frequency in human DSCR. Simple repeats such as CA dinucleotides were fourfold more abundant in the mouse

Table 2. Distribution of CpG Islands

Island	Start	End
Human		
cpg1	17,223	18,171
cpg2	235,811	236,724
cpg3	250,284	250,701
cpg4	259,444	260,296
cpg5	275,369	276,868
cpg6	342,297	343,933
cpg7	490,424	491,110
cpg8	527,489	527,943
cpg9	536,672	537,927
cpg10	635,663	638,003
cpg11	833,291	833,982
cpg12	945,063	945,364
cpg13	1,185,383	1,186,703
Mouse		
cpg1	10,903	11,644
cpg2	176,128	176,546
cpg3	199,152	199,672
cpg4	214,020	214,470
cpg5	255,699	257,061
cpg6	409,302	410,318
cpg7	452,020	453,003
cpg8	876,821	877,242

Human sequence corresponds to the region from 23,679,001 to 25,472,000 on NT011512.3 Both human and mouse sequences used in this study are available as supplements from our Web site (http://hgp.gsc.riken.go.jp).

Table 3. Distribution of Repetitive Elements in the Human DSCR and the Corresponding Mouse Region

	Human (1,792,995 bp) GC level 42.43%		Mouse (1,350,235 bp) GC level 45.35%	
	number of elements	percentage of sequence	number of elements	percentage of sequence
SINEs	982	12.97	799	8.53
Alus	719	10.81		
B1s			273	2.36
B2–B4			465	5.75
IDs			26	0.13
MIRs	263	2.16	35	0.30
LINEs	485	16.07	197	5.35
LINE1	316	13.34	176	5.13
LINE2	159	2.60	20	0.22
L3/CR1	10	0.13	1	0.00
LTR elements	327	7.83	353	8.14
MaLRs	198	4.19	269	5.46
ERVL	58	1.19	15	0.27
ERV_classI	54	2.10	7	0.10
ERV_classII	3	0.11	13	1.04
DNA elements	224	3.16	67	0.81
MER1_type	145	1.99	40	0.45
MER2_type	33	0.78	22	0.32
Unclassified	1	0.03	8	0.18
Total interspersed repeats		40.07		23.00
Small RNA	7	0.05	11	0.07
Simple repeats	243	0.98	694	3.96
Low complexity	200	0.57	156	0.80
Total		41.63		27.79

sequence. Overall, repetitive sequences represented 41.6% and 27.8% of the total sequence in the human and mouse, respectively. Most of this difference owes to the threefold higher frequency of LINE elements in human sequence. The higher frequency of repetitive elements in human contributes much of the size difference between the 1.35-Mb mouse sequence and the corresponding 1.8-Mb human DSCR (Table 1).

A dot-plot analysis of the two sequences (80%/100 bp) showed colinearity, with no changes in order or orientation (data not shown). About 10 kb was not seen in mouse relative to human in a region between *DSCR8* and *KCNJ15*. This region contained no predicted genes. Overall, the mouse and human sequences appear to have been well-conserved during evolution.

A BLAST search against public EST databases identified a significantly homologous sequence in the intergenic region between *DSCR4* and *KCNJ15* in human. The homologous sequence showed 90% identity over 600 bp with FLJ21347 mRNA (2589 bp; accession. no. NM_022827), which maps to human chromosome 17. The sequence identified here was shown to have various mutations disrupting the open reading frame (ORF), indicating that it is a pseudogene of FLJ21347.

"Conserved Features" Sequence Map

Mouse and human sequences were compared using the VISTA and CGAT programs (Fig. 1; Lund et al. 2000; Mayor et al. 2000). We adopted various conditions to search for significant sequence homology (conserved segments [CSs]) between the sequences (summarized in Table 4; for details, see Supplemental Fig. 1 and Supplemental Table 2, available at http://www.genome.org). We first identified 123 exons in the human sequence by merging and matching all cDNA and EST data in public databases with our genomic data and by defining all discrete matches as exons. The total length of exons defined by this method was 45,147 bp, or ~4.3%, of human nonrepetitive and 2.5% of total sequence.

We next searched for CSs under several conditions using the VISTA program (Table 4). A survey at a stringency of ≥75% nucleotide identity over 50 bp detected 110 of the known 123 exons. Under these conditions, most UTRs were not detected. When the sequences were compared at ≥80% identity over 100 bp, 254 CSs were detected, representing 53,428 bp, or 5.1%, of the nonrepetitive sequence. This analysis detected 90 of the 123 known exons in 110 CSs. The remaining 144 CSs, totaling 27.7 kb (2.6% of human nonrepetitive/1.5% of total sequence) did not correspond to known exons. Even at the highest stringency, CSs belonging to exons represented only about half of the total CSs detected between human and mouse (Table 4). Similar distribution patterns of CSs between exons and noncoding regions have been reported previously (Loots et al. 2000; Onyango et al. 2000; Frazer et al. 2001).

We further examined the distribution of 144 CSs from noncoding regions that were detected under the condition of ≥80% identity over 100 bp. Two thirds of these (102 CSs) were located in gene regions and one third (42 CSs) were in intergenic regions. Only five of the eight putative CpG islands found at corresponding positions in mouse and human (cpg1, cpg5, cpg6, cpg9, and cpg13) were detected as CSs under these conditions (Supplemental Table 2), showing an advantage of our multitiered search strategy. A total of 20 of the 144 CSs showed significant matches to anonymous ESTs (Table 5). Two of these 20 sequences matched three or more anonymous ESTs, strengthening the likelihood that they are bona

Table 4. Conserved Sequences (CSs) between the Human DSCR and the Corresponding Mouse Region

	Total No. of CSs	No. of CSs matched with known exons[a]	No. of exons matched with CSs (out of 123 known exons)
50 bp			
75% ID	1058 (113,607 bp)	178 (31,159 bp)	110
80% ID	674 (74,892 bp)	161 (26,751 bp)	109
85% ID	341 (39,592 bp)	132 (19,574 bp)	89
75 bp			
75% ID	526 (88,088 bp)	138 (30,983 bp)	104
80% ID	353 (59,173 bp)	123 (26,132 bp)	95
85% ID	217 (35,794 bp)	104 (19,529 bp)	84
100 bp			
75% ID	408 (86,136 bp)	128 (32,762 bp)	103
80% ID	254 (53,428 bp)	110 (25,773 bp)	90
85% ID	156 (31,809 bp)	84 (18,624 bp)	73

Numbers in parentheses indicate total nucleotide length of CSs. The data was obtained from the analysis using the VISTA program.
[a]One exon matched with more than two non-overlapping CSs.

fide transcripts. None of the 144 CSs contained a large ORF on either strand, and none showed similarity to protein sequences in public databases.

DISCUSSION

We describe the finished sequence and comparative analysis of a region of conserved synteny between MMU16 and the minimal human DSCR. The mouse sequence spans 1.35 Mb from *Sim2* to *Erg* and is substantially shorter than the corresponding 1.8 Mb of HSA21. This difference was mainly accounted for by a lower fraction of repetitive elements in the mouse. Human unique sequence represented 1.04 Mb, versus 0.97 Mb in the mouse, consistent with the results in several other studies (Martindale et al. 2000; Pletcher et al. 2000). The relative paucity of repeats in the mouse genome might be explained in part by reduced transposition activity of repetitive elements in rodents (Casavant et al. 2000).

Recombinational and physical mapping and alignment of draft mouse sequence with human sequence have shown a high degree of conservation between most of HSA21 and distal MMU16 (Pletcher et al. 2001). Previously, a single discrepancy in gene content across the extended MMU16/HSA21 region of conserved synteny had been shown. The *Itgb2l* gene on MMU16 has no human counterpart (Pletcher et al. 2001). Here, comparative analysis of finished sequence not only confirmed the overall conservation of gene order and content between this human DSCR and the corresponding mouse genomic region but also revealed previously annotated genes in human without clear counterparts on MMU16.

DSCR4 and *DSCR8* are categorized as genes based on the existence of spliced ESTs. The human genes have no paralogs

Table 5. New Chromosome 21 Transcripts Predicted from Human-Mouse Conserved Segments

	Human		Mouse				
	start	end	start	end	Length (bp)	% ID	Anonymous Expressed sequence tags[a]
1	19984	20144	14285	14441	161	84.5	T77624 (93, 93)
2	56011	56375	35235	35603	371	84.9	BF732521 (325,100), BE083161 (105, 99)
3	73238	73286	48886	49033	152	77.0	AA504947 (151,98), BF653739 (90, 90)
4	83323	83422	59940	60038	100	80.0	BG221805 (100, 100) BG221804 (94, 98)
5	87302	88266	64856	65812	970	85.3	BF195549 (335, 99)
6	114456	114802	82051	82391	347	81.6	AI417767 (347, 100)
7	140660	141040	103528	103911	384	84.1	BF662074 (224, 90), BF452804 (224, 90), AW610715 (224, 90)
8	188157	188326	136333	136497	174	78.2	AA372603 (145, 98)
9	421157	421256	312534	312633	100	80.0	H83304 (93, 100)
10	638015	638243	453023	453244	229	79.5	U81194 (81, 100)
11	888135	888294	636628	636780	161	75.2	BF522889 (58, 87)
12	892104	892283	640552	640723	184	79.3	AI839701 (106, 86)
13	1037840	1037941	764670	764768	102	80.4	AA580680 (78, 100)
14	1273256	1273370	946335	946448	115	79.1	R82099 (52, 96)
15	1526650	1526749	1128830	1128931	102	80.4	AA579369 (100, 100), AA579262 (100, 94), AW227088 (67, 88)
16	1559661	1559829	1155407	1155574	169	78.1	BE145433 (170, 98), BE145419 (170, 95)
17	1655061	1655306	1234095	1234342	249	79.5	AI969718 (220, 100)
18	1689699	1689830	1258685	1258820	136	80.1	BF760340 (132, 100), BI018151 (132, 99)
19	1745806	1746161	1309799	1310148	356	86.2	BE185421 (228, 99)
20	1773435	1773621	1335183	1335368	188	78.7	AL545844 (187, 94)

[a]GenBank accession nos. are indicated (match length, % identity).

in genomic or cDNA databases, nor are there related sequences in the mouse genome, based on the absence of significant sequence matches to mouse genomic sequence in public or private (Celera Genomics) databases or corresponding mouse cDNAs or ESTs. The possibility that cloning or sequencing artifacts led to an apparent absence of these genes from mouse was definitively eliminated. However, the predicted ORFs of both *DSCR4* and *DSCR8* are composed substantially of repetitive DNA sequences. An entire coding exon of *DSCR4* is an LTR-related repeat, and an Alu repeat provides a significant part of the ORF in *DSCR8*, raising questions about the models for these predicted genes. Further, although almost no conservation was detectable in the mouse region corresponding to the human *DSCR4/DSCR8* region under conditions routinely used to identify most exons of known genes, a low-stringency sequence comparison (65% identity/50 bp) to identify the boundaries of the "missing" mouse region revealed conservation across the entire segment, including the last exons of both *DSCR4* and *DSCR8*. Similarly, homology with the HSA21 segment containing *PRED69* was detectable at low stringency. The availability of sequence from a third species could indicate whether these are conserved transcribed sequences or merely regions where less divergence has occurred by chance between mouse and human.

This analysis points out some limitations of relying on cDNA, and especially EST databases, as a standard for identifying transcripts of "real" genes, and it indicates that the existence of genes cannot be excluded even when they are not strongly conserved between human and mouse. A case in point is the *XIST* gene. Despite its highly conserved function in X-inactivation, there is substantial divergence between mouse and human genes for this noncoding transcript. Further analysis of these marginal sequence matches could be instrumental in finding the next level of conserved function encoded by mammalian genomes.

We compared the gene predictions made by algorithms in mouse and human sequences. In contrast to the sensitivity and specificity of comparative sequence analysis, algorithmic predictions of the corresponding sequences returned very different results (Fig. 2). The first 277 kb of unique sequence from mouse and human was used as a test set for GENSCAN gene prediction. The test region contains four documented genes, plus parts of two others, and includes 72 exons. GENSCAN predicted 11 gene models in both sequences, but few of the human models corresponded closely to those in mouse. In neither case was there a strong correspondence to the known genes in the region. BLAST2 analysis identified 84 CSs between the two sequences. Thirty-four of these CSs overlapped a corresponding GENSCAN prediction in mouse and human sequences, and 33 of these 34 shared predicted exons represented previously identified transcribed sequences. Thus, comparative gene prediction, looking for commonalities of gene predictions made in the related sequences of two species, showed a low false-positive rate, 3% (1/34), but identified only a small percentage of conserved transcribed exons. The majority of GENSCAN predicted exons are not supported by sequence conservation and lead to very different gene models, despite the similarity in gene content in these conserved regions.

TWINSCAN is a gene annotation tool that incorporates

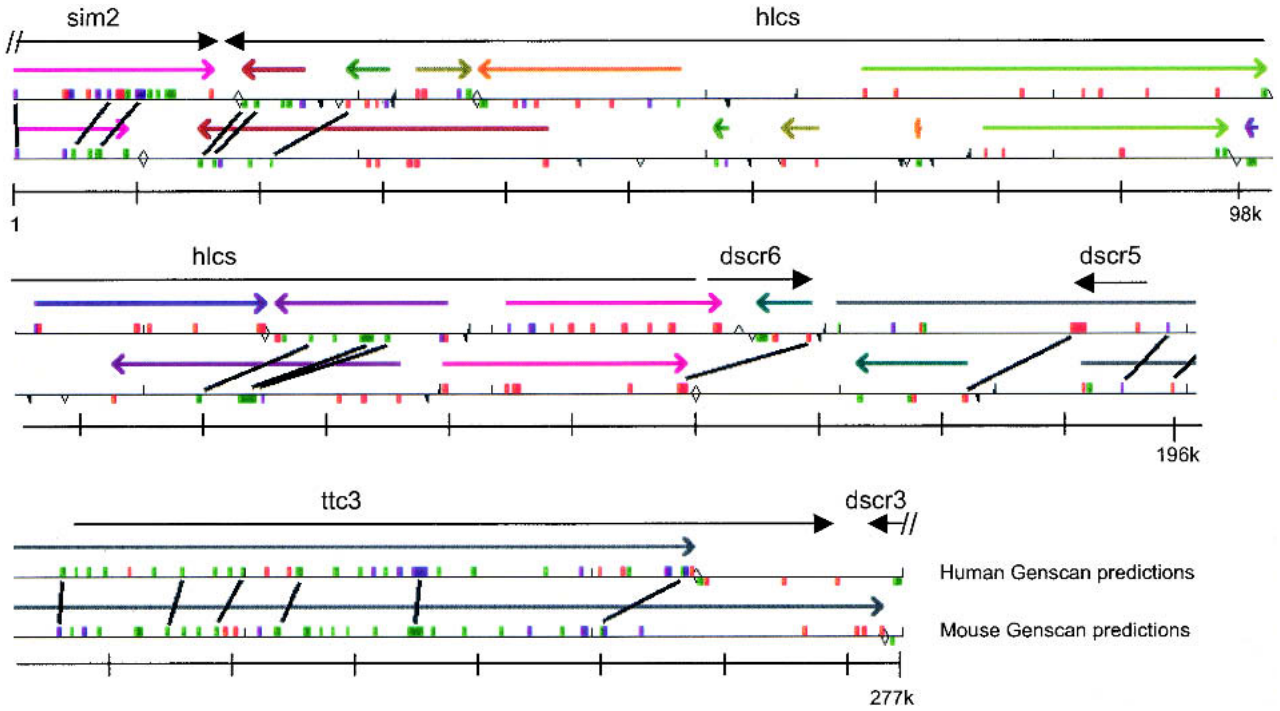

Figure 2 GENSCAN predictions from orthologous segments of mouse and human DNA do not closely correspond. Human and mouse sequences (1.8 and 1.3 Mb, respectively) were masked, and repeat sequences were removed to provide a closer correspondence in sequence length (1047 and 975 kb) for purposes of illustration. The first 277 kb of each sequence is shown here. This segment contains 84 conserved segments (based on BLAST2 sequence analysis with expected score ≥3e-18), but only 34 of 121 GENSCAN predicted human exons overlap one of these conserved sequences (indicated by black lines connecting the human and mouse sequence; only 20 diagonal lines appear because not all of the 34 matches resolve at this scale).

comparative sequence data into the GENSCAN prediction algorithm (Korf et al. 2001). TWINSCAN identified 18 gene models in this test region. Despite the consideration of sequence conservation in the TWINSCAN predictions, 13 of the 72 documented exons were missed. This included the entire *DSCR6* gene (Supplemental Table 2). The 18% false-negative rate of TWINSCAN, however, was an improvement over the 54% rate observed with the intersect of GENSCAN and comparative sequence analysis. TWINSCAN, like Genscan, had a high frequency of false-positive exon predictions. Both programs did well at predicting the positions of Sim2 exons, although only half of the TWINSCAN exons correspond to those actually in Sim2. TWINSCAN also predicted a large, multiexon gene spanning 80 kb within the 108 kb covered by *TTC3*. In all, 59 of 106 exons predicted by TWINSCAN were real, giving a false-positive rate of 46%.

Comparative analysis in this region is especially important to research in DS. A number of animal models have been studied based on mice with dosage imbalance of genes from this region of MMU16. Segmental trisomy 16 for this region in Ts65Dn and Ts1Cje mice produces phenotypes with direct parallels to those in DS (Reeves et al. 1995; Sago et al. 1998; Baxter et al. 2000; Richtsmeier et al. 2000, 2002; Cooper et al. 2001). The commonality of phenotypes indicates that the same developmental genetic pathways are disrupted by the corresponding dosage imbalance in the two species. It will also be of interest to investigate the function and occurrence in different mammalian taxa of genes that exist only in human or in mouse.

Recently, a hybridization-based whole chromosome resequencing strategy for chromosome 21 was reported by Frazer and colleagues (2001). An oligonucleotide wafer set representing the entire unique sequence of the chromosome was constructed. This technology is currently not practical for screening large numbers of individuals, as the chromosome 21 set is equivalent to ~800 standard Affymetrix chips. However, a conserved features chip containing all known exons and segments conserved over 80 Myr since the divergence of mouse and human would represent only 3% to 5% of this set. Such a resource could be extremely valuable in analysis of the contribution of allelic variation on chromosome 21 to the variable clinical presentation in DS.

Few studies have compared human and mouse genomic sequences >1 Mb in size (Loots et al. 2000; Onyango et al. 2000; Dehal et al. 2001), and none compared finished human and mouse sequence. The definitive sequences analyzed here provided substantially more power than our previous analyses using "complete" sequences (Pletcher et al. 2001). As substantial amounts of mouse genomic sequence data become available in public databases in the near future (Batzoglou et al. 2000), scientists will be challenged to narrow the gap between machine and human annotation based on comparative analyses.

METHODS

Sequencing and Data Assembly

Eleven PAC clones were screened against an 11.6× genomic-equivalent female mouse SV129 genomic library (RPCI-21; BAC/PAC Resources; www.chori.org/bacpac). PAC DNA was isolated using a standard alkali-SDS method in a PI-200 (Kurabo). The PAC DNA was purified by double cesium chloride centrifugation to remove most of *Escherichia coli* genomic DNA, and 2 μg of PAC DNA was mechanically sheared into 1.5- to 2.5-kb fragments using a Hydroshear apparatus (Genemachines) to prepare a shotgun clone library. Small fragments (<500 bp) were removed with Spin columns (Amersham-Pharmacia). The DNA fragments were blunt-ended and phosphorylated using a BKL kit (Takara) and subcloned into the dephosphorylated *Sma*I site of pUC18 vector (Amersham-Pharmacia). PCR-based template DNA preparation was performed as described (Hattori et al. 1997). After transformation, bacterial colonies were transferred by a colony picker (Flexys: Genomic Solutions) to a 384-format plate that contained 50 μL of L-broth in each well, and the suspension was incubated for 3 h to overnight at 37°C. An aliquot (0.1 μL) of the bacterial suspension was added to a 384-format plate, which contained 5 μL of PCR cocktail in each well. PCR-amplification of the insert DNA of shotgun clones was then performed using Ex-Taq polymerase (Takara). Excess PCR primers and unincorporated substrate nucleotides were digested by treating the PCR products with shrimp alkaline phosphatase and *E. coli* exonucleaseI (Amersham-Pharmacia) before sequencing. Plasmid subclones were prepared as described previously (Hattori et al. 1997). Dye-terminator cycle sequencing of the PCR products was performed using sequencing kits (Amersham-Pharmacia, ABI). Sequencing both ends of plasmid clones was also performed in a similar manner. Sequencing products were purified by gel filtration with Sephadex G50 (Millipore), and the products were run on automated capillary sequencers, MegaBACE 1000 and ABI 3700. Sequencing of shotgun and plasmid clones was performed to ~10× coverage of the estimated clone size. All the sequence data were transferred from sequencers to a UNIX platform for data assembly. Trace data was evaluated and assembled using Phred/Phrap software (Ewing et al. 1998). Finishing was performed with CONSED (Gordon et al. 1998) and Sequencher (Gene Codes Corp.). Gaps and ambiguities in the assembled data were experimentally resolved using several techniques such as nested deletion (Hattori et al. 1997), primer walk, and PCR-coupled primer walk. Direct sequencing of PAC DNA using appropriate primers was also used in the finishing process.

Global Analysis of Sequence

The Sim4 program (Florea et al. 1998) was used to align cDNA with genomic sequences. RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) was used to search human and mouse repetitive elements and to calculate GC content. Homology search was performed for masked sequences using BLAST program against nonredundant, EST, and protein databases (Altschul et al. 1990). Exon prediction was performed using GENSCAN (Burge and Karlin 1997). The VISTA program (Mayor et al. 2000; http://sichuan.lbl.gov/vista) and the CGAT program (Lund et al. 2000; http://inertia.bs.jhmi.edu/roger/CGAT/CGAT.html) were used for comparative sequence analysis with varying levels of stringency as described in the text.

Cloning of Mouse *Hlcs* cDNA

Mouse *Hlcs* cDNA was isolated by PCR using primers designed from the finished mouse genomic sequence. The PCR primer sets were as follows: a, 5′-AGCGTGGACAACTTCAGCAAGCT-3′; b, 5′-GGACAAATGGAATCCTCTGTCCC-3′; c, 5′-AGATGC CGCAGGAAATGGGCTTA-3′; and d, 5′-TGAAGTCTTTGGTT CATGTGAGGC-3′. Approximately 1.6 kb (primer a/b) and 2.8 kb (primer c/d) of PCR products were detected and isolated from a mouse testis cDNA library (Marathon-Ready cDNA, Clontech). Primer-walk for sequencing of the cDNA was performed by using appropriate primers. The accession no. of the mouse *Hlcs* cDNA sequence is AB066227.

ACKNOWLEDGMENTS

We thank all technical staff of the Human Genome Research Group in RIKEN-GSC for their contribution to this work. Chiharu Kawagoe (Hitachi, Ltd.) and Ranji Thekkil (J.H.U.) provided excellent support of computational data management. This work was supported by grants from the Ministry of Education, Science, Sports and Culture of Japan and by US Public Health Service Award HD38384 (R.H.R.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **21:5** 403–410.

Barlow, G.M., Chen, X.N., Shi, Z.Y., Lyons, G.E., Kurnit, D.M., Celle, L., Spinner, N.B., Zackai, E., Pettenati, M.J., Van Riper, A.J., et al. 2001. Down syndrome congenital heart disease: A narrowed region and a candidate gene. *Genet. Med.* **3:** 91–101.

Batzoglou, S., Pachter, L., Mesirov, J., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.

Baxter, L.L., Moran, T.H., Richtsmeier, J.T., Troncoso, J., and Reeves, R.H. 2000. Discovery and genetic localization of Down syndrome cerebellar phenotypes using the Ts65Dn mouse. *Hum. Mol. Genet.* **9:** 195–202.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structure in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Casavant, N.C., Scott, L., Cantrell, M.A., Wiggins, L.E., Baker, R.J., and Wichman, H.A. 2000 The end of the LINE?: Lack of recent L1 activity in a group of South American rodents. *Genetics* **154:** 1809–1817.

Cooper, J.D., Salehi, A., Delcroix, J.D., Howe, C.L., Belichenko, P.V., Chua-Couzens, J., Kilbridge, J.F., Carlson, E.J., Epstein, C.J., and Mobley, W.C. 2001. Failed retrograde transport of NGF in a mouse model of Down's syndrome: Reversal of cholinergic neurodegenerative phenotypes following NGF infusion. *Proc. Natl. Acad. Sci.* **98:** 10439–10444.

Dahmane, N., Ghezala, G.A., Gosset, P., Chamoun, Z., Dufresne-Zacharia, M.C., Lopes, C., Rabatel, N., Gassanova-Maugenre, S., Chettouh, Z., Abramowski, V., et al. 1998. Transcriptional map of the 2.5-Mb CBR-ERG region of chromosome 21 involved in Down syndrome. *Genomics* **48:** 12–23.

Dehal, P., Predki, P., Olsen, A.S., Kobayashi, A., Folta, P., Lucas, S., Land, M., Terry, A., Ecale Zhou, C.L., Rash, S., et al. 2001. Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. *Science* **293:** 104–111.

Epstein, C.J., Korenberg, J.R., Anneren, G., Antonarakis, S.E., Ayme, S., Courchesne, E., Epstein, L.B., Fowler, A., Groner, Y., Huret, J.L., et al. 1991. Protocols to establish genotype-phenotype correlations in Down syndrome. *Am. J. Hum. Genet.* **49:** 207–235.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using `Phred`, I: Accuracy assessment. *Genome Res.* **8:** 175–185.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.

Frazer, K.A., Sheehan, J.B., Stokowski, R.P., Chen, X., Hosseini, R., Cheng, J.F., Fodor, S.P., Cox, D.R., and Patil, N. 2001. Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* **11:** 1651–1659.

Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

Hassold, T., Abruzzo, M., Adkins, K., Griffin, D., Merrill, M., Millie, E., Saker, D., Shen, J., and Zaragoza, M. 1996. Human aneuploidy: Incidence, origin and etiology. *Environ. Mol. Mutagen* **28:** 167–175.

Hattori, M., Tsukahara, F., Furuhata, Y., Tanahashi, H., Hirose, M., Saito, M., Tsukuni, S., and Sakaki, Y. 1997. A novel method for making nested deletions and its application for sequencing of a 300 kb region of human APP locus. *Nucleic Acids Res.* **25:** 1802–1808.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.-S., Toyoda, A., Ishii, K. Totoki, Y., Choi, D.-K., et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405:** 311–319.

Korenberg, J.R., Chen, X.N., Schipper, R., Sun, Z., Gonsky, R., Gerwehr, S., Carpenter, N., Daumer, C., Dignan, P., Disteche, C., et al. 1994. Down syndrome phenotypes: The consequences of chromosomal imbalance. *Proc. Natl. Acad. Sci.* **91:** 4997–5001.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17:** S140–S148.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.

Lund, J., Chen, F., Hua, A., Roe, B., Budarf, M., Emanuel, B.S., and Reeves, R.H. 2000. Comparative sequence analysis of 634 kb of the mouse chromosome 16 region of conserved synteny with the human velocardiofacial syndrome region on chromosome 22q11.2. *Genomics* **63:** 374–383.

Mallon, A.M., Platzer, M., Bate, R., Gloeckner, G., Botcherby, M.R., Nordsiek, G., Strivens, M.A., Kioschis, P., Dangel, A., Cunningham, D., et al. 2000. Comparative genome sequence analysis of the Bpa/Str region in mouse and man. *Genome Res.* **10:** 758–775.

Martindale, D.W., Wilson, M.D., Wang, D., Burke, R.D., Chen, X., Duronio, V., and Koop, B.F. 2000. Comparative genomic sequence analysis of the Williams syndrome region (LIMK1-RFC2) of human chromosome 7q11.23. *Mamm. Genome* **11:** 890–898.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. 2000. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046–1047.

Ohira, M., Ichikawa, H., Suzuki, E., Iwaki, M., Suzuki, K., Saito-Ohara, F., Ikeuchi, T., Chumakov, I., Tanahashi, H., Tashiro, K., et al. 1996. A 1.6-Mb P_1-based physical map of the Down syndrome region on chromosome 21. *Genomics* **33:** 65–74.

Onyango, P., Miller, W., Lehoczky, J., Leung, C.T., Birren, B., Wheelan, S., Dewar, K., and Feinberg, A.P. 2000. Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. *Genome Res.* **10:** 1697–1710.

Osoegawa, K., Tateno, M., Woon, P.Y., Frengen, E., Mammoser, A.G., Catanese, J.J., Hayashizaki, Y., and de Jong, P.J. 2000. Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.* **10:** 116–128.

Pletcher, M., Roe, B.A. , Chen, F., Do, T., Do, A., Malaj, E., and Reeves, R.H. 2000. Chromosome evolution: The junction of mammalian chromosomes in the formation of mouse chromosome 10. *Genome Res.* **10:** 1463–1467.

Pletcher, M.T., Wiltshire, T., Cabin, D.E., Villanueva, M., and Reeves, R.H. 2001. Use of comparative physical and sequence mapping to annotate mouse chromosome 16 and human chromosome 21. *Genomics* **74:** 45–54.

Reeves, R.H., Irving, N.G., Moran, T.H., Wohn, A., Kitt, C., Sisodia, S.S., Schmidt, C., Bronson, R.T., and Davisson, M.T. 1995. A mouse model for Down syndrome exhibits learning and behavior deficits. *Nat. Genet.* **11:** 177–184.

Reeves, R.H., Baxter, L.L., and Richtsmeier, J.T. 2001. Too much of a good thing: Mechanisms of gene action in Down syndrome. *Trend Genet.* **17:** 83–88.

Richtsmeier, J.T., Baxter, L.L., and Reeves, R.H. 2000. Parallels of craniofacial maldevelopment in Down syndrome and Ts65Dn mice. *Dev. Dyn.* **217:** 137–145.

Richtsmeier, J.T., Epstein, C.J., Carlson, E., and Reeves, R.H. 2002. Craniofacial phenotypes in segmentally trisomic mouse models for Down syndrome. *Am. J. Med. Genet.* **107:** 317–324

Sago, H., Carlson, E.J., Smith, D.J., Kilbridge, J., Rubin, E.M., Mobley, W.C., Epstein, C.J., and Huang, T.T. 1998. Ts1Cej, a partial trisomy 16 mouse model for Down syndrome, exhibits learning and behavioral abnormalities. *Proc. Natl. Acad. Sci.* **95:** 6256–6261.

Shibuya, K., Kudoh, J., Minoshima, S., Kawasaki, K., Asakawa, S., and Shimizu, N. 2000. Isolation of two novel genes, DSCR5 and DSCR6, from Down syndrome critical region on human chromosome 21q22.2. *Biochem. Biophys. Res. Commun.* **19:** 693–698.

Sumarsono, S.H., Wilson, T.J., Tymms, M.J., Venter, D.J., Corrick, C.M., Kola, R., Lahoud, M.H., Papas, T.S., Seth, A., and Kola. I. 1996. Down's syndrome-like skeletal abnormalities in Ets2 transgenic mice. *Nature* **379:** 534–537.

WEB SITE REFERENCES

http://inertia.bs.jhmi.edu/roger/CGAT/CGAT.html; CGAT program.
http://hgp.gsc.riken.go.jp; Reeves Laboratory Web page with free download of CGAT comparative sequence software.
http://www.chori.org/bacpac; 11.6× genomic-equivalent female mouse SV129 genomic library.
http://ftp.genome.washington.edu/RM/RepeatMasker.html; RepeatMasker program.
http://genome.wustl.edu/Overview/finrulesname.php; rules for human genome sequencing.
http://sichuan.lbl.gov/vista; VISTA program.

Received February 1, 2002; accepted in revised form June 12, 2002.