

## Human Chromosome 7: DNA Sequence and Biology

Stephen W. Scherer,<sup>1,5\*</sup> Joseph Cheung,<sup>1</sup> Jeffrey R. MacDonald,<sup>1</sup> Lucy R. Osborne,<sup>6</sup> Kazuhiko Nakabayashi,<sup>1</sup> Jo-Anne Herbrick,<sup>1</sup> Andrew R. Carson,<sup>1</sup> Layla Parker-Katiraei,<sup>1,5</sup> Jennifer Skaug,<sup>1</sup> Razi Khaja,<sup>1</sup> Junjun Zhang,<sup>1</sup> Alexander K. Hudek,<sup>1</sup> Martin Li,<sup>1</sup> May Haddad,<sup>1</sup> Gavin E. Duggan,<sup>1</sup> Bridget A. Fernandez,<sup>7</sup> Emiko Kanematsu,<sup>1</sup> Simone Gentles,<sup>1</sup> Constantine C. Christopoulos,<sup>1</sup> Sanaa Choufani,<sup>1</sup> Dorota Kwasnicka,<sup>1</sup> Xiangqun H. Zheng,<sup>8</sup> Zhongwu Lai,<sup>8</sup> Deborah Nusskern,<sup>8</sup> Qing Zhang,<sup>8</sup> Zhiping Gu,<sup>8</sup> Fu Lu,<sup>8</sup> Susan Zeesman,<sup>9</sup> Malgorzata J. Nowaczyk,<sup>9</sup> Ikuko Teshima,<sup>1,2,11</sup> David Chitayat,<sup>2,11</sup> Cheryl Shuman,<sup>1,2,11</sup> Rosanna Weksberg,<sup>1,2,11</sup> Elaine H. Zackai,<sup>12</sup> Theresa A. Grebe,<sup>13</sup> Sarah R. Cox,<sup>13</sup> Susan J. Kirkpatrick,<sup>14</sup> Nazneen Rahman,<sup>15</sup> Jan M. Friedman,<sup>16</sup> Henry H. Q. Heng,<sup>17</sup> Pier Giuseppe Pelicci,<sup>18,19</sup> Francesco Lo-Coco,<sup>20</sup> Elena Belloni,<sup>18,19</sup> Lisa G. Shaffer,<sup>21</sup> Barbara Pober,<sup>22</sup> Cynthia C. Morton,<sup>23,24,26</sup> James F. Gusella,<sup>27</sup> Gail A. P. Bruns,<sup>28</sup> Bruce R. Korf,<sup>25,26</sup> Bradley J. Quade,<sup>24</sup> Azra H. Ligon,<sup>24</sup> Heather Ferguson,<sup>23</sup> Anne W. Higgins,<sup>23</sup> Natalia T. Leach,<sup>24</sup> Steven R. Herrick,<sup>24</sup> Emmanuelle Lemyre,<sup>23</sup> Chantal G. Farra,<sup>23</sup> Hyung-Goo Kim,<sup>27</sup> Anne M. Summers,<sup>29</sup> Karen W. Gripp,<sup>30</sup> Wendy Roberts,<sup>3</sup> Peter Szatmari,<sup>10</sup> Elizabeth J. T. Winsor,<sup>31</sup> Karl-Heinz Grzeschik,<sup>32</sup> Ahmed Teebi,<sup>2,11</sup> Berge A. Minassian,<sup>1,4</sup> Juha Kere,<sup>33</sup> Lluis Armengol,<sup>34</sup> Miguel Angel Pujana,<sup>34</sup> Xavier Estivill,<sup>34</sup> Michael D. Wilson,<sup>35</sup> Ben F. Koop,<sup>35</sup> Sabrina Tosi,<sup>36</sup> Gudrun E. Moore,<sup>37</sup> Andrew P. Boright,<sup>38</sup> Eitan Zlotorynski,<sup>39</sup> Batsheva Kerem,<sup>39</sup> Peter M. Kroisel,<sup>40</sup> Erwin Petek,<sup>40</sup> David G. Oscier,<sup>41</sup> Sarah J. Mould,<sup>41</sup> Hartmut Döhner,<sup>42</sup> Konstanze Döhner,<sup>42</sup> Johanna M. Rommens,<sup>1,5</sup> John B. Vincent,<sup>43</sup> J. Craig Venter,<sup>8</sup> Peter W. Li,<sup>8</sup> Richard J. Mural,<sup>8</sup> Mark D. Adams,<sup>8</sup> Lap-Chee Tsui<sup>1,5,†</sup>

DNA sequence and annotation of the entire human chromosome 7, encompassing nearly 158 million nucleotides of DNA and 1917 gene structures, are presented. To generate a higher order description, additional structural features such as imprinted genes, fragile sites, and segmental duplications were integrated at the level of the DNA sequence with medical genetic data, including 440 chromosome rearrangement breakpoints associated with disease. This approach enabled the discovery of candidate genes for developmental diseases including autism.

With the advent of the Human Genome Project (HGP), a wealth of resources including genetic (1), physical (2, 3), gene (4), and draft DNA sequence maps (5, 6) have facilitated the discovery of more than 360 disease-associated genes and loci on chromosome 7 (table S1).

Here we present a comprehensive assembly of 157,953,789 nucleotides (nt) of DNA covering human chromosome 7. About 85% of the content was derived from a subset of unpublished Celera whole-genome scaffolds

for chromosome 7 (7) based on updates of previous work (5). Another 15% was from new or updated clone-based sequences from the International Human Genome Sequencing Consortium (notably the Washington University Genome Sequencing Center) and other sources (supporting online text) (tables S2 and S3). The assembly (named CRA\_TCAGchr7.v1) is available at a public Web site ([www.chr7.org/](http://www.chr7.org/)) and in GenBank (7). To maximize the utility of the

sequence for discovery, we incorporated biological and medically relevant features

<sup>1</sup>Department of Genetics and Genomic Biology, <sup>2</sup>Division of Clinical and Metabolic Genetics, <sup>3</sup>The Child Development Centre, <sup>4</sup>Division of Neurology, Department of Paediatrics, The Hospital for Sick Children, Toronto, Ontario, Canada, M5G 1X8. <sup>5</sup>Department of Molecular and Medical Genetics, <sup>6</sup>Department of Medicine, University of Toronto, Toronto, Ontario, Canada, M5S 1A8. <sup>7</sup>Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada, A1B 3V6. <sup>8</sup>Celera Genomics, Rockville, MD 20850, USA. <sup>9</sup>Hamilton Health Sciences Centre and McMaster University, <sup>10</sup>Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario, Canada, L8N 3Z5. <sup>11</sup>Department of Pediatrics, University of Toronto, Toronto, Ontario, Canada, M5G 1X8. <sup>12</sup>Division of Human Genetics and Molecular Biology, The Children's Hospital of Philadelphia, Philadelphia, PA 19104-4301, USA. <sup>13</sup>University of Phoenix Genetics Program, Phoenix, AZ 85016, USA. <sup>14</sup>Department of Medical Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA. <sup>15</sup>Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey, SM2 5NG, UK. <sup>16</sup>Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada, V6H 3N1. <sup>17</sup>Wayne State University School of Medicine, Detroit, MI 48202, USA. <sup>18</sup>European Institute of Oncology, Department of Experimental Oncology, 20141 Milan, Italy. <sup>19</sup>Firc Institute for Molecular Oncology, Cancer Genetics Unit, 20134 Milan, Italy. <sup>20</sup>Universita di Roma Tor Vergata, Dipartimento di Biopatologia e Diagnostica per Immagini, 00133 Rome, Italy. <sup>21</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. <sup>22</sup>Department of Genetics, Yale University School of Medicine, New Haven, CT 06520-8005, USA. <sup>23</sup>Department of Obstetrics, Gynecology and Reproductive Biology, <sup>24</sup>Department of Pathology, <sup>25</sup>Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA. <sup>26</sup>Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston, MA 02115, USA. <sup>27</sup>Molecular Neurogenetics Laboratory, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129, USA. <sup>28</sup>Department of Pediatrics, The Children's Hospital, Harvard Medical School, Boston, MA 02115, USA. <sup>29</sup>Department of Genetics, North York General Hospital, Toronto, Ontario, Canada, M2K 1E1. <sup>30</sup>Division of Medical Genetics, A. I. duPont Hospital for Children, Wilmington, DE 19899, USA. <sup>31</sup>Prenatal Diagnosis Program and Department of Laboratory Medicine and Pathobiology, University Health Network, The University of Toronto, Toronto, Ontario, Canada, M5G 1X5. <sup>32</sup>Medizinisches Zentrum für Humangenetik der Universität Marburg, D35037 Marburg, Germany. <sup>33</sup>Department of Biosciences, Karolinska Institute, at Novum and Clinical Research Centre, Huddinge University Hospital, S-141 57 Stockholm, Sweden. <sup>34</sup>Program in Genes and Disease, Centre for Genomic Regulation, 08003 Barcelona, Catalonia, Spain. <sup>35</sup>Department of Biology, University of Victoria, Victoria, British Columbia, Canada, V8W 3N5. <sup>36</sup>MRC Molecular Haematology Unit, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, UK. <sup>37</sup>Department of Fetal and Maternal Medicine, Institute of Reproductive and Developmental Biology, Imperial College, Faculty of Medicine, Hammersmith Campus, London W12 0NN, UK. <sup>38</sup>Division of Endocrinology, Department of Medicine, University Health Network, University of Toronto, Toronto, Ontario, Canada, M5G 2C4. <sup>39</sup>Department of Genetics, The Life Sciences Institute, The Hebrew University, Jerusalem, 91904 Israel. <sup>40</sup>Institute of Medical Biology and Human Genetics, Karl-Franzens University of Graz, A-8010 Graz, Austria. <sup>41</sup>Department of Haematology, Royal Bournemouth Hospital, Bournemouth, BH7 7DW UK. <sup>42</sup>Department of Internal Medicine III, University Hospital of Ulm, Ulm, Germany, 89081. <sup>43</sup>Centre for Addiction and Mental Health, Clarke Institute and Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada, M5T 1R8.

\*To whom correspondence should be addressed. E-mail: [steve@genet.sickkids.on.ca](mailto:steve@genet.sickkids.on.ca)

†Present address: The University of Hong Kong, Pokfulam Road, Hong Kong.

RESEARCH ARTICLE

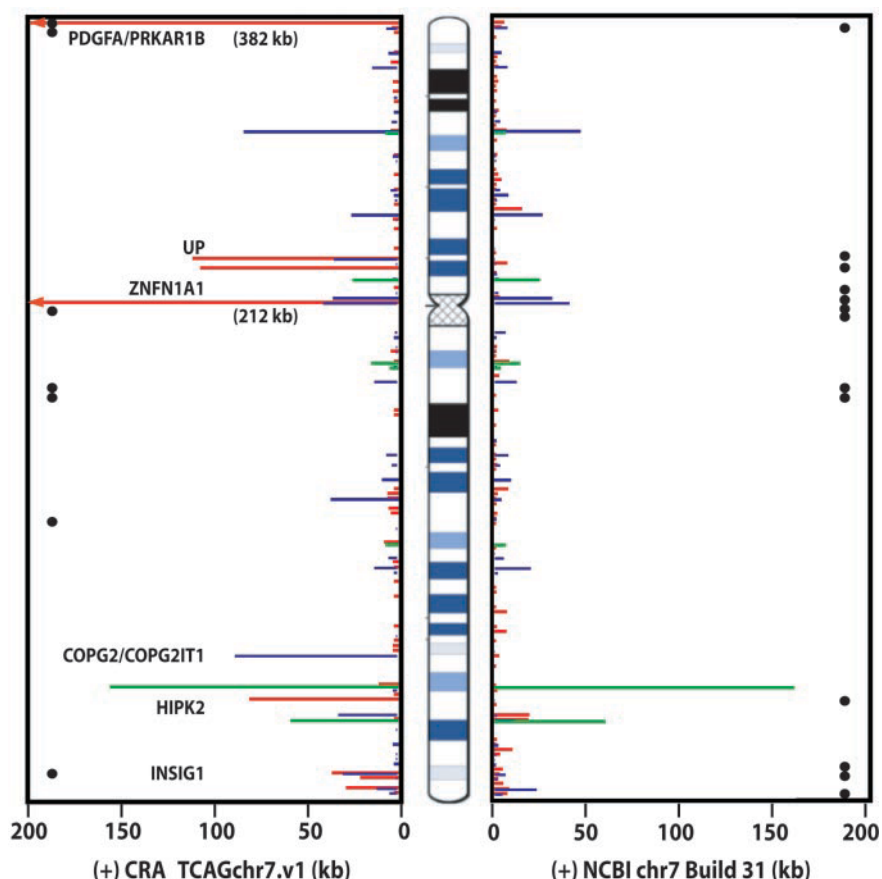
from all available databases, the literature, and our data (7). Wherever possible, computer-based annotations of the sequence were exam-

ined manually and validated experimentally. Moreover, we included patient analysis as an aspect of the sequence annotation to increase

knowledge of the function and regulation of genes. The Generic Model Organism Database (8) and its Genome Browser function were implemented to display all mapping, sequencing, structural, and clinical data to provide a mechanism and dynamic platform for human chromosome 7 annotation.

The assembled sequences were positioned to cytogenetic bands on chromosome 7 by fluorescence in situ hybridization (FISH) with 1440 genomic clones (7). The FISH resource also assisted in confirming order and copy number in chromosomal regions containing low-copy or complex repeats (9, 10). For the 770 bacterial genomic clones displayed in the Genome Browser, FISH experiments were reproduced more than once in at least two laboratories to allow accurate cytogenetic boundaries to be established. The sequence assembly reached both telomere ends and encompassed the apparent junction sequences between the euchromatic arms, and the *D7Z2* and *D7Z1* centromeric satellites on 7p and 7q, respectively. Because the centromere is polymorphic (ranging in size from 1500 to 3800 kb at *D7Z1* and 100 to 500 kb at *D7Z2*) (11, 12), 2,700,000 nucleotides (nt) were substituted to represent an average-sized chromosome 7.

We tested all available genomic data against our assembly, including the latest National Center for Biotechnology Information (NCBI) chromosome 7 sequence database (Build 31) (supporting online text). Using the PatternHunter program (13) to compare CRA\_TCAGchr7.v1 and Build 31, we found (i) a total of 1,186,913 nts of unmatched sequence between the assemblies, (ii) 132 other sites (encompassing 508,332 nt) where different sequences were found at the same relative chromosomal positions (termed sequence variations), and (iii) 10 equivalent DNA segments placed in an inverted orientation between the two assemblies (Fig. 1; figs. S1 and



**Fig. 1.** DNA sequence comparison of CRA\_TCAGchr7.v1 against NCBI Build 31. Black circles represent the sites of physical gaps. The sites and extent of unmatched sequences present in one assembly but not the other are shown in red, sequence variations in blue, and inversions in green. Genes present in CRA\_TCAGchr7.v1, but absent in Build 31, are shown (see table S4; complete dataset is at [www.chr7.org/](http://www.chr7.org/)).

**Table 1.** Chromosome 7 gene summary. *TCRB* encompasses 67 gene and 16 pseudogene segments. *TCRG* includes 14 gene segments and 8 pseudogene segments. NCBI Refseq and Ensembl use different criteria; for

comparison, we have grouped similar categories in the same row. The complete list of genes is at [www.chr7.org/](http://www.chr7.org/), and it will continue to be updated.

Categories of genes	No. of genes	Gene length (bp)	Transcript length (bp)	Exon size (bp)	No. of exons per gene	CpG islands (-2000 to 1000 bp)	No. of chromosome 7 genes from other projects	
							NCBI RefSeq	Ensembl
Known genes	863	69,877	2,639	261	10.1	541 (63%)	355 (reviewed)	1,053 (known)
Novel genes	71	50,103	1,989	386	5.2	27 (38%)	276 (provisional)	297 (novel)
Partial genes	40	42,964	1,850	339	5.5	20 (45%)	172 (predicted)	
Predicted genes	481	14,573	1,026	326	3.1	102 (21%)	520 (model)	
Putative and noncoding RNA genes	213	17,501	1,638	629	2.6	45 (21%)		
<b>Total genes</b>	<b>1,668</b>						<b>1,323</b>	<b>1,350</b>
TCR (gene segments)	81	565	290					
TCR pseudogene segments	24	454	349					
Pseudogenes	144	1,439	1,056					
<b>Total structures</b>	<b>1,917</b>							

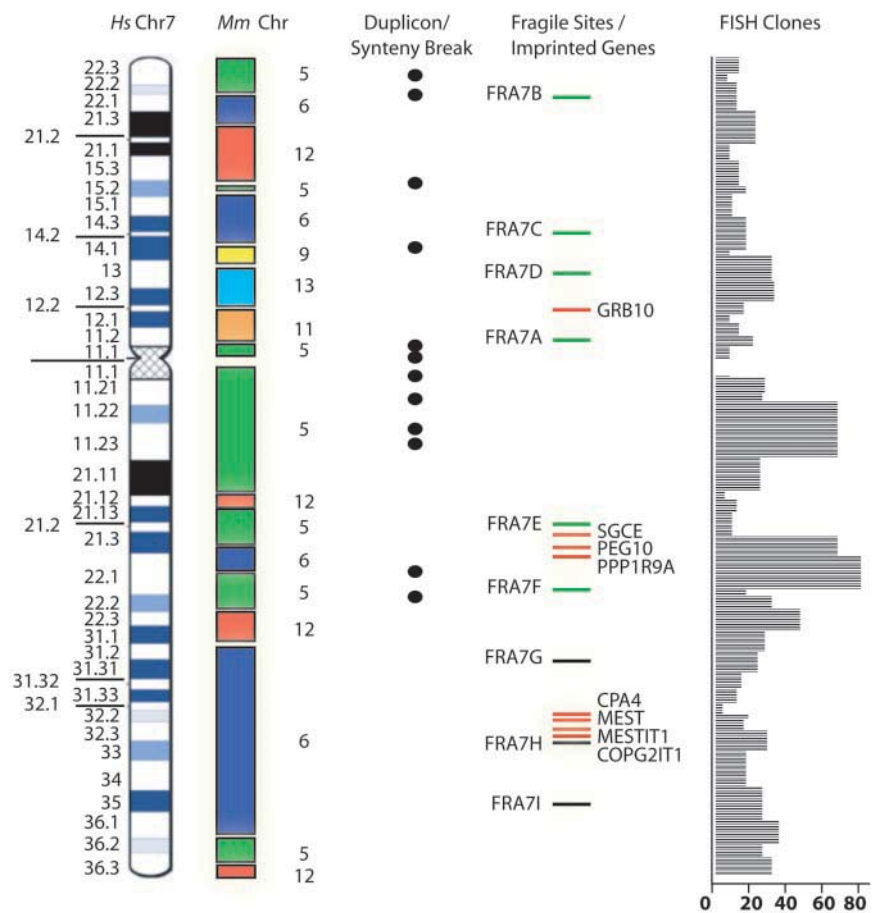
S2, table S4). The differences detected could be due to rearrangements arising during cloning, assembly mistakes, or polymorphism between the source chromosomal DNA (no correlation was observed between inverted regions and known genomic polymorphism or discrepancies in genetic maps).

**Chromosome 7 functional and structural features.** Through comparison of the Celera mouse genome sequence to our assembly, 21,859 syntenic anchor points (14) from six murine chromosomes (5, 6, 9, 11, 12, and 13) were identified (Fig. 2; table S5, a and b). The syntenic anchor points were grouped into 36 blocks, 14 of which had not been reported before (15, 16).

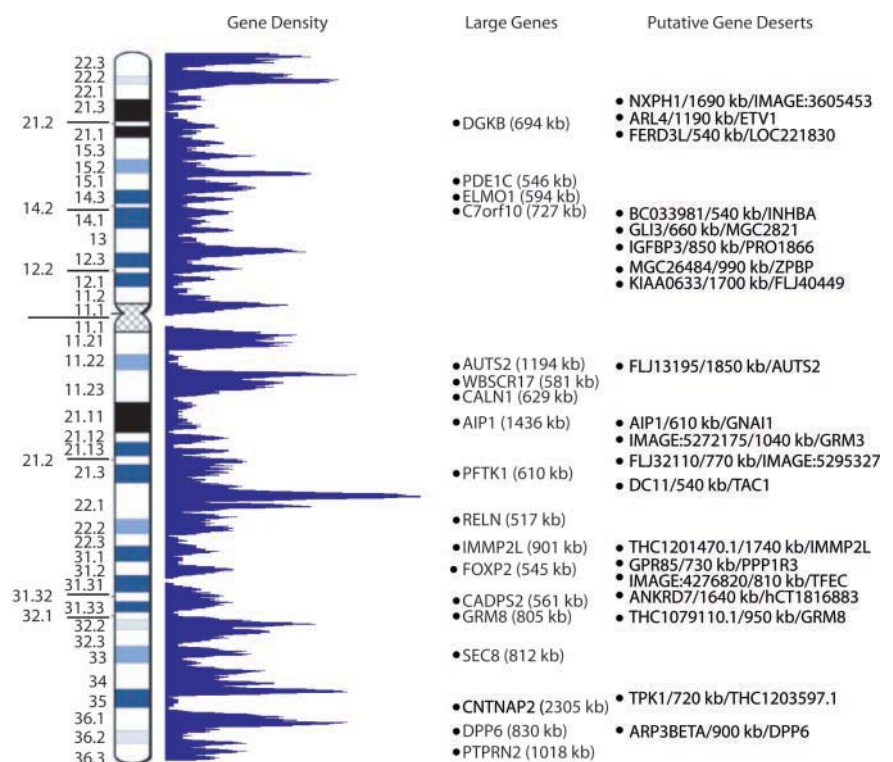
To generate the most complete description of genes on chromosome 7, we used computer-based annotation in conjunction with extensive laboratory experimentation (7). A team of reviewers scrutinized data and, by comparison to the reference DNA sequence, defined 1917 gene structures (Table 1). Their distribution along chromosome 7 is shown in Fig. 3. The description of 297 of these gene units was either exclusive to our dataset, or present in a more complete or different form, as seen from comparison with other databases.

The gene structures were grouped into eight categories: (i) 863 known genes or human full-length cDNA sequences present in LocusLink or HUGO databases; (ii) 71 novel genes, full-length cDNA, or expressed sequence tag (EST) clusters that contain an open reading frame (ORF) (>100 amino acids) without a formal name; (iii) 40 partial genes, human cDNA, or EST clusters with an incomplete ORF (missing the start or stop codon); (iv) 481 predicted genes or predicted gene models, for which at least one exon matches supportive evidence (EST, protein homology, or mouse sequence) of nonoverlapping NCBI, RefSeq, or Ensembl entries; (v) 213 putative and noncoding RNA genes, human cDNA, or EST clusters that do not contain an apparent ORF (51 that have homology in mouse); (vi) 81 gene segments from the two T cell receptor (*TCR*) loci; (vii) 24 *TCR* pseudogene segments; and (viii) 144 pseudogenes. Overall, our data suggest that 1455 potential protein coding genes (known, novel, partial, predicted) and 213 putative and noncoding RNA genes reside on chromosome 7. Extrapolating these chromosome 7 numbers to the human genome, one would predict that there are about 29,000 protein coding genes and 3700 putative and noncoding RNA genes, consistent with some other estimates (5, 6, 17). Of the known genes, 474 of 863 (55%) were found to have one or more alternatively spliced forms, comparable to those observed on chromosomes 14 (54%) and 22 (59%) (18, 19).

The average gene size on chromosome 7 was 69.9 kb, exceeding what was reported previously (5, 6). There were 18 genes greater than 500 kb in size (Fig. 3); the largest,



**Fig. 2.** Six mouse chromosomes with synteny to human chromosome 7, 12 syntenic breakpoints overlapping segmental duplications, 9 fragile sites (*FRA7E*, *FRA7G*, *FRA7H*, *FRA7I* being cloned), 8 imprinted genes (7), and 770 bacterial clones anchored to the sequence.



**Fig. 3.** Distribution of 1917 gene structures and 20 putative gene deserts on chromosome 7.

## RESEARCH ARTICLE

*CNTNAP2*, spanned 2300 kb. The q22 Giemsa light band had the highest gene density. At one site (coordinates 98.2 to 99.2 Mb), 56 genes were found; that exceeds the mean of 10.7 genes/Mb for the rest of chromosome 7. If the 1749 annotated genes (excluding pseudogenes) are considered, they cover a total of 72.9 Mb of sequence (intragenic regions), which suggests at least 46.5% of chromosome 7 is transcribed (the average intergenic distance was 42.4 kb). A total of 1335 CpG islands were identified on chromosome 7, of which 63% (541 of 863) resided in the 5' end of a known gene.

Overlapping genes (total 100) were identified through comparison of the sequence coordinates for each known, novel, or partial gene. The pairs were then categorized on the basis of the type of sequence overlap and transcriptional orientation (7). After excluding splice variants, 38 sense-antisense gene pairs (with direct sequence overlap) were identified, 8 and 18 of which were in head-to-head or tail-to-tail orientation, respectively (table S6). The median size of sequence overlap between these transcripts was 238 bp, and for 23 out of 38 (61%), this occurred in the coding sequence. There were also 18 sense-sense

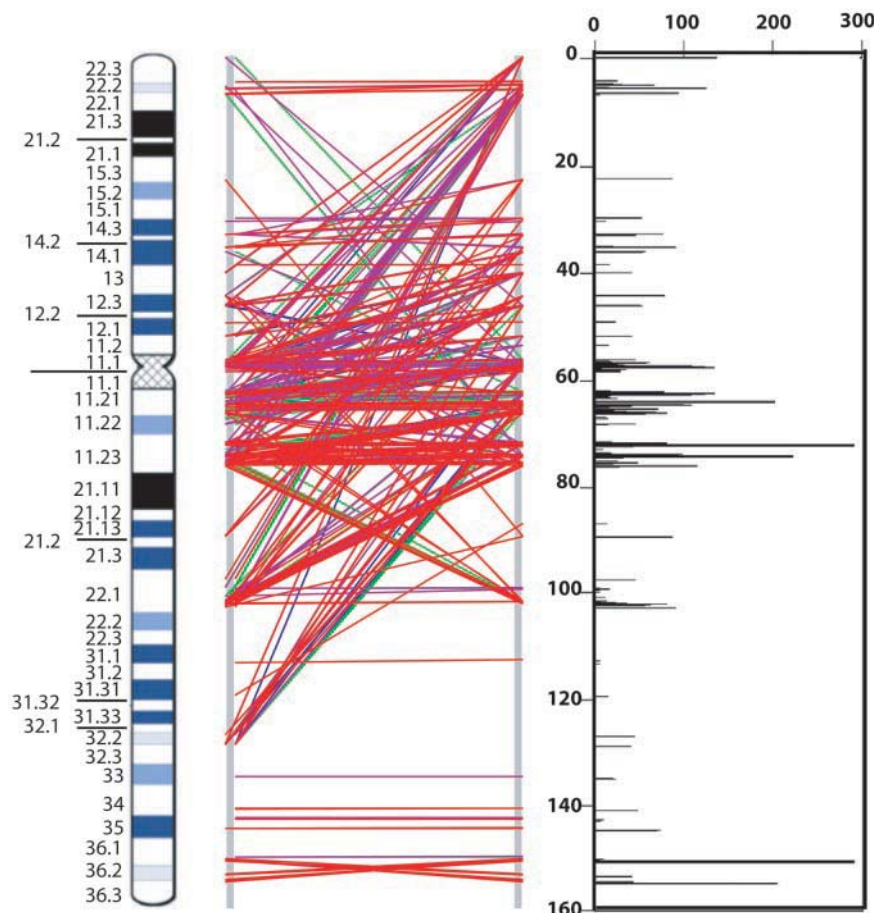
(same strand) and 43 sense-antisense overlapping gene pairs that did not share sequence overlap, but occupied the same genomic domain. We found no bias in the number or position of CpG islands near overlapping transcripts. We did, however, observe that a disproportionate number of known imprinted genes, 5 of 8 on chromosome 7 (Fig. 2), had an overlapping transcript (table S6) relative to the total number of transcripts on chromosome 7.

We discovered 20 euchromatic regions each greater than 500 kb in size where no known, novel, or partial gene was found (named putative gene deserts; Fig. 3, table S7). These intervals, which were mostly (16 out of 20) located within or at the boundary of Giemsa dark bands, covered 20.5 Mb (13%) of chromosome 7; the largest was 1850 kb. They contained a low number of CpGs (0.8/Mb vs. 9.8/Mb control) and short interspersed nuclear elements (SINEs) (7.5 versus 16%), a high density of long interspersed nuclear elements (LINEs) (28.1 versus 18.8%), and a decreased mouse syntenic anchor point density (4.2 versus 7.2%). Gene-poor regions described on chromosomes 14, 20, and 21 (18, 20, 21) also exhibited similar characteristics (in all cases, each category yielded statistically sig-

nificant results when compared with controls,  $P < 0.0001$ ). Moreover, our analysis of the equivalent regions in mouse also did not yield any new genes, which suggests that these 20 deserts occurred before the divergence of the human and mouse genomes (table S7). As in humans, the mouse region contained low SINE (3.8 versus 13.6% average) and high LINE (30.2 versus 21%) content. However, the observation that 14 of 20 of these deserts were larger (average increase in size was 19%) in mouse was intriguing given the estimates of a 15% compression in overall size of the murine genome compared with the human (14, 16).

As part of this project, we had shown previously that segmental duplications (duplicons) on chromosome 7 can be targets for nonallelic homologous recombination leading to genomic deletions or inversions in Williams-Beuren syndrome (WBS) (10), or gene conversion events in Shwachman-Diamond syndrome (22). Moreover, our analysis of the entire human genome (using NCBI Build 31) revealed that chromosome 7 contained the largest amount of intrachromosomal duplication (23), which suggests that there could be other disease associations. To complete a more refined analysis of chromosome 7, we compared the CRA\_TCAGchr7.v1 assembly with itself to search for all recent (>90% sequence identity) and large (>5 kb) intrachromosomal duplications. Overall, 146 distinct segments were identified, which composed 5.3% (8.3 out of 157.9 Mb) of the chromosome (Fig. 4).

Large duplicons (>100 kb) were identified at 7p22, 7p14-p15, the pericentromeric region, 7q11.21, 7q11.23, 7q22, and 7q36. We identified segmental duplications on chromosome 7 contained in 37 bacterial artificial clones (BACs) (confirmed by FISH) not described before (24). Notwithstanding, sequence analysis and metaphase FISH alone would not allow detection of all segmental duplications, and our finding that duplicons were present at 4 of the 7 remaining physical gaps on chromosome 7 suggests additional complexities (Fig. 1). Near-identical duplicons situated directly adjacent to each other on the chromosome could also be missed. Using high-resolution FISH analysis, we discovered one such segmental duplication about 1 Mb in size just telomeric to the WBS region (spanning *D7S2470* to *D7S2545*) (7). Before we can add this to the sequence assembly, the exact boundaries of duplication will need to be determined. Additional WBS-like duplicons were observed at 7q22.1, where 29 rearrangement breakpoints were mapped; 24 were involved in malignancy (Fig. 5). Finally, our observation that 12 of 35 (34.2%) of mouse synteny breaks coincide with a recent segmental duplication on human chromosome 7 ( $P < 0.0001$ ) (7) supports the idea synteny breaks (genomic rearrangements) do not always occur as random events (table S8).



**Fig. 4.** Recent segmental duplications on chromosome 7. Graphical views (using GenomePixelizer) of paralogous relationships between segmental duplications. Each line pairs two related sequences; red, 99 to 100% identity; purple, 96 to 98%; green, 93 to 95%; and blue, 90 to 92%. The size of segmental duplications (kb) is plotted against the length of chromosome 7.

**Medical annotation of the chromosome 7 DNA sequence.** To facilitate positional cloning studies and genotype-phenotype correlation, we have incorporated all medically relevant data into the DNA sequence map (supporting online text). From this ongoing initiative, we positioned 70 additional microsatellite markers, collated 1440 clinical karyotypes, and gathered numerous structural data, based on our FISH resource, that have been distributed worldwide. We have also cloned the *FRA7G*, *FRA7H*, and *FRA7E* fragile sites and have identified imprinted and differentially expressed genes (7). We studied rearrangement breakpoints from patients who have chromosome 7 anomalies and could place 440 on the sequence map (Fig. 5) (7). For more than 100 of these, molecular data were not available previously. Examples of new breakpoints mapped within a genomic region that contains a disease locus (gene not yet identified) associated with the patient's phenotype (e.g., acute myeloid leukemia at 7q22, cavernous malformation syndrome at 7p15, splenic lymphoma at 7q33) are summarized in table S9. Therefore, studies of the sequence at the rearrangement breakpoint(s) could provide insight into the regulation and function of genes, as is described below for three different developmental diseases.

Split-hand split-foot syndrome [Online Mendelian Inheritance of Man (OMIM) 183600], also known as ectrodactyly or lobster claw deformity, is a human developmental condition that is genetically heterogeneous and demonstrates variable expression of phenotype. These genetic characteristics are problematic in the clinical setting, making it impossible to predict carrier status and severity of the disease. Examples of such pedigrees (the first reported in 1908) were even used to argue against the applicability of Mendelian genetics to humans. Our analysis of chromosomal deletions in patients with the syndrome led to the mapping of a disease locus within a 1.2-Mb critical region at 7q21.3 (named *SHFM1*) (Fig. 6), but in a decade-long search, no disease gene could be identified. We positioned a balanced translocation or inversion breakpoint from 12 unrelated ectrodactyly patients and found them to be scattered throughout the critical region. The breakpoints did not collectively interrupt any single gene, which suggests a "position effect" mutation might be involved. Reports that double-null knockouts of two murine *Distalless* homeobox genes, *Dlx5* and *Dlx6*, exhibit ectrodactyly (25, 26) strongly implicate the human orthologs (which map within the *SHFM1* region). In the simplest explanation, the chromosomal breakpoints located up to 1 Mb centromeric of human *DLX5* and *DLX6* separate critical regulatory elements from the genes and lead to their dysregulation during development. From our characterization of other patients (represented in the Genome Browser), position-effect mutations are known for other developmental genes on chromosome 7: *GLI3*, *TWIST*, *SHH*, and

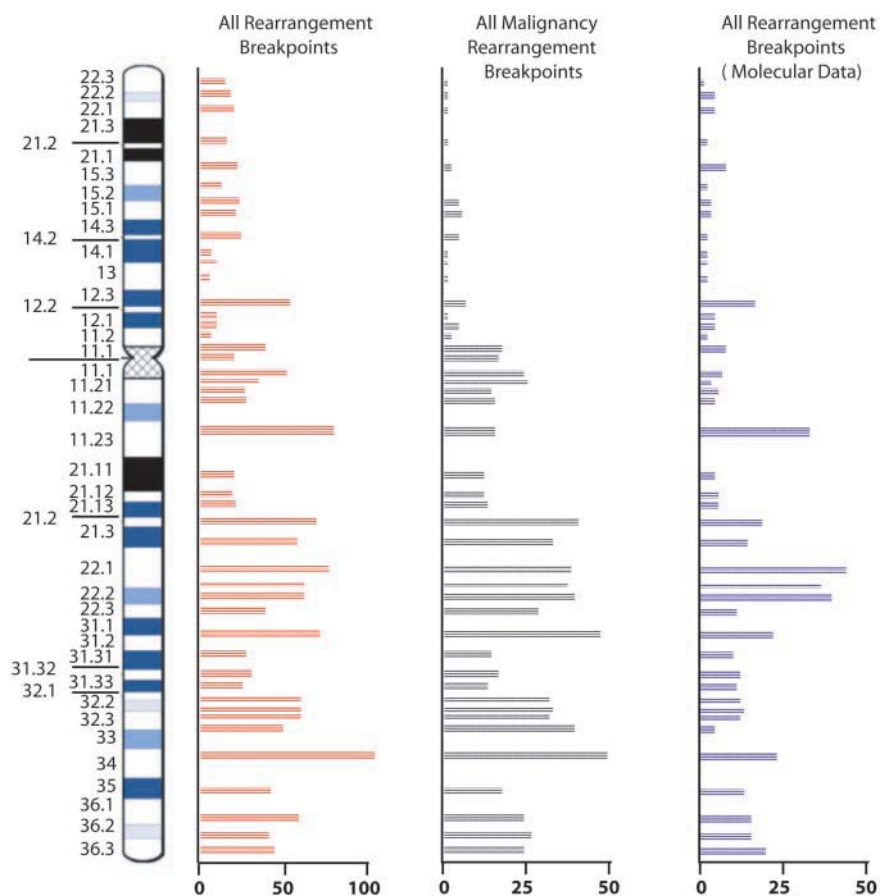
*CDK6* in Greig's cephalopolysyndactyly (up to 15 kb-3'), Saethre-Choetzen (up to 100 kb-5'), holoprosencephaly (up to 250 kb-5') and triphalangeal thumb (up to 1000 kb-5'), and splenic marginal zone lymphoma (up to 66 kb-5'), respectively. Further study of the breakpoints including comparative DNA analysis will guide experiments to identify candidate regulatory sequences for testing in functional assays.

We followed a line of investigation similar to that used to study chromosomal rearrangements in autism patients to assess candidate genes for the susceptibility locus (*AUTS1*) mapped to 7q (27). Using the sequence as a guide, we fine-mapped the chromosome 7 derivative-translocation breakpoints from three new autism patients within 7q22-q31 (the region demonstrating maximum linkage). The breakpoints in cases 16724, 18667, and 11550 (table S9) were positioned within BACs at 7q22.1, 7q31.2, and 7q31.3, respectively. The most interesting finding was that the breakpoint in case 11550 overlapped with a rearrangement from an unrelated patient (10893502\_2), diagnosed with speech and language disorder (a common component of the autism phenotype). The last breakpoint was anchored to the sequence by using data from the literature (28). Both breakpoints disrupt the

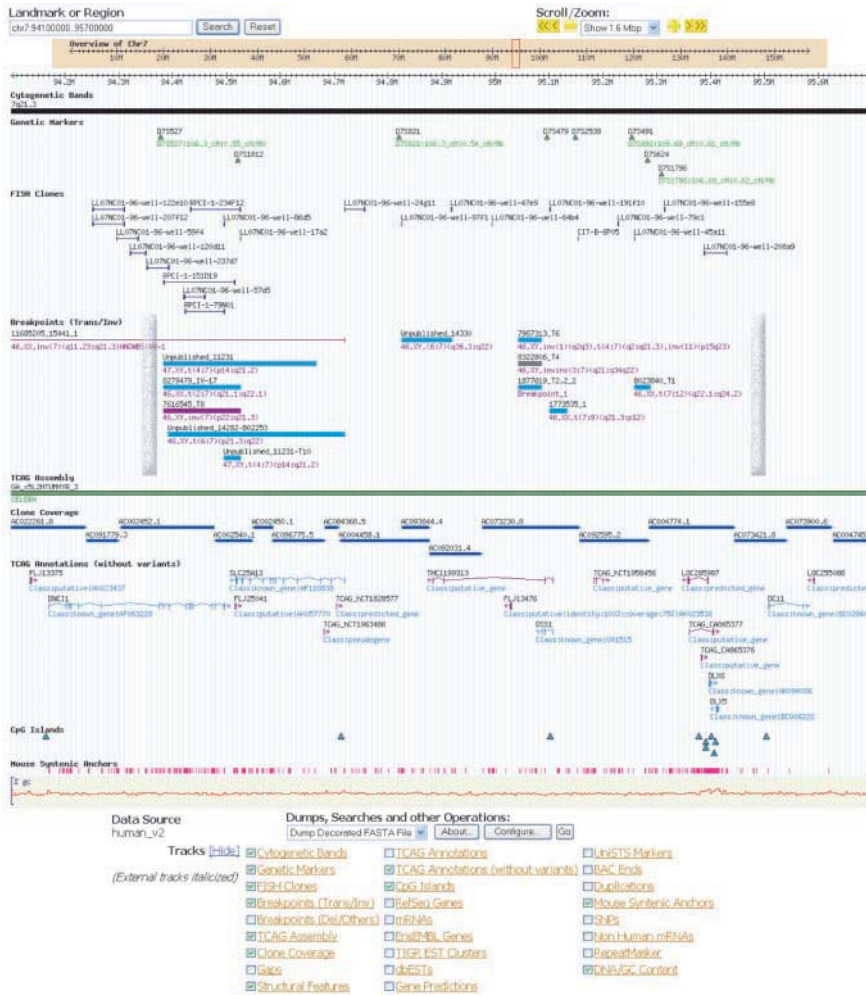
same, apparently noncoding, RNA transcript (TCAG\_4133353; GenBank CB338058), which is composed of at least 5 exons spanning 288 kb and is not found in the mouse genome.

For autism case 18667, the translocation breakpoint mapped near to the *FOXP2* gene, which was shown previously to cause a form of speech and language disorder (29). The child inherited the translocation from the mother, who had speech delay, which suggests that *FOXP2* might also be involved in autism. Seven isoforms of *FOXP2* (spanning 545 kb) were characterized, but all mapped at least 680 kb 3' of the translocation breakpoint (in a gene-poor region), once again raising the possibility of a position-effect mutation. Finally, in autism case 16724, the breakpoint was nearest to the neuronal pentraxin 2 (*NPTX2*) gene, which is thought to be involved in excitatory synaptogenesis and could, therefore, be considered a functional candidate for autism. The TCAG\_4133353, *FOXP2*, and *NPTX2* are now being examined for mutations in autism families.

In the third example, for two WBS patients (17430 and 16724) who do not carry the common 1.6-Mb hemizygous microdeletion found in 95% of affected individuals, we have discovered a new 500-kb inversion variant associated with the



**Fig. 5.** The distribution of 1570 cytogenetic rearrangement breakpoints (850 constitutional and 720 malignancy-associated) from 1440 patients with defined phenotypes; 440 rearrangement breakpoints have been characterized at the molecular level for disease identification studies (7).



**Fig. 6.** The *SHFM1* region at 7q21.3, as an example of information in the chromosome 7 Genome Browser. The positions of nine translocations, two inversions, and two breakpoints from an insertion (all from *SHFM1* patients) are shown to map within the minimal critical region defined by deletion breakpoints (vertical bars) (30). Besides the known genes (*DNC11*, *SLC25A13*, *DSS1*, *DLX5*, *DLX6*) (31), we identified two novel transcripts (TCAG\_CA865376 and TCAG\_CA865377) at the *DLX* locus, a longer variant of *DSS1* (THC1199313), and additional putative genes.

disease (supporting online text). Unlike the majority of other WBS deletions or inversions (10), the breakpoints in these individuals do not appear to be associated with segmental duplication-mediated events because they map to unique sequences. The *CYLN2* and *GTF2IRD1* genes and the *SCYA24* and *SCYA26* genes are closest to the centromeric and telomeric inversion breakpoints, respectively.

**Future studies of human chromosome 7 and community-based annotation.** Our goal of establishing a complete reference sequence benefited from data derived from both whole-genome shotgun and clone-based projects. Although some experimentation will be necessary to resolve minor discrepancies in the assembly, with the current framework in hand, the focus of work can be turned to confirming representation, testing for polymorphism, finalizing the gene map, and applying the information for disease study. For the last, we will continue to incorporate as much biomedical information as possible

into the DNA sequence map. As we demonstrated, the approach of studying chromosomal rearrangements en masse enabled rapid identification of candidate genes for monogenic and complex diseases and facilitated many functional and structural studies of chromosome 7.

Throughout our study, we found differences and inconsistencies between databases, arguing strongly for the need of additional community involvement in establishing and annotating the consensus sequence of the human genome. We have established a user-friendly database and organized our results into standardized files, in the spirit that this compilation of chromosome 7 information will not only be used as a primary source, but also be incorporated into other projects.

**References and Notes**

1. C. Dib et al., *Nature* **380**, 152 (1996).
2. J. Kunz et al., *Genomics* **22**, 439 (1994).
3. G. C. Bouffard et al., *Genome Res.* **7**, 59 (1997).
4. G. D. Schuler et al., *Science* **274**, 540 (1996).

5. J. C. Venter et al., *Science* **291**, 1304 (2001).
6. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
7. Materials and methods are available as supporting material on Science Online. The sequence assembly is at [www.chr7.org/](http://www.chr7.org/) and in the Third Party Annotation Section of the DDBJ/EMBL/GenBank databases under the accession number TPA: BL000001. The scaffolds are in DDBJ/EMBL/GenBank under the project accession number AACCC0000000. The version described in this paper is the first version, AACCC01000000. Individual accession numbers of the scaffolds are AACCC01000001, AACCC01000002, AACCC01000003, AACCC01000004, AACCC01000005, AACCC01000006, AACCC01000007, AACCC01000008, AACCC01000009, AACCC01000010, AACCC01000011, AACCC01000012, AACCC01000013, AACCC01000014, AACCC01000015, AACCC01000016, AACCC01000017, AACCC01000018, AACCC01000019, AACCC01000020, AACCC01000021, AACCC01000022, AACCC01000023, AACCC01000024, AACCC01000025, and AACCC01000026. The annotation data and analyses based on the CRA\_TCAGchr7.v1 assembly described in this paper are shown (and are archived) as the March 2003 database freeze (see [www.chr7.org/](http://www.chr7.org/)). Additional annotations or updates to the sequence assembly will be available as subsequent freezes. The Washington University Genome Sequencing Center has also produced an assembly and analysis of human chromosome 7 (L. Hillier et al., *Nature*, in press).
8. L. D. Stein et al., *Genome Res.* **10**, 1599 (2002).
9. L. R. Osborne et al., *Genomics* **45**, 402 (1997).
10. L. R. Osborne et al., *Nature Genet.* **29**, 321 (2001).
11. R. Wevrick, H. F. Willard, *Nucleic Acids Res.* **19**, 2295 (1991).
12. A. de la Puente et al., *Cytogenet. Cell Genet.* **83**, 176 (1998).
13. B. Ma, J. Tromp, M. Li, *Bioinformatics* **18**, 440 (2002).
14. R. J. Mural et al., *Science* **296**, 1661 (2002).
15. P. Pevzner, G. Tesler, *Genome Res.* **13**, 37 (2003).
16. Mouse Genome Sequencing Consortium, *Nature* **420**, 520 (2002).
17. The Fantom Consortium, *Nature* **420**, 523 (2002).
18. R. Heilig et al., *Nature* **421**, 601 (2003).
19. I. Dunham et al., *Nature* **402**, 489 (1999).
20. P. Deloukas et al., *Nature* **414**, 865 (2001).
21. M. Hattori et al., *Nature* **405**, 311 (2000).
22. G. R. Boocock et al., *Nature Genet.* **33**, 97 (2003).
23. J. Cheung et al., *Genome Biology* **4**, R25 (2003).
24. J. A. Bailey et al., *Science* **297**, 1003 (2002).
25. G. R. Merlo et al., *Genesis* **33**, 97 (2002).
26. R. F. Robledo, L. Rajan, X. Li, T. Lufkin, *Genes Dev.* **16**, 1089 (2002).
27. International Molecular Genetic Study of Autism Consortium, *Human Mol. Genet.* **3**, 571 (1998).
28. C. S. Lai et al., *Am. J. Hum. Genet.* **67**, 357 (2000).
29. C. S. Lai, S. E. Fisher, J. A. Hurst, F. Vargha-Khadem, A. P. Monaco, *Nature* **413**, 519 (2001).
30. S. W. Scherer et al., *Hum. Mol. Genet.* **3**, 1345 (1994).
31. K. Kobayashi et al., *Nature Genet.* **2**, 159 (1999).
32. We thank The Centre for Applied Genomics at The Hospital for Sick Children (HSC) and Celera Genomics, as well as clinical collaborators and families. Supported by Genome Canada, the Canadian Institutes of Health Research, the Canadian Genetic Diseases Network, the Howard Hughes Medical Institute International Scholar Program (to S.W.S.), and the HSC Foundation. We also thank groups worldwide for contributing genomic information to databases.

**Supporting Online Material**  
[www.sciencemag.org/cgi/content/full/1083423/DC1](http://www.sciencemag.org/cgi/content/full/1083423/DC1)  
 Materials and Methods  
 SOM Text  
 Figs. S1 and S2  
 Tables S1 to S9  
 References

13 February 2003; accepted 25 March 2003  
 Published online 10 April 2003;  
 10.1126/science.1083423  
 Include this information when citing this paper.

## Supporting Online Material, Scherer *et al.* 10.1126/science.1083423

### Material and Methods

*Fluorescence in situ hybridization (FISH) mapping of genomic clones.* Each of the 770 bacterial genomic clones shown in Fig. 2 was FISH mapped to cytogenetic bands in at least two different consortium laboratories by using published protocols (a total of 10 metaphases from at least four different individuals were typically tested) (S1). Over the last 5 years, representative clones from the FISH panel have also been sent to over 350 labs worldwide, and wherever possible, the mapping data were gathered and compared to our results. Overall, in 12 cases, we could not resolve inconsistencies between mapping data, so these clones were not included in the database. Much of the FISH mapping information for the yeast artificial chromosome (YAC) clones has been described previously (S2). Any new data were generated by using the same protocols. All genomic FISH probes described in this manuscript or in the chromosome 7 database are available upon request.

*Establishing synteny blocks and association with segmental duplications.* Syntenic blocks were defined as regions of orthologous DNA encompassed by at least 10 consecutive syntenic anchors (Table S5a) aligned in the same continuous orientation. A total of 35 breaks in synteny were identified, which constituted a jump of greater than 1 Mb in corresponding sequence alignment, a change in orthologous chromosome, or a change in orientation of alignment (inversion) with the neighboring syntenic block (Table S5b). For each synteny break region (within 5 kb of distance surrounding the break), we searched for the presence of large segmental duplications (each >50 kb in size; there are a total of 68 such duplications on chromosome 7) (Table S8). For measure of significance, we divided the chromosome into 50-kb blocks (3148 in total), and counted the number of blocks that contain synteny breaks (82), duplications (167), neither (2849), and both (50). A chi-square test was performed that yielded a value of 201 (with  $P < 0.0001$ ).

*Gene annotation.* We used a combined approach of computer-based gene prediction followed by laboratory experimentation to confirm gene models and extend transcripts. The data have been collected over a 10-year period, and the experiments were conducted in different stages. Our group effort has led to the initial description and characterization of over 100 full-length genes on chromosome 7 and sequencing of thousands of cDNAs. In the earliest stages, most of the information was generated from targeted positional cloning studies, and the cDNA or gene sequences were subsequently submitted to the databases. In the past 5 years, large-scale sequencing of cDNA libraries was completed, and some of these data were submitted to public databases. Other information is being released publicly for the first time. Full-length cDNA clones have been generated by screening conventional cDNA libraries and by using 5' and 3' extension protocols, as required. Also, wherever possible, comparative genomic DNA sequence data were used to guide our work. On a monthly basis, we test our most recent gene index of chromosome 7 for potential revision that might arise because of new additions from our sequencing efforts or from new sequences in the database. At The Centre for Applied Genomics (TCAG; <http://tcag.bioinfo.sickkids.on.ca/>), we used a DNA sequence annotation pipeline we developed named Genescript (S3). Genescript combines a compendium of analysis programs to perform EST clustering, gene predictions, homology, and comparative similarity searches to estimate gene models. We also used Celera's Otto gene annotation system [(S4); a detailed description of the gene annotation process is at <http://www.celera.com/>]. All of the data are displayed in appropriate tracks in the Genome Browser at <http://www.chr7.org>. Ultimately, scientists at TCAG and Celera manually examined every gene structure model that was

generated by computer algorithm and grouped them into the categories described in Table 1 of the manuscript. The selection of categories was modeled after the gene annotation projects for chromosomes 14 (S5) and 22 (S6). All of the data were also compared with the annotations of Ensembl, NCBI, and UCSC, as displayed in the chromosome 7 Genome Browser. We will continue to update the gene annotation records for chromosome 7 monthly.

*Defining CpG islands.* In all of our analyses in this study, CpG islands were detected by using EMBOSS *cpGREport* running default parameters and a stringent cutoff of 200.

*Imprinted genes and fragile sites.* To identify the genes on human chromosome 7 that are differentially expressed from parental chromosomes (imprinted), we used various techniques including somatic cell hybrid expression, methylation-sensitive restriction enzyme analysis, and bisulfite sequencing. The imprinting status of any candidate gene identified by these approaches was confirmed in fetal tissue expression assays. The scientific protocols for these experiments are detailed in a previous manuscript (S7). As part of our study, we have identified the *MEST11* [paternally expressed (S7)], *CPA4* (maternally expressed, new data), and *PPPIR9A* (maternally expressed, new data) genes to be imprinted in certain tissues. We have also been examining differential expression patterns of chromosome 7 genes to search for different patterns of parent-of-origin methylation (S8). In a long-term project, we have been working to map each of the nine fragile sites on chromosome 7 at the level of the DNA sequence. The protocols have been described in detail (S9, S10). So far, *FRA7E* (new data), *FRA7G* (S9), *FRA7H* (S10), and *FRA7I* (S11) are characterized at the molecular level. All of the data mentioned above are represented as a feature of the chromosome 7 sequence in the Genome Browser (<http://www.chr7.org>).

*Detection of new segmental duplications.* In a study of patients with Williams-Beuren syndrome (WBS), we had detected a genomic region (starting at marker *D7S1440*) just telomeric to the WBS microdeletion region that, on the basis of DNA sequence analysis alone, would have been predicted to be unique in the human genome. FISH analysis with BACs (e.g., RP11-229D13) encompassing this region consistently gave two signals beside each other at 7q11.23 (in metaphase and interphase analyses). Subsequent experiments with additional BAC probes extending telomeric (RP11-1129E22, RP11-275G11, RP11-441N19, RP11-792H4) also gave two hybridization signals when we used metaphase and interphase FISH in 10 control (non-WBS) samples. All of our data suggest that the region between *D7S1440* and *D7S1478* (~1 Mb) is part of a larger, previously undetected, segmental duplication at 7q11.23. Subsequent analysis will be required to confirm the finding and to determine the precise boundaries, at which time the sequence will be added to the assembly.

*Mapping of chromosome breakpoints.* The mapping of chromosome 7 breakpoints followed several different courses depending on the laboratory involved and the sample being examined. In general, microsatellite markers were tested against patient and parental DNA samples (whenever available). After narrowing the breakpoint regions, FISH experimentation was performed by using the protocols described above. For translocations and inversions YACs and BACs were tested directly by FISH. When somatic cell hybrids were available, probing and blot-hybridization or PCR was performed. In some cases, the breakpoint was cloned and sequenced. Detailed clinical information on all cases studied is available at the chromosome 7 Web site. Samples from the Developmental Genome Anatomy Project (DGAP) can also be found at [www.bwhpathology.org/dgap](http://www.bwhpathology.org/dgap) (funded by NIH GM61354 to C.C.M.).

## Supplementary (SOM )Text

*DNA sequence assembly and characterization.* As a backbone, 26 scaffolds from the previously unpublished Celera whole-genome assembly (HR26) were used. The HR26 sequence is available by subscription, see <http://www.celera.com/>; as part of this publication the HR26 chromosome 7 scaffold sequences are being released publicly. The HR26 sequence is a component assembly (Celera components are sets of regional sequences that are linked together through overlaps and mate-pairs) that represents an update to the assembly presented previously (*S4*), with the scaffold N50 improved from 2.96 to 14.37 Mb. These scaffolds, which encompassed 133,269,991 bp, were selected because our experimental analysis (Fig. S1) and others (*S12*) had determined they contained high-quality sequence with exceptional internal marker order consistency and coverage. To extend scaffolds, we took end sequences and tested them in an iterative manner against all other Celera data and the public nonredundant and high-throughput genomic sequence databases. At loci containing large segmental duplications (e.g., 7p22, 7cen, 7q11.23, 7q22, and 7q36; see below) or clustered highly related gene family members (e.g., the T cell receptor loci, mucin genes), we relied exclusively on clone-based sequence to provide the most representative coverage. Members of our group had previously mapped or sequenced or already annotated many of these regions [e.g. (*S9*, *S10*, *S13–S17*)]. In total, 209 genomic clones covering 21,781,798 nt were used. After merging all data, 1509 gaps remained (median size 255 bp). In the next step, library screening, PCR-based sequencing, and database comparisons resulted in an addition of 2,209,832 bp that resulted in a composite sequence assembly encompassing 157,953,789 nt of chromosome 7 assembled together by nonredundant concatenation of overlapping sequences (Table S2). Table S3 shows all components used in the assembly. In the end, for 138 small intrascaffold (estimated average size of 1.9 kb) and six physical gaps (two at 7p22 and 7q11.23, one each at 7q22 and 7q36), sequences have not yet been obtained (Tables S2 and S3); the sites of the gaps are displayed in the Genome Browser. At the time of analysis, we could not find any additional sequences in other databases to fill these gaps. Additional BAC and cosmid library screenings were completed to increase clone coverage, especially around segmental duplication regions and the centromere. Clones were end-sequenced to confirm identity and, subsequently, were used as a template for PCR or cloning. The final assembly of chromosome 7 presented in this manuscript was named CRA\_TCAGchr7.v1. We intend to continue to examine this assembly and if there are changes, they will be released regularly at the Chromosome 7 Web site. Any new sequences generated will also continue to be submitted to GenBank.

To test for representation and accuracy, we continually tested all genomic data available against the assembly. In our initial analysis of the sequence using electronic PCR, we could find 6445 markers (6012 were unique) known to be on chromosome 7, but 76 were missing. We examined these closer, and when accounting for potential primer-sequence polymorphisms, we could identify the remaining markers. We then tested the 863 known genes on chromosome 7 and were able to detect each one along the chromosome in the order predicted from the mRNA sequence. We also compared marker orders with 246 microsatellites from the deCODE genetic map and only found the marker combination *D7S1795* (37.6 cM) to *D7S2458* (37.38 cM), which are separated by 16.5 kb, to be inverted between the two maps. The deCODE genetic data for each microsatellite marker is displayed in the chromosome 7 Genome Browser. Finally, using pair-wise marker-to-marker and nucleotide-to-nucleotide analysis, we examined CRA\_TCAGchr7.v1 against each release of the NCBI's chromosome 7 sequence, which was based on sequences originating from the hierarchical, genomic clone–sequencing strategy. As

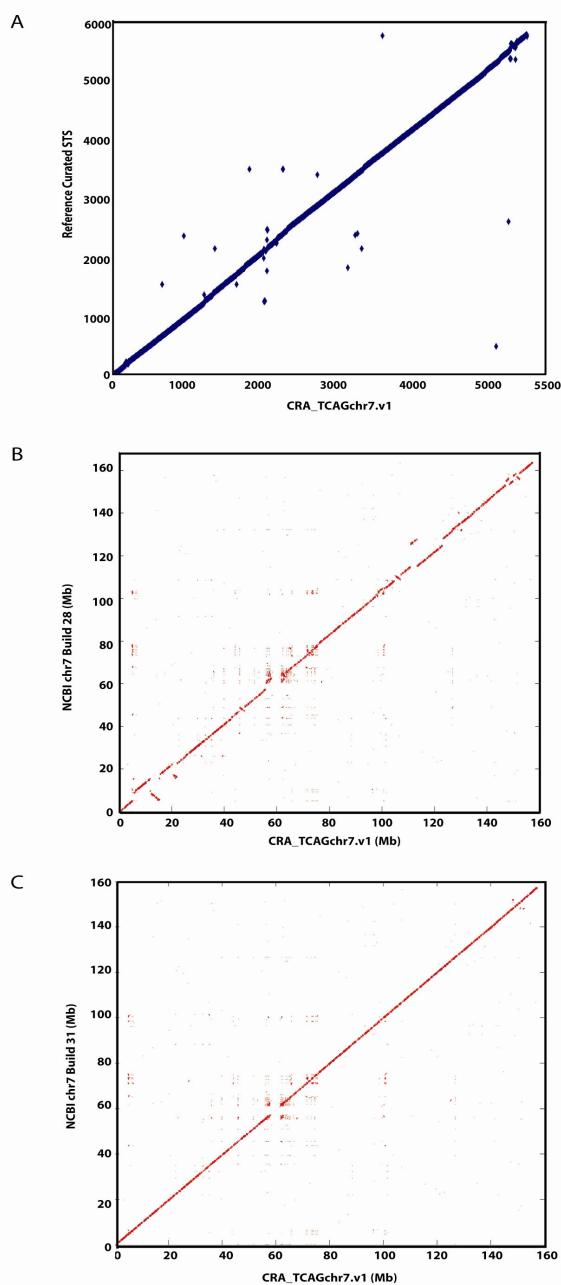
shown in Fig. S1, the more recent NCBI builds (e.g., NCBI Build 31) bear closer resemblance to our assembly, but there were differences.

We performed sequence comparisons of the different assemblies using the PatternHunter program (*S18*), which allowed the identification of unmatched sequences, sequence variations, and sequence inversions (summary in Table S4; complete dataset can be obtained at <http://www.chr7.org>). A total of 1,022,295 nt were identified in CRA\_TCAGchr7.v1 (the category was termed 'unmatched') that were not found in NCBI Build 31. In most cases, when a sequence was present in CRA\_TCAGchr7.v1, but missing from the public data, we could confirm its chromosome 7 origin on the basis of the sequence tagged sites (STSs) it contained. In other cases, we developed new STSs and performed PCR mapping against somatic cell hybrids to confirm the chromosome 7 origin (Fig. S2). These STSs have been and will continue to be submitted to the databases as landmarks. STS markers BV012541, BV006765, BV006768, BV006769, BV006770, BV006771, BV012528, BV012540, BV012529, BV012530, BV012531, BV007444, BV006773, BV006774, BV006795, BV007445, BV012532, BV006775, BV006792, BV012533, BV006796 have so far been submitted to GenBank. Examples of larger segments missing from NCBI Build 31 at the time of analysis include 74 kb, 304 kb, 104 kb, and 108 kb at the *PRKAR1B*, *PDGFA*, *ZNF1A1*, and *UP* gene loci, respectively (Fig. 1). About 164 kb present in the NCBI assembly could not be found in CRA\_TCAGchr7.v1, and we are examining this sequence further. At 132 other sites (encompassing 508 kb), when we compared the CRA\_TCAGchr7.v1 assembly to Build 30, different sequences (termed sequence variations) were found at the same relative chromosomal positions (Table S4). Finally, 10 equivalent DNA segments were positioned in an inverted orientation between the two assemblies (Fig. 1). The inverted regions did not correspond to the locus discrepant with the deCODE map nor to any segment of chromosome 7 so far known to demonstrate genomic polymorphism. As discussed in the manuscript, differences detected between the two assemblies (highlighted in the chromosome 7 Genome Browser) could be due to polymorphism in the source chromosomal DNA, rearrangements introduced during the cloning process, or improper assembly. It is also important to note that although all data suggest that the CRA\_TCAGchr7.v1 assembly is an accurate description, there could be undetected anomalies such as large, nearly identical stretches of DNA present in two or more copies on chromosome 7 that are currently only represented as a single locus. The multi-step strategy we followed, however, should have allowed us to find at the start the majority of the problem regions on chromosome 7 or discrepancies with other sequence assemblies. We are currently using interphase FISH and genomic microarray analysis to scan for previously undetected segmental duplications.

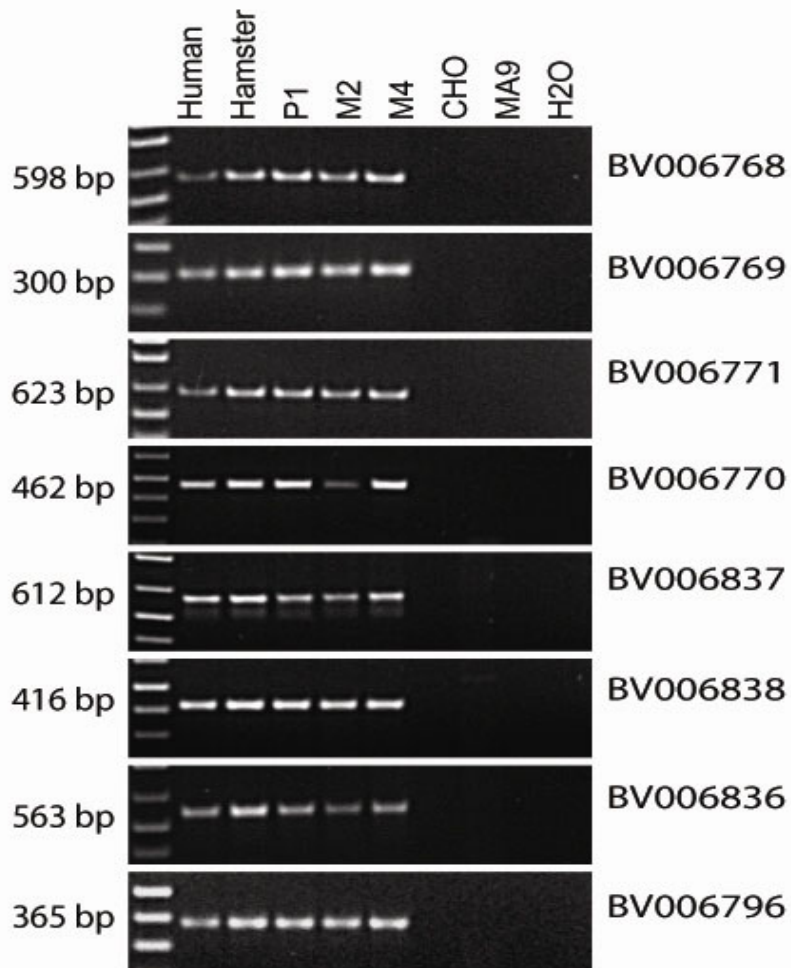
*Medical annotation of the chromosome 7 DNA sequence.* We are reviewing every paper published on chromosome 7 and extracting information on molecular markers (that were not previously submitted to databases), clinical cases having a defined phenotype with chromosome 7 rearrangements, genomic variation, and other biological phenomena. To assist in extrapolating information generated before the PCR era, we converted restriction fragment length polymorphism markers and probes into STSs, which allowed us to anchor them to the sequence. We also determined end sequences of uncharacterized genomic clones used in FISH-based chromosome analyses. This curation process will continue and we are accepting data submissions from other groups. The characterization of chromosomal breakpoints was part of targeted studies aimed at specific disease genes, as well as more general projects examining chromosome 7 genotype-phenotype correlation at the chromosomal level. We are also continuing to characterize breakpoints from the existing cases in the database and are accepting new ones. New information from all sources will be updated regularly.

*Williams-Beuren syndrome and new inversion variants.* Williams-Beuren syndrome (WBS) is developmental condition that arises from the hemizygous deletion (*S19*) of 1.6 Mb of DNA encompassing at least 23 genes on human chromosome 7q11.23. The WBS phenotype, occurring in 1 of every 20,000 individuals worldwide, includes congenital vascular and heart disease, dysmorphic facies, growth deficiency, infantile hypercalcemia, mental retardation, unique cognitive profile, and a characteristic personality. Diagnosis of WBS includes testing for hemizyosity at 7q11.23 by FISH with a probe encompassing the *ELN* gene. In more than 95% of cases, there is a defined 1.6-Mb deletion, but for the remaining individuals, no cytogenetically detectable chromosome rearrangement was known. Recently, we identified that inversions of the same genomic interval usually deleted in WBS individuals can also be associated with the disease (*S17*). Moreover, in ~30% of parents with a WBS child having a deletion the same inversion is found on the disease-transmitting chromosome. The common microdeletion (*S20*) and inversions (*S17*) arise because of unequal meiotic recombination, mediated by the highly homologous segmental duplications flanking the region. In the current study, we identify a new inversion variant in two unrelated WBS patients (17430 and 16724) who had apparently normal chromosomes upon standard cytogenetic analysis. As shown in Table S9 and the Genome Browser, the breakpoints in both of these inversions are ~500 kb in size and occur between *D7S613/D7S1870* and *D7S2490/D7S1440*. The *CYLN2* and *GTF2IRD1* genes and the *SCYA24* and *SCYA26* genes are closest to the centromeric and telomeric inversion breakpoints, respectively. The breakpoints in this inversion reside in apparently unique sequences, which suggests that the mechanism of rearrangement is not mediated by repeats. We have not observed the inversion in controls (50 unrelated from non-WBS families).

## Supplementary Figures



**Fig. S1.** Assessment of the CRA\_TCAFchr7.v1 chromosome 7 DNA sequence assembly. **(A)** The order of 5344 chromosome 7 DNA markers present in the CRA\_TCAGchr7.v1 assembly was almost entirely consistent with the marker order established by hand-curated data derived from radiation and somatic cell hybrid, YAC and BAC, and genetic mapping experiments. **(B)** and **(C)** Dot-plot analysis of the nucleotide sequence of CRA\_TCAGchr7.v1 versus NCBI Build 28 (December 2001) and 31 (November 2002). It is important to note that the nucleotide and clone content for chromosome 7 between NCBI Builds 30 and 31 are essentially unchanged. Each plot was generated by using MegaBLAST, and all hits with e-values less than  $10^{-120}$  are shown. Misoriented lines represent reversed sequence. Outlying dots can represent repeats and small sequence rearrangements. Gaps in lines represent unmatched sequences.



**Fig. S2.** Validation of the origin of DNA sequences from chromosome 7. Loci-specific PCR primers from eight regions along chromosome 7 from the CRA\_TCAGchr7.v1 assembly not represented in NCBI Build 31 were tested against DNA from human, human chromosome 7 hamster (CHO) hybrid, mouse hybrid containing a paternal (P1) or maternal (M2, M4) human chromosome 7, hamster (CHO), and mouse A9. The accession numbers of the DNA sequence being amplified are shown on the right, and the amplified product size is on the left.

## Supplementary Tables

**Table S1.** Disease genes and loci on human chromosome 7. The table includes all information from OMIM (<http://www.ncbi.nlm.nih.gov/Omim/>), as well as additional information from our group and that gathered from searches of the literature. The table will be regularly updated and shown at <http://www.chr7.org>.

Gene #	Symbol	Full name	OMIM#	Disorder	Location
1	AASS	Aminoacidase-semialdehyde synthase	605113	Saccharopinuria, 268700; Hyperlysinemia, 238700	7q31.3
2	ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1	171050	Colchicine resistance	7q21.1
3	ABCB4	ATP-binding cassette, sub-family B (MDR/TAP), member 4	171060	Cholestasis, progressive familial intrahepatic 3, 602347	7q21.1
4	ABCB8	ATP-binding cassette, sub-family B (MDR/TAP), member 8	605464		7q36
5	ABP1	Amiloride binding protein 1 (amine oxidase (copper-containing))	104610		7q34-q36
6	ACHE	Acetylcholinesterase (YT blood group)	100740	Acute sensitivity to anti-acetylcholinesterases [OMIM:100740]	7q22
7	ACRPS	Acropectoral syndrome	605967	Acropectoral syndrome	7q36
8	ACTB	Actin, beta	102630		7p15-p12
9	ADAM22	A disintegrin and metalloproteinase domain 22	603709		7q21
10	ADCY1	Adenylate cyclase 1 (brain)	103072		7p13-p12
11	ADCYAP1R1	Adenylate cyclase-activating polypeptide 1 (pituitary) receptor type I	102981		7p14
12	AGR2	Anterior gradient 2 homolog ( <i>Xenopus laevis</i> )	606358		7p21.3
13	AHR	Aryl hydrocarbon receptor	600253		7p15
14	AIB	Aneurysm, intracranial berry	105800	Aneurysm, intracranial berry	7q11.2
15	AIP1	Atrophin-1-interacting protein 1; activin receptor-interacting protein 1	606382		7q21
16	AKR1B1	Aldo-keto reductase family 1, member B1 (aldose reductase)	103880		7q35
17	AMPH	Amphiphysin (stiff-man syndrome with breast cancer 128-kD autoantigen)	600418	Paraneoplastic stiff-man syndrome [OMIM:600418]	7p14-p13
18	AOAH	Acyloxyacyl hydrolase (neutrophil)	102593		7p14-p12
19	AP1S1	Adaptor-related protein complex 1, sigma 1 subunit	603531		7q11.22
20	APS	Adaptor protein with pleckstrin homology and Src homology 2 domains	605300		7q22
21	AQP1	Aquaporin 1 (channel-forming integral protein, 28 kD)	107776	Blood group-Coltan, 110450; [Aquaporin-1 deficiency]	7p14
22	ARF5	ADP-ribosylation factor 5	103188		7q31.3
23	ARHGEF5	Rho guanine nucleotide exchange factor (GEF) 5	600888		7q33-q35 7p21-
24	ARL4	ADP-ribosylation factor-like 4	604786		p15.3
25	ASB4	Ankyrin repeat and SOCS box-containing 4	605761		7q21-q22
26	ASK	Activator of S-phase kinase	604281		7q21.3 7cen-
27	ASL	Argininosuccinate lyase	207900	Argininosuccinicaciduria	q11.2
28	ASNS	Asparagine synthetase	108370		7q21.3
29	ATP6V0A4	ATPase, H <sup>+</sup> transporting, lysosomal V0 subunit a isoform 4	605239	Renal tubular acidosis, distal, autosomal recessive, 602722	7q33-q34
30	AUTS1	Autism susceptibility 1	209850	{Autism, susceptibility to}	7q
31	AZGP1	Alpha-2-glycoprotein 1, zinc	194460		7q22.1
32	BAZ1B	Bromodomain adjacent to zinc finger domain, 1B	605681		7q11.23

Gene #	Symbol	Full name	OMIM#	Disorder	Location
33	BCL7B	B-cell CLL/lymphoma 7B	605846		7q11.23
34	BLVRA	Biliverdin reductase A	109750		7p14-cen
35	BMIQ1	Body mass index QTL on chromosome 7	606642	Body mass index, 606641	7q32.3
36	BPGM	2,3-Bisphosphoglycerate mutase	222800	Hemolytic anemia due to bisphosphoglycerate mutase deficiency	7q31-q34
37	BRAF	V-raf murine sarcoma viral oncogene homolog B1	164757		7q34
38	C7orf2	Chromosome 7 open reading frame 2	605522	Acheiropody, 200500	7q36
39	CACNA2D1	Calcium channel, voltage-dependent, alpha 2/delta subunit 1	114204		7q21-q22
40	CALCR	Calcitonin receptor	114131	Osteoporosis, involutional, 166710	7q21.3
41	CALD1	Caldesmon 1	114213		7q33
42	CALML1	Calmodulin-like 1	114181		7pter-p13
43	CALU	Calumenin	603420		7q32 7q31.2- q31.3
44	CAPZA2	Capping protein (actin filament) muscle Z-line, alpha 2	601571		7p15-p14
45	CARD4	Caspase recruitment domain family, member 4	605980		7q34-q35
46	CASP2	Caspase 2, apoptosis-related cysteine protease	600639		7q32
47	CATR1	CATR tumorigenicity conversion 1	600676		7q31.1
48	CAV1	Caveolin 1, caveolae protein, 22 kD	601047		7q31.1
49	CAV2	Caveolin 2	601048		7q31.1
50	CCM1	Cerebral cavernous malformations 1	604214	Cerebral cavernous malformations 1, 116860	7q21-q22
51	CCM2	Cerebral cavernous malformation 2	603284	Cerebral cavernous malformations-2	7p15-p13
52	CD36	CD36 antigen (collagen type I receptor, thrombospondin receptor)	173510	Platelet glycoprotein IV deficiency; [Macrothrombocytopenia]	7q11.2
53	CDK5	Cyclin-dependent kinase 5	123831		7q36
54	CDK6	Cyclin-dependent kinase 6	603368	Splenic lymphoma [OMIM:603368]	7q21-q22
55	CFTR	Cystic fibrosis transmembrane conductance regulator, ATP-binding cassette	602421	Vas deferens, congenital bilateral aplasia of; Cystic fibrosis; Sweat chloride elevation without CF; {Hypertrypsinemia, neonatal} ; {Pancreatitis, idiopathic}	7q31.2
56	CGRP-RCP	Calcitonin gene-related peptide-receptor component protein	606121		7p13
57	CHDM	Chordoma	215400	Chordoma	7q33
58	CHN2	Chimerin (chimaerin) 2	602857		7p15.3
59	CHRM2	Cholinergic receptor, muscarinic 2	118493		7q31-q35
60	CLCN1	Chloride channel 1, skeletal muscle (Thomsen disease, autosomal dominant)	118425	Myotonia, generalized, Myotonia congenita, autosomal dominant, Myotonia levior, recessive	7q35
61	CLDN3	Claudin 3	602910		7q11.23
62	CLECSF5	C-type (calcium dependent, carbohydrate-recognition domain) lectin, superfamily member 5	604987		7q33
63	CMH6	Cardiomyopathy, hypertrophic 6	600858	Cardiomyopathy, familial hypertrophic with Wolff-Parkinson-White syndrome	7q31-qter
64	CMT2D	Charcot-Marie-Tooth neuropathy 2D	601472	Charcot-Marie-Tooth neuropathy, type 2D	7p14
65	CMT2F	Charcot-Marie-Tooth disease, axonal, F	606595	Charcot-Marie-Tooth disease, type 2F	7q11-q21
66	CNTNAP2	Contactin associated protein-like 2	604569		7q35-q36
67	COG5	Component of oligomeric Golgi complex 5	606821		7q31
68	COL1A2	Collagen, type I, alpha 2	120160	Osteoporosis, involutional, Ehlers-Danlos syndrome, type VII, autosomal dominant, Marfan syndrome, atypical	7q22.1
69	COPG2	Coatmer protein complex, subunit gamma 2	604355		7q32
70	CPA1	Carboxypeptidase A1 (pancreatic)	114850		7q32

Gene #	Symbol	Full name	OMIM#	Disorder	Location
71	CPA2	Carboxypeptidase A2 (pancreatic)	600688		7q32
72	CPSF4	Cleavage and polyadenylation specific factor 4, 30-kD subunit	603052		7q22.1
73	CRHR2	Corticotropin-releasing hormone receptor 2	602034		7p15.3
74	CRIP1	Cysteine-rich protein 1 (intestinal)	123875		7q11.23
75	CROT	Carnitine O-octanoyltransferase	606090		7q21.1
76	CRS	Craniosynostosis	123100	Craniosynostosis, type 1	7p21
77	CUTL1	Cut-like 1, CCAAT displacement protein ( <i>Drosophila</i> )	116896		7q22
78	CYLN2	Cytoplasmic linker 2	603432		7q11.23
79	CYP3A4	Cytochrome P450, subfamily IIIA (niphedipine oxidase), polypeptide 4	124010		7q21.1
80	CYP3A43	Cytochrome P450, subfamily IIIA, polypeptide 43	606534		7q21.1
81	CYP3A5	Cytochrome P450, subfamily IIIA (niphedipine oxidase), polypeptide 5	605325		7q21.1
82	CYP51	Cytochrome P450, 51 (lanosterol 14-alpha-demethylase)	601637		7q21.2-q21.3
83	DDC	Dopa decarboxylase (aromatic L-amino acid decarboxylase)	107930		7p11
84	DFNA5	Deafness, autosomal dominant 5	600994	Deafness, autosomal dominant 5	7p15
85	DFNB13	Deafness, autosomal recessive 13	603098	Deafness, autosomal recessive 13	7q34-q36
86	DFNB14	Deafness, autosomal recessive 14	603678	Deafness, autosomal recessive 14	7q31
87	DFNB17	Deafness, autosomal recessive 17	603010	Deafness, autosomal recessive 17	7q31
88	DGKB	Diacylglycerol kinase, beta (90 kD)	604070		7p22.1
89	DGKI	Diacylglycerol kinase, iota	604072		7q32.3-q33
90	DIA2	Diaphorase-2	125870		7
91	DLD	Dihydrolipoamide dehydrogenase	246900	Lipoamide dehydrogenase deficiency	7q31-q32
92	DLX5	Distal-less homeo box 5	600028		7q22
93	DLX6	Distal-less homeo box 6	600030		7q22
94	DNAH11	Dynein, axonemal, heavy polypeptide 11	603339		7p21
95	DNCL1	Dynein, cytoplasmic, intermediate polypeptide 1	603772		7q21.3-q22.1
96	DNHBL	Dynein, heavy chain beta-like	603299		7p15
97	DPP6	Dipeptidylpeptidase VI	126141		7q36.1-q36.2
98	DSS1	Deleted in split-hand/split-foot 1 region	601285		7q21.3-q22.1
99	EEC1	Ectrodactyly, ectodermal dysplasia and cleft lip/palate syndrome 1	129900	EEC syndrome-1	7q11.2-q21.3
100	EGFR	Epidermal growth factor receptor	131550		7p12
101	ELMO1	Engulfment and cell motility 1 (ced-12 homolog, <i>C. elegans</i> )	606420		7p15.3-p15.2
102	ELN	Elastin (supravalvular aortic stenosis, Williams-Beuren syndrome)	130160	Williams-Beuren syndrome, Supravalvular aortic stenosis, Cutis laxa,	7q11.23
103	EN2	Engrailed homolog 2	131310		7q36
104	EPHA1	EphA1	179610		7q32-q36
105	EPHB4	EphB4	600011		7q22
106	EPHB6	EphB6	602757		7q33-q35
107	EPIM	Epimorphin	132350		7
108	EPO	Erythropoietin	133170	Erythremia	7q22

Gene #	Symbol	Full name	OMIM#	Disorder	Location
109	ERV3	Endogenous retroviral sequence 3 (includes zinc finger protein H-plk/HPF9)	131170		7p11.2
110	ERVK6	Endogenous retroviral sequence K, 6	605626		7
111	ERVWE1	Endogenous retroviral family W, env(C7), member 1 (syncytin)	604659		7q21-q22
112	ETV1	Ets variant gene 1	600541		7p22
113	EVX1	Eve, even-skipped homeo box homolog 1 ( <i>Drosophila</i> )	142996		7p15-p14
114	EZH2	Enhancer of zeste homolog 2 ( <i>Drosophila</i> )	601573		7q35-q36
115	FDPSL2	Farnesyl diphosphate synthase-like 2 (farnesyl pyrophosphate synthetase-like 2)	134632		7
116	FGL2	Fibrinogen-like 2	605351		7q11.23
117	FHA2	Hyperaldosteronism, familial, type II	605635	Hyperaldosteronism, familial, type II	7p22
118	FKBP6	FK506-binding protein 6 (36 kD)	604839		7q11.23
119	FLNC	Filamin C, gamma (actin-binding protein 280)	102565		7q32-q35
120	FOXP2	Forkhead box P2	605317	Specific language impairment, 602081	7q31
121	FTSJ2	FtsJ homolog 2 (E. coli)	606906		7p22
122	FZD1	Frizzled homolog 1 ( <i>Drosophila</i> )	603408		7q21
123	FZD9	Frizzled homolog 9 ( <i>Drosophila</i> )	601766		7q11.23
124	G7P1	Kinase-like protein	148750		7q22-q32
125	GABPB1	GA-binding protein transcription factor, beta subunit 1 (53 kD)	600610		7q11.2
126	GARS	Glycyl-tRNA synthetase	600287		7p15
127	GBAS	Glioblastoma amplified sequence	603004		7p12
128	GBX1	Gastrulation brain homeo box 1	603354		7q36
129	GCF1	Growth control factor 1	139220		7
130	GCK	Glucokinase (hexokinase 4, maturity onset diabetes of the young 2)	138079	Hyperinsulinism, Maturity-onset diabetes of the young, type II,	7p15.3-p15.1
131	GCTG	Gamma-glutamylcyclotransferase	137170		7pter-p14
132	GHRHR	Growth hormone-releasing hormone receptor	139191	Growth hormone deficient dwarfism	7p14
133	GLC1F	Glaucoma 1, open angle, F (adult-onset)	603383	Glaucoma 1F	7q35-q36
134	GLI3	GLI-Kruppel family member GLI3 (Greig cephalopolysyndactyly syndrome)	165240	Greig cephalopolysyndactyly syndrome, Polydactyly, preaxial IV, Polydactyly, postaxial, type a1, Pallister-Hall syndrome,	7p13
135	GNAI1	Guanine nucleotide-binding protein (G protein), alpha inhibiting activity polypeptide 1	139310		7q21
136	GNB2	Guanine nucleotide-binding protein (G protein), beta polypeptide 2	139390		7q22
137	GNGT1	Guanine nucleotide-binding protein (G protein), gamma-transducing activity polypeptide 1	189970		7q21.3
138	GPDS1	Glaucoma-related pigment dispersion syndrome 1	600510	Pigment dispersion syndrome	7q35-q36
139	GPR22	G protein-coupled receptor 22	601910		7q22-q31.1
140	GPR30	G protein-coupled receptor 30	601805		7p22
141	GPR37	G protein-coupled receptor 37 (endothelin receptor type B-like)	602583		7q31

Gene #	Symbol	Full name	OMIM#	Disorder	Location
142	GPR85	G protein-coupled receptor 85	605188		7q31
143	GRB10	Growth factor receptor-bound protein 10	601523		7p12-p11.2
144	GRM3	Glutamate receptor, metabotropic 3	601115		7q21.1-q21.2
145	GRM8	Glutamate receptor, metabotropic 8	601116		7q31.3-q32.1
146	GSBS	G-substrate	604088		7p15
147	GTF2I	General transcription factor II, i	601679		7q11.23
148	GTF2IRD1	GTF2I repeat domain-containing 1	604318		7q11.23
149	GUSB	Glucuronidase, beta	253220	Mucopolysaccharidosis VII	7q21.11
150	HAKAI	Hypothetical protein FLJ23109	606872		7q21.11
151	HDAC9	Histone deacetylase 9	606543		7p21-p15
152	HGF	Hepatocyte growth factor (hepapoietin A; scatter factor)	142409		7q21.1
153	HIP1	Huntingtin-interacting protein 1	601767		7q11.23
154	HIPK2	Homeodomain-interacting protein kinase 2	606868		7q32-q34
155	HLXB9	Homeobox HB9	176450	Currarino syndrome, 176450; Sacral agenesis-1	7q36
156	HNRPA2B1	Heterogeneous nuclear ribonucleoprotein A2/B1	600124		7p15
157	HOXA1	Homeobox A1	142955		7p15.3
158	HOXA10	Homeobox A10	142957		7p15-p14
159	HOXA11	Homeobox A11	142958	Radioulnar synostosis with amegakaryocytic thrombocytopenia, 605432	7p15-p14
160	HOXA13	Homeobox A13	142959	Hand-foot-uterus syndrome, 140000; Guttacher syndrome, 176305	7p15-p14
161	HOXA3	Homeobox A3	142954		7p15-p14
162	HOXA4	Homeobox A4	142953		7p15-p14
163	HOXA5	Homeobox A5	142952		7p15-p14
164	HOXA6	Homeobox A6	142951		7p15-p14
165	HOXA7	Homeobox A7	142950		7p15-p14
166	HOXA9	Homeobox A9	142956		7p15-p14
167	HPFH2	Hereditary persistence of fetal hemoglobin, heterocellular, Indian type	142335	Hereditary persistence of fetal hemoglobin, heterocellular, Indian-type	7q36
168	HPVC1	Human papillomavirus (type 18) E5 central sequence-like 1	600762		7p14-p13
169	HRX	Hyperreflexia	145290		7q
170	HSPB1	Heat shock 27-kD protein 1	602195		7p12.3
171	HTR5A	5-Hydroxytryptamine (serotonin) receptor 5A	601305		7q36.1
172	HUS1	HUS1 checkpoint homolog (S. pombe)	603760		7p13-p12
173	HYAL4	Hyaluronoglucosaminidase 4	604510		7q31.3
174	ICA1	Islet cell autoantigen 1 (69 kD)	147625		7p22
175	IFRD1	Interferon-related developmental regulator 1	603502		7q22-q31
176	IGFBP1	Insulin-like growth factor-binding protein 1	146730		7p13-p12
177	IGFBP3	Insulin-like growth factor-binding protein 3	146732		7p13-p12
178	IL6	Interleukin 6 (interferon, beta 2)	147620	Osteoporosis, involutional, 166710; Kaposi sarcoma, 148000	7p21
179	IMMP2L	Inner mitochondrial membrane peptidase 2 like	605977	Gilles de la Tourette syndrome [candidate OMIM:605977]	7q31

Gene #	Symbol	Full name	OMIM#	Disorder	Location
180	IMPDH1	IMP (inosine monophosphate) dehydrogenase 1	146690		7q31.3-q32
181	INHBA	Inhibin, beta A (activin A, activin AB alpha polypeptide)	147290		7p15-p13
182	INMT	Indolethylamine N-methyltransferase	604854		7p15.3-p15.2
183	INSIG1	Insulin induced gene 1	602055		7q36
184	JAZF1	Juxtaposed with another zinc finger gene 1	606246	Endometrial stromal tumors	7p15
185	JTV1	JTV1 gene	600859		7p22
186	KCND2	Potassium voltage-gated channel, Shal-related subfamily, member 2	605410		7q31-q32
187	KCNH2	Potassium voltage-gated channel, subfamily H (eag-related), member 2	152427	Long QT syndrome-2	7q35-q36
188	KEL	Kell blood group	110900		7q33
189	LAMB1	Laminin, beta 1	150240	Cutis laxa, marfanoid neonatal type	7q22
190	LEP	Leptin (obesity homolog, mouse)	164160	Obesity, morbid, with hypogonadism; Obesity, severe, due to leptin deficiency	7q31.3
191	LFNG	Lunatic fringe homolog ( <i>Drosophila</i> )	602576		7p22
192	LGMD1D	Limb girdle muscular dystrophy 1D (autosomal dominant)	603511	Muscular dystrophy, limb-girdle, type 1D	7q
193	LIMK1	LIM domain kinase 1	601329		7q11.23
194	MAD1L1	MAD1 mitotic arrest deficient-like 1 (yeast)	602686	Prostate cancer, 176807; Lymphoma, somatic	7p22
195	MAFK	V-maf musculoaponeurotic fibrosarcoma oncogene homolog K (avian)	600197		7p22
196	MAP2K2	Mitogen-activated protein kinase kinase 2	601263		7q32
197	MCM7	MCM7 minichromosome maintenance deficient 7 ( <i>S. cerevisiae</i> )	600592		7q21.3-q22.1
198	MDDC		153880	Macular dystrophy, dominant cystoid	7p21-p15
199	MDH2	Malate dehydrogenase 2, NAD (mitochondrial)	154100		7p12.3-q11.2
200	MEOX2	Mesenchyme homeo box 2 (growth arrest-specific homeo box)	600535		7p22.1-p21.3
201	MEST	Mesoderm specific transcript homolog (mouse)	601029		7q32
202	MET	Met proto-oncogene (hepatocyte growth factor receptor)	164860	Renal cell carcinoma, papillary, 605074	7q31
203	MGAM	Maltase-glucoamylase (alpha-glucosidase)	154360		7q32.3
204	MHS3	Malignant hyperthermia susceptibility 3	154276	{Malignant hyperthermia susceptibility 3}	7q21-q22
205	MKLN1	Muskelin 1, intracellular mediator-containing kelch motifs	605623		7q32
206	MTERF	Transcription termination factor, mitochondrial	602318		7q21-q22
207	MTPN	Myotrophin	606484		7q33-q35
208	MUC11	Mucin 11	604608		7q22
209	MUC12	Mucin 12	604609		7q22
210	MUC3A	Mucin 3A, intestinal	158371	Ulcerative colitis, susceptibility to, 191390	7q22
211	MUC3B	Mucin 3B	605633		7q22
212	MYCLK1	V-myc myelocytomatosis viral oncogene homolog (avian)-like 1	164865		7p15
213	NCF1	Neutrophil cytosolic factor 1 (47 kD, chronic granulomatous disease, autosomal 1)	233700	Chronic granulomatous disease due to deficiency of NCF-1	7q11.23
214	NDUFA5	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 5 (13 kD, B13)	601677		7q32
215	NFE2L3	Nuclear factor (erythroid-derived 2)-like 3	604135		7p15-p14
216	NHCP2	Non-histone chromosome protein 2	118880		7

Gene #	Symbol	Full name	OMIM#	Disorder	Location
217	NM	Neutrophil migration	162820		7q22-qter
218	NOS3	Nitric oxide synthase 3 (endothelial cell)	163729	Hypertension, pregnancy-induced (PEE 1) [OMIM:189800]	7q36 7q21.3-q22.1
219	NPTX2	Neuronal pentraxin II	600750		7p15.1 7q31.1-q31.2
220	NPY	Neuropeptide Y	162640		7q32
221	NRCAM	Neuronal cell adhesion molecule	601581		7p15.3
222	NRF1	Nuclear respiratory factor 1	600879		7p22
223	NT5C3	5'-nucleotidase, cytosolic III	606224	Uridine 5-prime monophosphate hydrolase deficiency, hemolytic anemia due to	7p13-p11
224	NUDT1	Nudix (nucleoside diphosphate linked moiety X)-type motif 1	600312		7p14-p13
225	OCM	Oncomodulin	164795		7q31.3-q32
226	OGDH	Oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)	203740	Alpha-ketoglutarate dehydrogenase deficiency	7q22.1
227	OPN1SW	Opsin 1 (cone pigments), short-wave-sensitive (color blindness, tritan)	190900	Color blindness, tritan	7q34-q36
228	ORCSL	Origin recognition complex, subunit 5-like (yeast)	602331		7q31.3
229	OTSC2	Otosclerosis 2	605727	Otosclerosis-2	7q32
230	p100	EBNA-2 co-activator (100 kD)	602181		7q11.23-q21.3
231	PAX4	Paired box gene 4	167413		7q22.1
232	PCLO	Piccolo (presynaptic cytomatrix protein)	604918		7q21.3
233	PCOLCE	Procollagen C-endopeptidase enhancer	600270		7q22
234	PDAP1	PDGFA associated protein 1	607075		7q11.21
235	PDGFA	Platelet-derived growth factor alpha polypeptide	173430		7p22 7q21.3-q22.1
236	PDK4	Pyruvate dehydrogenase kinase, isoenzyme 4	602527		7q21-q22
237	PEX1	Peroxisome biogenesis factor 1	602136	Refsum disease, infantile form, Zellweger syndrome, Adrenoleukodystrophy, autosomal neonatal form,	
238	PGAM2	Phosphoglycerate mutase 2 (muscle)	261670	Myopathy due to phosphoglycerate mutase deficiency	7p13-p12
239	PHKG1	Phosphorylase kinase, gamma 1 (muscle)	172470		7p12-q21
240	PIK3CG	Phosphoinositide-3-kinase, catalytic, gamma polypeptide	601232		7q21.11
241	PILR	Paired immunoglobulin-like receptor alpha	605341		7q22
242	PILR	Paired immunoglobulin-like receptor beta	605342		7q22
243	PIP	Prolactin-induced protein	176720		7q34
244	PLOD3	Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3	603066		7q22
245	PLXNA4	Plexin A4	604280		7
246	PMPCB	Peptidase (mitochondrial processing) beta	603131		7q22-q32
247	PMS2	PMS2 postmeiotic segregation increased 2 ( <i>S. cerevisiae</i> )	600259	Turcot syndrome, 276300; Colorectal cancer, hereditary nonpolyposis, 114500	7p22
248	PODXL	Podocalyxin-like	602632		7q32-q33
249	POLD2	Polymerase (DNA directed), delta 2, regulatory subunit (50 kD)	600815		7p15.1
250	POLR2J	Polymerase (RNA) II (DNA directed) polypeptide J (13.3 kD)	604150		7q11.2
251	POMZP3	POM (POM121 rat homolog) and ZP3 fusion	600587		7q11.23
252	PON1	Paraoxonase 1	168820	Coronary artery disease, susceptibility to; Coronary arteryspasm, susceptibility to	7q21.3
253	PON2	Paraoxonase 2	602447	{Coronary artery disease, susceptibility to}	7q21.3

Gene #	Symbol	Full name	OMIM#	Disorder	Location
254	PON3	Paraoxonase 3	602720		7q21.3
255	POR	P450 (cytochrome) oxidoreductase	124015		7q11.2
256	PPD2	Polydactyly, preaxial II	174500	Polydactyly, preaxial II	7q36
257	PPIA	Peptidylprolyl isomerase A (cyclophilin A)	123840		7p13-p11.2
258	PPP1R3A	Protein phosphatase 1, regulatory (inhibitor) subunit 3A	600917	Diabetes mellitus, insulin-resistant, with acanthosis nigricans and hypertension,	7q11.23
259	PRES	Prestin (motor protein)	604943		7q22
260	PRKAG2	Protein kinase, AMP-activated, gamma 2 non-catalytic subunit	602743	Cardiomyopathy, familial hypertrophic, with Wolff-Parkinson-White syndrome, Wolff-Parkinson-White syndrome,	7q35-q36
261	PRKAR1B	Protein kinase, cAMP-dependent, regulatory, type I, beta	176911		7pter-p22
262	PRKAR2B	Protein kinase, cAMP-dependent, regulatory, type II, beta	176912		7q22-q31.1
263	PRSS1	Protease, serine, 1 (trypsin 1)	276000	Pancreatitis, hereditary, 167800; Trypsinogen deficiency	7q34
264	PRSS2	Protease, serine, 2 (trypsin 2)	601564		7q34
265	PSCD3	Pleckstrin homology, Sec7 and coiled/coil domains 3	605081		7p22.3-p22.2
266	PSMC2	Proteasome (prosome, macropain) 26S subunit, ATPase, 2	154365		7q22.1-q22.3
267	PSPH	Phosphoserine phosphatase	172480		7p15.2-p15.1
268	PSPHL	Phosphoserine phosphatase-like	604239		7q11.2
269	PTN	Pleiotrophin (heparin-binding growth factor 8, neurite growth-promoting factor 1)	162095		7q33-q34
270	PTPN12	Protein tyrosine phosphatase, non-receptor type 12	600079	Colon cancer	7q11.23
271	PTPRN2	Protein tyrosine phosphatase, receptor type, N polypeptide 2	601698		7q36
272	PTPRZ1	Protein tyrosine phosphatase, receptor-type, Z polypeptide 1	176891		7q31.3
273	RALA	V-ral simian leukemia viral oncogene homolog A (ras related)	179550		7p22-p15
274	RAMP3	Receptor (calcitonin) activity-modifying protein 3	605155		7p13-p12
275	RELN	Reelin	600514	Lissencephaly syndrome, norman-roberts type, 257320	7q22
276	RFC2	Replication factor C (activator 1) 2 (40 kD)	600404		7q11.23
277	RHEB2	Ras homolog enriched in brain 2	601293		7q36
278	RNY1	RNA, Y1 small cytoplasmic (associated with Ro protein)	601821		7q36
279	RNY3	RNA, Y3 small cytoplasmic (associated with Ro protein)	601822		7q36
280	RNY4	RNA, Y4 small cytoplasmic (associated with Ro protein)	601823		7q36
281	RNY5	RNA, Y5 small cytoplasmic (associated with Ro protein)	601824		7q36
282	RP10	Retinitis pigmentosa 10 (autosomal dominant)	180105	Retinitis pigmentosa-10	7q31-q35
283	RP9	Retinitis pigmentosa 9 (autosomal dominant)	180104	Retinitis pigmentosa-9	7p14.3
284	RPA3	Replication protein A3 (14 kD)	179837		7p22
285	RPP20	POP7 (processing of precursor, <i>S. cerevisiae</i> ) homolog	606113		7q22
286	SCAP2	Src family associated phosphoprotein 2	605215		7p21-p15
287	SCYA24	Small inducible cytokine subfamily A (Cys-Cys), member 24	602495		7q11.23
288	SDS	Shwachman-Diamond syndrome	260400	Shwachman-Diamond syndrome	7p11-q11

Gene #	Symbol	Full name	OMIM#	Disorder	Location
289	SERPINE1	Serine (or cysteine) proteinase inhibitor, clade E	173360	Thrombophilia due to excessive plasminogen activator inhibitor ; Hemorrhagic diathesis due to PAII deficiency	7p14.1 7q21-q22
290	SFRP4	Secreted frizzled-related protein 4	606570		7q21.3-q22.1 7q21.2-q21.3
291	SGCE	Sarcoglycan, epsilon	604149	Myoclonic dystonia, 159900	7q36
292	SHFM1	Split hand/foot malformation (ectrodactyly) type 1	183600	Split hand/foot malformation, type 1	7q36
293	SHFM1D	Split hand/foot malformation type 1 with deafness	605617	Split hand/foot malformation type 1 with deafness	
294	SHH	Sonic hedgehog homolog ( <i>Drosophila</i> )	600725	Solitary median maxillary central incisor, Holoprosencephaly 3, Basal cell carcinoma	7q31-q32 7q33
295	SLC13A1	Solute carrier family 13 (sodium/sulfate symporters), member 1	606193		7q21.3
296	SLC13A4	Solute carrier family 13 (sodium/sulfate symporters), member 4	604309		7q31
297	SLC25A13	Solute carrier family 25, member 13 (citrin)	603859	Citrullinemia, type II, adult-onset, 603471	7q31
298	SLC26A3	Solute carrier family 26, member 3	126650	Chloride diarrhea, familial, 214700; Colon cancer	
299	SLC26A4	Solute carrier family 26, member 4	605646	Deafness, neurosensory, autosomal recessive 4, Pendred syndrome, Enlarged vestibular aqueduct	7 7q35-q36
300	SLC30A3	Solute carrier family 30 (zinc transporter), member 3	602878		7p
301	SLC4A2	Solute carrier family 4, anion exchanger, member 2	109280		7q35-q36
302	SMAD1		600794	Spinal muscular atrophy, distal, with upper limb predominance	
303	SMARCD3	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 3	601737		7q32.3 7q21.1-q31.1
304	SMOH	Smoothened homolog ( <i>Drosophila</i> )	601500	Basal cell carcinoma, sporadic	7p22
305	SMURF1	E3 ubiquitin ligase SMURF1	605568		7p21.3
306	SNL	Singed-like (fascin homolog, sea urchin) ( <i>Drosophila</i> )	602689		7p15
307	SNX13	Sorting nexin 13	606589		7q31
308	SP4	Sp4 transcription factor	600540		
309	SPAM1	Sperm adhesion molecule 1 (PH-20 hyaluronidase, zona pellucida binding)	600930		
310	SRI	Sorcin	182520		7q21.1
311	SSBP1	Single-stranded DNA binding protein	600439	Colon cancer [OMIM:600833]	7q34 7q31.1-q31.3
312	ST7	Suppression of tumorigenicity 7	600833		7q21
313	STEAP	Six transmembrane epithelial antigen of the prostate	604415		7p21.3-p21.2
314	STK31	Serine/threonine kinase 31	605790		7q
315	STQTL7	Stature QTL on chromosome 7	606256	[Stature QTL]	
316	STX1A	Syntaxin 1A (brain)	186590		7q11.23
317	TAC1	Tachykinin, precursor 1	162320		7q21-q22
318	TAF6	TAF6 RNA polymerase II, TATA box-binding protein (TBP)-associated factor, 80 kD	602955		7q11.1
319	TAS2R16	Taste receptor, type 2, member 16	604867		7q31.1-q31.3
320	TAS2R3	Taste receptor, type 2, member 3	604868		7q31.3-q32

Gene #	Symbol	Full name	OMIM#	Disorder	Location
321	TAS2R4	Taste receptor, type 2, member 4	604869		7q31.3-q32
322	TAS2R5	Taste receptor, type 2, member 5	605062		7q31.3-q32
323	TAX1BP1	Tax1 (human T-cell leukemia virus type I)-binding protein 1	605326		7p15
324	TBL2	Transducin (beta)-like 2	605842		7q11.23
325	TBX20	T-box 20	606061		7p15-p14
326	TBXAS1	Thromboxane A synthase 1 (platelet, cytochrome P450, subfamily V)	274180	Thromboxane synthase deficiency	7q34-q35
327	TCF6L1	Transcription factor 6-like 1 (mitochondrial transcription factor 1-like)	157670		7pter-cen
328	TEM6	Tumor endothelial marker 6	606825		7p15.1
329	TES	Testis derived transcript (3 LIM domains)	606085		7q31.2
330	TFPI2	Tissue factor pathway inhibitor 2	600033		7q22
331	TFR2	Transferrin receptor 2	604720	Hemochromatosis, type 3, 604250	7q22
332	TIF1	Transcriptional intermediary factor 1	603406	Thyroid carcinoma, papillary, 188550	7q32-q34
333	TPK1	Thiamin pyrophosphokinase 1	606370		7q34-q35
334	TPST1	Tyrosylprotein sulfotransferase 1	603125		7q11.21
335	TPPTS	Triphalangeal thumb-polysyndactyly syndrome	190605	Triphalangeal thumb-polysyndactyly syndrome	7q36
336	TRB@	T cell receptor beta locus	186930		7q34
337	TRG@	T cell receptor gamma locus	186970		7p15-p14
338	TRIP6	Thyroid hormone receptor interactor 6	602933		7q22 7q21.2-q22.1
339	TRRAP	Transformation/transcription domain-associated protein	603015		7p12-cen
340	TTIM1	T-cell tumor invasion and metastasis 1	147830		7p21.2
341	TWIST	Twist homolog (acrocephalosyndactyly 3; Saethre-Chotzen syndrome)	601622	Saethre-Chotzen syndrome, 101400	7q32
342	UBE2H	Ubiquitin-conjugating enzyme E2H (UBC8 homolog, yeast)	601082		7
343	UP	Uridine phosphorylase	191730		7q22
344	VGf	VGF nerve growth factor inducible	602186		7q36.3
345	VIPR2	Vasoactive intestinal peptide receptor 2	601970		7p14-p13
346	VPS41	Vacuolar protein sorting 41 (yeast)	605485		7q11.23
347	WBSCR1	Williams-Beuren syndrome chromosome region 1	603431		7q11.23
348	WBSCR14	Williams Beuren syndrome chromosome region 14	605678		7q11.23
349	WBSCR5	Williams-Beuren syndrome chromosome region 5	605719		7q11.23
350	WNT2	Wingless-type MMTV integration site family member 2	147870		7p15-p11.2
351	WTSL	Wilms tumor suppressor locus	601583	{Wilms tumor susceptibility-5}	7q36.1
352	XRCC2	X-ray repair complementing defective repair in Chinese hamster cells 2	600375		7q11.23
353	YWHAG	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma polypeptide	605356		7q22
354	ZAN	Zonadhesin	602372		7q11.2
355	ZNF107	Zinc finger protein 107 (Y8)	603989		7q11.2
356	ZNF117	Zinc finger protein 117 (HPF9)	194624		7q11.2

Gene #	Symbol	Full name	OMIM#	Disorder	Location
357	ZNF12	Zinc finger protein 12 (KOX 3)	194536		7p22-p21
358	ZNF138	Zinc finger protein 138 (clone pHZ-32)	604080		7q11.21- q11.23
359	ZNF212	Zinc finger protein 212	602386		7q36.1
360	ZNF277	Zinc finger protein 277	605465		7q31.1
361	ZNF36	Zinc finger protein 36 (KOX 18)	601260		7q21.3-q22.1
362	ZNF38	Zinc finger protein 38 (KOX 25)	601261		7q21-q22
363	ZNFN1A1	Zinc finger protein, subfamily 1A, 1 (Ikaros)	603023	Leukemia, acute lymphoblastic	7p13-p11.1
364	ZP3A	Zona pellucida glycoprotein 3A (sperm receptor)	182889		7q11.23
365	ZP3B	Zona pellucida glycoprotein 3B (sperm receptor)	195002		7
366	ZRF1	Zuotin related factor 1	605502		7q22-q32
367	ZYX	Zyxin	602002		7q32

**Table S2.** Summary of components that compose the CRA\_TCAGchr7.v1 chromosome 7 DNA sequence assembly. The 26 Celera scaffolds ranged from 195 kb to 24.9 Mb in size with an average of 4.9 Mb. The size of the remaining intrascaffold gaps could be estimated based on mate-pair information. The estimated size of the intrascaffold gaps are listed in brackets, because they are already included in the total nucleotide content within each scaffold. An estimated 202,000 N's were substituted into the remaining six physical gaps and 2,700,000 N's were substituted at the centromere.

Component	Number	Length of Sequence (bp)	Percent
Celera scaffolds	26	133,269,991	84.4
(Intrascaffold gaps)	(138)	(218,684)	(0.14)
Genomic clones	209	21,781,798	13.8
Centromere	1	2,700,000	1.7
Physical Gaps	6	202,000	0.1
Total	242	157,953,789	100

**Table S3.** Components of the chromosome 7 DNA sequence assembly. The complete DNA sequence assembly can be downloaded at <http://www.chr7.org> and is also at GenBank [see text reference (7)]. Of the 209 components that are not from the updated Celera human genome assembly, 196 are clone-based sequences from the Genome Sequencing Center at Washington University, 9 are from the University of Washington Genome Center, 2 are from the German Sequencing Project (Jena), 1 is from the Lawrence Livermore National Laboratory Human Genome Center, and 1 is from the University of Oklahoma Genome Center.

Component #	Chr. 7 Start Coordinate	Chr. 7 End Coordinate	Component Name	Component Start	Component End	Orientation	Length	# Gaps	Total Size of Gaps
1	1	1120	AC099546.2	1	1120	+	1120		
2	1121	51120	50kN	1	50000	+	50000	1	50000
3	51121	96590	AC097647.2	1	45470	+	45470		
4	96591	105103	AC112517.1	33129	41641	+	8513		
5	105104	114912	AC097645.2	31764	41572	+	9809		
6	114913	127224	AC097646.2	30650	42961	+	12312		
7	127225	157442	AC093686.6	3821	34038	+	30218		
8	157443	170367	AC093666.4	1	12925	+	12925		
9	170368	177612	AC093627.4	1	7245	+	7245		
10	177613	656460	GA_x39G1GC42LN_4	1	478848	+	478848	25	74891
11	656461	706460	50kN	1	50000	+	50000	1	50000
12	706461	1189758	GA_x39G1FKSFPU_4	1	483298	+	483298	5	7515
13	1189759	4656132	GA_x5HB7VCL7H0_3	3805	3470178	+	3466374	4	716
14	4656133	4808169	AC072054.10	1	152037	+	152037		
15	4808170	4847992	AC092428.4	1	39823	+	39823		
16	4847993	5008010	AC092610.1	1	160018	+	160018		
17	5008011	5030248	AC053546.8	1	22238	+	22238		
18	5030249	5206646	AC092032.5	1	176398	+	176398		
19	5206647	5282813	AC092028.4	2001	78167	+	76167		
20	5282814	5456975	AC093376.5	2001	176162	+	174162		
21	5456976	5526158	AC093620.3	2001	71183	+	69183		
22	5526159	5609290	AC092171.4	2001	85132	+	83132		
23	5609291	5803697	AC006483.3	1	194407	-	194407		
24	5803698	5944060	AC008167.5	1	140363	-	140363		
25	5944061	6066881	AC004983.2	2001	124821	+	122821		
26	6066882	6146691	AC005995.3	1	79810	-	79810		
27	6146692	6806357	GA_x5HB7VCL73F_3	56781	716446	+	659666	4	2239
28	6806358	6892575	AC073343.6	87750	173967	+	86218		
29	6892576	6950002	AC079882.6	2001	59427	+	57427		
30	6950003	7102375	AC079804.5	2001	154373	+	152373		
31	7102376	29799627	GA_x5HB7VCL79P_4	1	22697252	-	22697252	11	15124
32	29799628	29859417	AC007276.3	39110	98899	+	59790		
33	29859418	32673246	GA_x5HB7VCL779_3	133858	2947686	-	2813829	0	0
34	32673247	32842480	AC018633.2	1	169234	+	169234		
35	32842481	32927451	AC018645.4	24644	109614	+	84971		
36	32927452	35043431	GA_x5HB7VCL78J_3	1	2115980	+	2115980	1	3254
37	35043432	35073644	AC004711.1	11369	41581	+	30213		

38	35073645	35184293	AC010085.3	5624	116272	-	110649
----	----------	----------	------------	------	--------	---	--------

Component #	Chr. 7 Start Coordinate	Chr. 7 End Coordinate	Component Name	Component Start	Component End	Orientation	Length	# Gaps	Total Size of Gaps
39	35184294	35359268	AC006379.2	201	175175	+	174975		
40	35359269	44080811	GA_x5HB7VCL77B_3	1	8721543	+	8721543	5	20715
41	44080812	44083066	AC004985.2	1	2255	-	2255		
42	44083067	44134859	AC004951.5	1	51793	-	51793		
43	44134860	44159323	AC017116.7	2001	26464	+	24464		
44	44159324	45843882	GA_x5HB7VCL761_3	1	1684559	-	1684559	2	48
45	45843883	45870425	AC091439.5	77233	103775	+	26543		
46	45870426	45975559	AC096582.1	2001	107134	+	105134		
47	45975560	48208174	GA_x5HB7VCL7FH_6	1	2232615	+	2232615	2	120
48	48208175	51673296	GA_x54KREB6MY7_6	1	3465122	-	3465122	8	12468
49	51673297	53217371	GA_x54KREB72W2_6	12762	1556836	+	1544075	1	2200
50	53217372	53379831	AC073318.8	1	162460	+	162460		
51	53379832	55765198	GA_x54KREB72Y9_3	67820	2453186	+	2385367	2	40
52	55765199	55819489	AC073237.3	1	54291	+	54291		
53	55819490	55915276	AC099681.4	2001	97787	+	95787		
54	55915277	56036240	AC091812.4	1980	122943	+	120964		
55	56036241	56198131	AC092647.2	2001	163891	+	161891		
56	56198132	56315089	AC092579.3	2001	118958	+	116958		
57	56315090	56351964	AC092101.5	2001	38875	+	36875		
58	56351965	56486071	AC006970.6	1992	136098	+	134107		
59	56486072	56566667	AC073136.6	201	80796	+	80596		
60	56566668	56664264	AC093392.4	2001	99597	+	97597		
61	56664265	56687199	AC092423.5	2001	24935	+	22935		
62	56687200	56861746	AC092447.5	2001	176547	+	174547		
63	56861747	57152262	GA_x5HB7VC7SFH_3	95227	385742	+	290516	0	0
64	57152263	57175518	AC118758.3	127558	150813	+	23256		
65	57175519	57308614	AC069152.5	1	133096	-	133096		
66	57308615	57384445	AC122133.4	115273	191103	+	75831		
67	57384446	57386833	AC125231.2	2001	4388	+	2388		
68	57386834	57534090	AC099654.4	1	147257	+	147257		
69	57534091	57712195	AC073057.6	1	178105	+	178105		
70	57712196	57810840	AC092175.5	2001	100645	+	98645		
71	57810841	57953804	AC064862.6	1	142964	-	142964		
72	57953805	57955798	AC103878.4	2001	3994	+	1994		
73	57955799	58151435	AC023141.6	2001	197637	+	195637		
74	58151436	58344712	AC017075.8	1	193277	+	193277		
75	58344713	61044712	2700kN	1	2700000	+	2700000	1	2700000
76	61044713	61229994	AC019063.4	2001	187282	-	185282		

Component #	Chr. 7 Start Coordinate	Chr. 7 End Coordinate	Component Name	Component Start	Component End	Orientation	Length	# Gaps	Total Size of Gaps
77	61229995	61266885	AC110784.3	1	36891	+	36891		
78	61266886	61290392	AC118943.2	1	23507	+	23507		
79	61290393	61430440	AC104789.4	1	140048	+	140048		
80	61430441	61567886	AC092585.2	1	137446	+	137446		
81	61567887	61741840	AC069285.8	1999	175952	+	173954		
82	61741841	61766196	AC092001.4	2001	26356	+	24356		
83	61766197	61903580	AC006455.2	2001	139384	+	137384		
84	61903581	61966708	AC073188.10	2001	65128	+	63128		
85	61966709	62098896	AC006457.4	2000	134187	+	132188		
86	62098897	62211113	AC006015.5	201	112417	+	112217		
87	62211114	62349354	AC079355.6	2001	140241	+	138241		
88	62349355	62541499	AC092634.3	1998	194142	+	192145		
89	62541500	62655415	AC115220.1	2001	115916	+	113916		
90	62655416	62721714	AC091685.4	2001	68299	+	66299		
91	62721715	62758151	AC104094.4	2001	38437	+	36437		
92	62758152	62944830	AC073270.6	2001	188679	+	186679		
93	62944831	63039727	AC022202.12	201	95097	+	94897		
94	63039728	63243099	AC016769.10	201	203572	+	203372		
95	63243100	63289779	AC091799.5	2001	48680	+	46680		
96	63289780	63428640	AC073349.11	2001	140861	+	138861		
97	63428641	63474451	AC092161.3	2001	47811	+	45811		
98	63474452	63675847	AC073210.8	2001	203396	+	201396		
99	63675848	63852517	AC104073.3	2001	178670	+	176670		
100	63852518	63949798	AC092685.2	85983	183263	+	97281		
101	63949799	64082137	AC114501.1	2001	134339	+	132339		
102	64082138	64083572	AC104092.4	2000	3434	+	1435		
103	64083573	64163915	AC104057.4	1	80343	+	80343		
104	64163916	64356092	AC073107.7	1	192177	+	192177		
105	64356093	64447316	AC093582.3	1	91224	+	91224		
106	64447317	64464514	AC093485.4	2001	19198	+	17198		
107	64464515	64555917	AC073261.8	2001	93403	+	91403		
108	64555918	65220597	GA_x5HB7VC11UB_3	434098	1098777	-	664680	1	654
109	65220598	65408804	AC027644.9	1	188207	-	188207		
110	65408805	65475079	AC073335.5	2001	68275	+	66275		
111	65475080	65603363	AC079920.6	2001	130284	+	128284		
112	65603364	65773151	AC073089.5	2001	171788	+	169788		
113	65773152	65930903	AC006480.3	2001	159752	+	157752		

114	65930904	71002633	GA_x5L2HTUEKN6_3	98299	5170028	+	5071730	7	401
-----	----------	----------	------------------	-------	---------	---	---------	---	-----

Component #	Chr. 7 Start Coordinate	Chr. 7 End Coordinate	Component Name	Component Start	Component End	Orientation	Length	# Gaps	Total Size of Gaps
115	71002634	71135388	AC092536.3	1	132755	+	132755		
116	71135389	71265316	AC091738.4	2001	131928	+	129928		
117	71265317	71297325	AC105447.4	2001	34009	+	32009		
118	71297326	71309187	AC114816.4	2001	13862	+	11862		
119	71309188	71314261	AC114731.3	1	5074	+	5074		
120	71314262	71399854	AC005236.4	2001	87593	+	85593		
121	71399855	71585391	AC005488.2	201	185737	+	185537		
122	71585392	71756092	AC006995.5	1	170701	-	170701		
123	71756093	71811480	AC073841.9	1	55388	-	55388		
124	71811481	71918208	AC005049.2	1	106728	-	106728		
125	71918209	71968612	AC005074.1	1	50404	-	50404		
126	71968613	72127431	AC005089.3	1	158819	-	158819		
127	72127432	72248435	AC073846.6	1	121004	-	121004		
128	72248436	72395124	AC093168.3	2001	148689	+	146689		
129	72395125	72479435	AC099398.5	2001	86311	+	84311		
130	72479436	72574982	AC005056.2	652	96198	-	95547		
131	72574983	72674352	AC005057.2	1	99370	-	99370		
132	72674353	72766893	AC005081.3	87550	180090	-	92541		
133	72766894	72986828	AC005015.2	1	219935	-	219935		
134	72986829	73058255	AC004851.2	1	71427	-	71427		
135	73058256	73157036	AC005231.3	3132	101912	-	98781		
136	73157037	73326640	AC083884.6	1	169604	+	169604		
137	73326641	73448244	AC004867.5	1	121604	-	121604		
138	73448245	73494600	AC004166.12	228842	275197	+	46356		
139	73494601	73495600	lkn	1	1000	+	1000	1	1000
140	73495601	73726152	AC005098.2	1	230552		230552		
141	73726153	73727152	lkn	1	1000	+	1000	1	1000
142	73727153	73814185	AC118138.2	1	87033	+	87033		
143	73814186	73908215	AC004878.3	16655	110684	-	94030		
144	73908216	74035974	AC006014.2	1	127759	+	127759		
145	74035975	74170753	AC018720.5	1	134779	-	134779		
146	74170754	74346392	AC004491.1	201	175839	+	175639		
147	74346393	74351609	AC006025.2	1	5217	-	5217		
148	74351610	74457917	AC005102.1	1	106308	-	106308		
149	74457918	74547165	AC005067.2	1	89248	-	89248		
150	74547166	74643264	AC006330.5	1	96099	-	96099		

151	74643265	74883443	AC005077.5	1	240179	-	240179
152	74883444	74942484	AC006388.3	1	59041	-	59041

Component #	Chr. 7 Start Coordinate	Chr. 7 End Coordinate	Component Name	Component Start	Component End	Orientation	Length	# Gaps	Total Size of Gaps
153	74942485	75091889	AC005522.2	1	149405	-	149405		
154	75091890	75106164	AC007078.4	1	14275	-	14275		
155	75106165	75313458	AC004980.5	1499	208792	+	207294		
156	75313459	75484995	AC006972.2	1	171537	+	171537		
157	75484996	75518882	AC007003.5	201	34087	+	33887		
158	75518883	75614109	AC114737.3	2001	97227	+	95227		
159	75614110	75810769	AC007000.2	1	196660	+	196660		
160	75810770	75869115	AC098851.5	103125	161470	+	58346		
161	75869116	75944742	AC073635.8	55196	130822	+	75627		
162	75944743	76094874	AC004921.1	1	150132	-	150132		
163	76094875	76240640	AC006451.5	201	145966	+	145766		
164	76240641	76283835	AC090421.5	2001	45195	+	43195		
165	76283836	76402392	AC004955.3	2001	120557	+	118557		
166	76402393	76487411	AC073520.6	201	85219	+	85019		
167	76487412	96426521	GA_x5L2HTUMHYR_3	28900	19968009	-	19939110	7	15767
168	96426522	96524077	AC079781.7	1	97556	+	97556		
169	96524078	96537266	AC004967.3	2001	15189	+	13189		
170	96537267	99246376	GA_x5L2HTUMHW3_6	294676	3003785	-	2709110	5	4354
171	99246377	99345052	AC009488.5	1	98676	+	98676		
172	99345053	99517409	AC011895.4	1	172357	+	172357		
173	99517410	99547605	AC118759.3	1	30196	+	30196		
174	99547606	99597605	50kN	1	50000	+	50000	1	50000
175	99597606	99737122	AC105446.4	1	139517	+	139517		
176	99737123	99740808	AC112241.3	2001	5686	+	3686		
177	99740809	99865268	AC004876.2	2003	126462	+	124460		
178	99865269	99865767	AC007008.3	201	699	+	499		
179	99865768	100010820	AC006329.5	201	145253	+	145053		
180	100010821	100181378	AC004965.2	201	170758	+	170558		
181	100181379	100327585	AC004953.2	201	146407	+	146207		
182	100327586	100512761	AC005096.4	201	185376	+	185176		
183	100512762	100553305	AC092788.3	2001	42544	+	40544		
184	100553306	100620672	AC005072.2	2001	69367	+	67367		
185	100620673	100769285	AF047825.1	1	148613	-	148613	2	2
186	100769286	100839186	AC005086.2	59686	129586	+	69901		
187	100839187	100902969	AC005088.2	201	63983	+	63783		

188	100902970	101019978	AC091390.1	3158	120166	-	117009
189	101019979	101178332	AC093668.4	1	158354	+	158354
190	101178333	101220255	AC004084.1	51351	93273	-	41923

Component #	Chr. 7 Start Coordinate	Chr. 7 End Coordinate	Component Name	Component Start	Component End	Orientation	Length	# Gaps	Total Size of Gaps
191	101220256	101344337	AC105052.3	1	124082	+	124082		
192	101344338	101465857	AC006477.4	1	121520	+	121520		
193	101465858	101559993	AC005250.1	201	94336	+	94136		
194	101559994	101713835	AC073127.11	2001	155842	+	153842		
195	101713836	101722555	AC108167.3	2001	10720	+	8720		
196	101722556	101876290	AC007683.5	1	153735	-	153735		
197	101876291	101988751	AC004668.1	1	112461	-	112461		
198	101988752	102037917	AC093701.4	2001	51166	+	49166		
199	102037918	126983441	GA_x54KREB2RL0_6	133042	25078565	-	24945524	15	18046
200	126983442	127148531	AC010655.7	1	165090	+	165090		
201	127148532	127238253	AC090114.5	2001	91722	+	89722		
202	127238254	127379569	AC093183.3	32818	174133	+	141316		
203	127379570	140750156	GA_x54KREB3J6B_4	4089	13374675	-	13370587	16	26526
204	140750157	140919524	AC091742.5	1	169368	+	169368		
205	140919525	141186680	U66059.1	1	267156	+	267156		
206	141186681	141392464	U66060.1	9639	215422	+	205784		
207	141392465	141604497	U66061.1	20618	232650	+	212033		
208	141604498	142871475	GA_x54KREB44CJ_5	806657	2073634	+	1266978	3	592
209	142871476	142970821	AC004889.1	30685	130030	+	99346		
210	142970822	143027366	AC074386.6	10066	66610	-	56545		
211	143027367	148558089	GA_x54KREB449V_4	1	5530723	+	5530723	2	437
212	148558090	148699621	AC092681.3	7287	148818	+	141532		
213	148699622	148834555	AC092666.2	2001	136934	+	134934		
214	148834556	148859477	AC099647.5	2001	26922	+	24922		
215	148859478	148915031	AC006008.2	1	55554	-	55554		
216	148915032	148957331	AC005586.2	89651	131950	-	42300		
217	148957332	150326678	GA_x54KREB44AV_3	1	1369347	+	1369347	1	790
218	150326679	150421442	AC006358.6	1	94764	-	94764		
219	150421443	150597913	AC093583.3	2001	178471	+	176471		
220	150597914	150651619	AC074257.5	132342	186047	+	53706		
221	150651620	150665772	AC099345.3	2001	16153	+	14153		
222	150665773	150826328	AC006017.2	2001	162556	+	160556		
223	150826329	150895846	AC006474.3	201	69718	+	69518		
224	150895847	151042452	AC104692.1	2001	148606	+	146606		

225	151042453	151184225	AC005631.4	1	141773	-	141773
226	151184226	151308851	AC104843.3	2001	126626	+	124626
227	151308852	151334724	AC003109.1	1	25873	-	25873
228	151334725	151418005	AC092180.5	1995	85275	+	83281

Component #	Chr. 7 Start Coordinate	Chr. 7 End Coordinate	Component Name	Component Start	Component End	Orientation	Length	# Gaps	Total Size of Gaps
229	151418006	151556155	AC072057.8	1	138150	-	138150		
230	151556156	151691838	AC006348.3	1	135683	-	135683		
231	151691839	152120511	GA_x5HB7VC9F37_3	275971	704643	+	428673	1	20
232	152120512	152139608	AC091744.5	3936	23032	+	19097		
233	152139609	152356339	AC092033.4	2001	218731	+	216731		
234	152356340	152452722	AC005998.3	1	96383	-	96383		
235	152452723	152523002	AC006973.2	1	70280	-	70280		
236	152523003	152642285	AC005588.1	1	119283	-	119283		
237	152642286	152766700	AC006019.2	1	124415	-	124415		
238	152766701	152919365	AC104594.2	2001	154665	+	152665		
239	152919366	152969304	AC099341.4	2001	51939	+	49939		
240	152969305	153955934	GA_x5J8B7Q6JNY_6	155406	1142035	+	986630	3	3689
241	153955935	154005934	50kN	1	50000	+	50000	1	50000
242	154005935	157953789	GA_x5J8B7Q6WBE_6	1	3947855	+	3947855	7	8078

**Table S4.** Differences in DNA sequence assemblies of chromosome 7. "Unmatched sequence" is a DNA sequence that is present in the CRA\_TCAGchr7.v1 or NCBI Build 31 with no corresponding sequence at the same relative position in the other. "Sequence variation" is DNA sequence present in the CRA\_TCAGchr7.v1 or NCBI Build 31 assembly that has a different sequence in the same relative position in the other. Bracketed numbers indicate the number of intrascaffold gaps contained within the unmatched sequences identified and the estimated nucleotide content they encompass. Because of intrascaffold sequence gaps and variation in size of the unique sequences, the total amount of sequence variation detected between the two assemblies will not be equal. Ten segments were identified between the two assemblies that contained the same DNA sequence at the same relative site but in an inverted orientation. The sizes of the inversions are 770, 1,185, 2,001, 5,272, 5,459, 12,686, 22,616, 30,037, 56,345 and 121,729 nt. The complete dataset listing all differences detected is available at <http://www.chr7.org>. As expected, the PatternHunter analysis found that CRA\_TCAGchr7.v1 and NCBI Build 31 had identical sequences between the 209 BAC clones used in each.

Range (bp)	Unmatched sequence				Sequence variation			
	+CRA_TCAGchr7.v1		+NCBI Build 31		CRA_TCAGchr.v1		NCBI Build 31	
	No.	bp	No.	bp	No.	bp	No.	bp
<100	85 (5)	1,857 (108)	65	2,210	45 (2)	1,140 (40)	52	1,421
100–999	59 (18)	20,230 (360)	79	28,532	50 (31)	22,003 (3,569)	46	16,163
1,000–9,999	14 (1)	42,045 (1,113)	30	86,391	27 (20)	92,673 (56,675)	26	85,676
10,000–99,999	5 (13)	229,907 (20,355)	3	47,486	10 (13)	392,516 (49,065)	8	182,391
≥100,000	4 (31)	728,255 (87,319)	0	0	0 (0)	0 (0)	0	0
Total	167 (68)	1,022,294 (109,255)	177	164,619	132 (66)	508,332 (109,349)	132	285,651

**Table S5a.** List of 21,859 syntenic anchor points. Refer to larger file for all sequences and identifiers. The header before each sequence details the name of the anchor (CRA|hmSA...), the unique syntenic anchor identifier (sa\_uid=...), organism (*Homo sapiens*), unique Celera scaffold Identifier (ga\_uid=...) alignment within the Celera scaffold (scf align=...), the chromosome 7 alignment in Celera assembly (chr7 align =). The next section details the information for the mouse. Syntenic anchor unique Id, organism, unique Celera mouse scaffold identifier, mouse Celera scaffold alignment coordinates, mouse chromosome, and Celera mouse chromosome alignment coordinates. This is followed by the sequence in FASTA format. [Link to Table 5a here]

**Table S5b.** Blocks of conserved synteny between human chromosome 7 and the mouse genome. Syntenic blocks were defined as regions of orthologous DNA encompassed by at least 10 consecutive syntenic anchors aligned in the same continuous orientation. A break of synteny is a jump of greater than 1 Mb in the corresponding mouse chromosome sequence, a change in orthologous chromosome, or a change in orientation of alignment (inversion) with the neighboring syntenic block. A total of 36 mouse synteny blocks (35 breaks in synteny) were identified.

Chr7 Start Coordinate	Chr7 End Coordinate	Length	Orientation	Mouse Chr	Mouse Start Coordinate	Mouse End Coordinate	Length
1	2601237	2601236	-	5	132777741	134675959	1898218
2601238	2733078	131840	+	5	134778640	134681192	97448
2733079	5959499	3226420	-	5	134802969	137158488	2355519
5959500	7003407	1043907	+	5	138007949	137284703	723246
7003408	12666896	5663488	-	6	4961557	10366641	5405084
12666897	22687170	10020273	+	12	38562457	29343939	9218518
22687171	22939941	252770	+	5	24371468	24530267	158799
22939942	23383958	444016	-	5	17821513	18143575	322062
23383959	33029045	9645086	+	6	46530862	54560566	8029704
33029046	36468691	3439645	+	9	16761250	19865347	3104097
36468692	36636017	167325	-	9	16714387	16603219	111168
36636018	43860101	7224083	+	13	18288275	11533702	6754573
43860102	53314879	9454777	+	11	3197799	12084996	8887197
53314880	54019071	704191	-	11	12915367	12222820	692547
54019072	55785691	1766619	+	11	12984533	14285847	1301314
56204509	57273981	1069473	-	5	123760291	124511348	751057
64537706	65902874	1365168	-	5	123813487	124351993	538506
65902875	71455213	5552338	+	5	128046743	124280123	3766620
71455214	73542272	2087058	+	5	129395788	128168037	1227751
73542273	74981009	1438736	-	5	129398591	130045997	647406
74981009	81119663	6138654	+	5	15419585	9733554	5686031
81137847	84069994	2932147	-	12	22239707	19349788	2889919
84069995	90270805	6200810	-	5	7776558	1141338	6635220
90270806	90539540	268734	+	5	845457	1022114	176657
90539541	91446058	906517	-	5	794881	195763	599118
91446059	96303447	4857388	-	6	522392	4797387	4274995
96303448	98554334	2250886	-	5	138006726	139285757	1279031
98554335	100891170	2336835	+	5	131798383	130143630	1654753
100891171	103854898	2963727	-	5	15402465	17768168	2365703
103854899	106405991	2551092	+	12	31827242	29593469	2233773
106405992	110777289	4371297	+	12	42964399	38605813	4358586
110777290	147090204	36312914	-	6	10390573	44493899	34103326
147090205	149000925	1910720	+	6	45329049	46418592	1089543
149000926	151087512	2086586	-	5	18239080	19842075	1602995
151087513	155528034	4440521	+	5	20697112	24274339	3577227
155528035	157953789	2471965	-	12	66493706	65229425	1264281

**Table S6.** List of overlapping transcripts. The position of the two overlapping transcriptional units and their relative orientation to each other is described (determined by NCBI Blast-2-seq). (Imprinted genes are shown in bold, with overlapping transcript pairs shaded grey). The genomic context of the gene structures are displayed in the Genome Browser at <http://www.chr7.org>.

Chromosome Position	Orientation	Gene Symbol	Gene Class	Chromosome Position	Orientation	Gene Symbol	Gene Class	Category
Chr7:1072186..1212070	-	MGC11257	Known_Gene	Chr7:1132972..1134729	+	LOC115330	Known_Gene	Structurally overlapping
Chr7:1072186..1212070	-	MGC11257	Known_Gene	Chr7:1161880..1168792	+	GPR30	Known_Gene	Structurally overlapping
Chr7:1616251..1618687	-	FLJ39498	Known_Gene	Chr7:1618665..1628730	-	FLJ90562	Known_Gene	Sense overlapping
Chr7:1888729..2299454	-	MAD1L1	Known_Gene	Chr7:1911520..1922768	+	FLJ38166	Novel_Gene	Structurally overlapping
Chr7:4922852..5011853	-	KIAA1849	Known_Gene	Chr7:4985936..4990193	-	PAPOLB	Known_Gene	Structurally overlapping
Chr7:5434992..5499425	-	KIAA1856	Known_Gene	Chr7:5441309..5448806	-	TNRC18	Known_Gene	Structurally overlapping
Chr7:6137439..6152023	+	JTV1	Known_Gene	Chr7:6150437..6187349	-	HRI	Known_Gene	Antisense overlapping
Chr7:21678924..220375	+	DNAH11	Known_Gene	Chr7:22036667..2208162	-	FLJ38646	Known_Gene	Antisense overlapping
Chr7:22254001..223295	-	GFR	Known_Gene	Chr7:22327755..2249249	-	DKFZp667A016	Known_Gene	Sense overlapping
Chr7:27274711..272954	-	HOXA3_variant_2	Known_Gene	Chr7:27283871..2729148	+	FLJ31668	Novel_Gene	Antisense overlapping
Chr7:27283871..272918	+	FLJ31668	Novel_Gene	Chr7:27290440..2729724	+	IMAGE:5180590	Known_Gene	Sense overlapping
Chr7:27290440..27297284	+	IMAGE:5180590	Known_Gene	Chr7:27297056..27299257	-	HOXA4	Known_Gene	Antisense overlapping
Chr7:27314014..27316267	-	HOXA6	Known_Gene	Chr7:27315680..27322829	+	FLJ34614	Known_Gene	Antisense overlapping
Chr7:32658122..32746050	+	FLJ33300	Known_Gene	Chr7:32742821..32878692	-	LOC89231	Known_Gene	Structurally overlapping
Chr7:36479947..36545804	+	KIAA0895	Known_Gene	Chr7:36545588..36608835	+	ANLN	Known_Gene	Antisense overlapping
Chr7:43255311..43706073	+	KIAA0322	Known_Gene	Chr7:43260611..43305907	+	FLJ35943	Novel_Gene	Structurally overlapping
Chr7:47918480..48086584	-	PKD1L1	Known_Gene	Chr7:47938691..47963215	+	FLJ21075	Known_Gene	Structurally overlapping
Chr7:47918480..48086584	-	PKD1L1	Known_Gene	Chr7:47973969..47976209	+	FLJ34738	Novel_Gene	Antisense overlapping
Chr7:50628221..50735080	-	DDC	Known_Gene	Chr7:50701496..50713198	+	FLJ32838	Novel_Gene	Antisense overlapping
Chr7:50759821..50962692	-	<b>GRB10</b>	Known_Gene	Chr7:50952775..50955639	+	FLJ40093	Novel_Gene	Structurally overlapping
Chr7:63552100..63556800	-	ZNF117	Known_Gene	Chr7:63553297..63566352	-	H-plk	Known_Gene	Sense overlapping
Chr7:64450022..64531566	+	IMAGE:5093832	Known_Gene	Chr7:64511168..64514965	+	FLJ31298	Novel_Gene	Structurally overlapping
Chr7:65573825..65816501	+	SDCR2A	Known_Gene	Chr7:65769747..65876413	-	hPMS6	Known_Gene	Structurally overlapping
Chr7:69704302..69706697	-	FLJ39672	Novel_Gene	Chr7:69706185..70286910	+	WBSCR17	Known_Gene	Antisense overlapping
Chr7:71457198..71529242	+	POM121	Known_Gene	Chr7:71526094..71532522	-	WBSCR20C_variant_1	Known_Gene	Antisense overlapping
Chr7:71457198..71529242	+	POM121	Known_Gene	Chr7:71526094..71527969	-	WBSCR20C_variant_4	Known_Gene	Antisense overlapping
Chr7:73715487..73913228	-	PMS2L14	Known_Gene	Chr7:73745682..73761094	-	hPMS4	Known_Gene	Sense overlapping
Chr7:73715487..73913228	-	PMS2L14	Known_Gene	Chr7:73794064..73797525	-	BC022013	Novel_Gene	Structurally overlapping
Chr7:73715487..73913228	-	PMS2L14	Known_Gene	Chr7:73812721..73818901	+	WBSCR20B	Known_Gene	Structurally overlapping
Chr7:73715487..73913228	-	PMS2L14	Known_Gene	Chr7:73909902..73930009	-	PMS2L9	Known_Gene	Sense overlapping
Chr7:73715487..73913228	-	PMS2L14	Known_Gene	Chr7:73909903..73916945	-	PMS5	Known_Gene	Sense overlapping
Chr7:75524766..75697354	+	IMAGE:5273288	Known_Gene	Chr7:75598324..75601977	-	FGL2	Known_Gene	Structurally overlapping
Chr7:76419229..77855250	-	AIP1	Known_Gene	Chr7:76749041..76761258	+	MGC:34774	Novel_Gene	Structurally overlapping
Chr7:76419229..77855250	-	AIP1	Known_Gene	Chr7:77716248..77717105	-	AF521131	Partial_Gene	Structurally overlapping
Chr7:85554013..85597925	+	DMTF1	Known_Gene	Chr7:85597761..85621315	-	MGC4175	Known_Gene	Antisense overlapping
Chr7:85905465..86114658	-	ABCB1	Known_Gene	Chr7:86029816..86231847	+	RPIB9	Known_Gene	Structurally overlapping

Chromosome Position	Orientation	Gene Symbol	Gene Class	Chromosome Position	Orientation	Gene Symbol	Gene Class	Category
Chr7:86237517..86277761	-	MCFP	Known_Gene	Chr7:86277635..86309783	+	ASK	Known_Gene	Antisense overlapping
Chr7:87160633..87736011	+	FLJ32110	Novel_Gene	Chr7:87195300..87196912	-	MGC:26647	Known_Gene	Structurally overlapping
Chr7:92191760..92309350	+	GNGT1	Known_Gene	Chr7:92284110..92289030	-	TFPI2	Known_Gene	Antisense overlapping
Chr7:93308678..96690908	+	<b>PPP1R9A*</b>	Known_Gene	Chr7:93552675..93611807	-	FLJ33602	Novel_Gene	Structurally overlapping
Chr7:96688452..96795016	-	IMAGE:3842949	Known_Gene	Chr7:96703094..96705651	+	FLJ30064	Novel_Gene	Structurally overlapping
Chr7:96688452..96795016	-	IMAGE:3842949	Known_Gene	Chr7:96779934..96780796	-	IMAGE:3350750	Novel_Gene	Structurally overlapping
Chr7:97273047..97407740	+	TRRAP	Known_Gene	Chr7:97407688..97431013	+	FLJ10671	Novel_Gene	Sense overlapping
Chr7:97407688..97431013	+	FLJ10671	Novel_Gene	Chr7:97421976..97539001	-	KIAA1625	Known_Gene	Antisense overlapping
Chr7:97807643..97818271	+	G10	Known_Gene	Chr7:97815393..97837402	-	KIAA0632	Known_Gene	Antisense overlapping
Chr7:98404777..98420049	+	ZNF38	Known_Gene	Chr7:98418857..98432430	-	ZNF3_variant_1	Known_Gene	Antisense overlapping
Chr7:98456566..98462188	+	AP4M1	Known_Gene	Chr7:98462087..98474366	-	TAF6_variant_1	Known_Gene	Antisense overlapping
Chr7:98532695..98569357	+	STAG3	Known_Gene	Chr7:98537697..98595329	+	DKFZp434F086	Known_Gene	Sense overlapping
Chr7:98532695..98569357	+	STAG3	Known_Gene	Chr7:98563914..98568111	-	FLJ34099	Novel_Gene	Antisense overlapping
Chr7:98537697..98595329	+	DKFZp434F086	Known_Gene	Chr7:98563914..98568111	-	FLJ34099	Novel_Gene	Structurally overlapping
Chr7:98537697..98595329	+	DKFZp434F086	Known_Gene	Chr7:98585680..98587206	+	PILRB	Known_Gene	Sense overlapping
Chr7:98766464..98792864	+	RAB-R	Known_Gene	Chr7:98792858..98795434	+	IRS3L	Known_Gene	Sense overlapping
Chr7:98813600..98828329	+	FBXO24_variant_1	Known_Gene	Chr7:98816614..98831251	-	FLJ40386	Novel_Gene	Antisense overlapping
Chr7:98816614..98831251	-	FLJ40386	Novel_Gene	Chr7:98829609..98835384	+	PCOLCE	Known_Gene	Antisense overlapping
Chr7:99182807..99243180	+	MUC3B	Known_Gene	Chr7:99238438..99242971	-	FLJ39484	Novel_Gene	Antisense overlapping
Chr7:100637811..100671335	+	FLJ40957	Novel_Gene	Chr7:100670292..100700596	+	FLJ13902	Known_Gene	Sense overlapping
Chr7:100815465..100846548	-	POLR2J2B	Known_Gene	Chr7:100831956..100834744	+	TCAG_hCT1815882	Partial_Gene	Structurally overlapping
Chr7:101087224..101348560	-	FLJ40218	Known_Gene	Chr7:101186996..101218941	+	P37NB	Known_Gene	Antisense overlapping
Chr7:101087224..101348560	-	FLJ40218	Known_Gene	Chr7:101247513..101262848	+	IMAGE:5404753	Partial_Gene	Structurally overlapping
Chr7:101449124..101554211	-	DKFZp434E092	Known_Gene	Chr7:101535991..101536300	+	S100A14	Known_Gene	Structurally overlapping
Chr7:101571443..101586551	+	PMPCB	Known_Gene	Chr7:101586465..101618658	-	ZRF1	Known_Gene	Antisense overlapping
Chr7:103805949..103841380	+	FLJ11785	Known_Gene	Chr7:103838843..103855426	-	TCAG_hCT1818545	Known_Gene	Antisense overlapping
Chr7:105477292..105838353	-	COG5	Known_Gene	Chr7:105748136..105749992	+	GPR22	Known_Gene	Structurally overlapping
Chr7:106298393..106390534	-	TCAG_hCT1953978	Known_Gene	Chr7:106336732..106340213	-	LAMB4	Known_Gene	Sense overlapping
Chr7:108943035..109843890	-	IMMP2L	Known_Gene	Chr7:109371032..109403545	+	FLJ11129	Partial_Gene	Structurally overlapping
Chr7:108943035..109843890	-	IMMP2L	Known_Gene	Chr7:109691260..109691761	+	TCAG_hCT1645734	Known_Gene	Structurally overlapping
Chr7:115256868..115258757	-	ST7OT1	Known_Gene	Chr7:115257763..115534422	+	RAY1_variant_a	Known_Gene	Antisense overlapping
Chr7:115257763..115534422	+	RAY1_variant_a	Known_Gene	Chr7:115417004..115450273	-	ST7OT2_variant_1	Known_Gene	Antisense overlapping
Chr7:115324720..115534500	+	ST7_variant_a	Known_Gene	Chr7:115487394..115514336	+	ST7OT3	Known_Gene	Sense overlapping
Chr7:120624366..121185595	-	CADPS2	Known_Gene	Chr7:121002696..121004139	-	LOC168433	Known_Gene	Structurally overlapping
Chr7:120624366..121185595	-	CADPS2	Known_Gene	Chr7:121006650..121007952	-	MGC:35222	Novel_Gene	Structurally overlapping
Chr7:125945922..126386024	+	p100	Known_Gene	Chr7:126320491..126324370	-	NAG14	Known_Gene	Structurally overlapping
Chr7:127156485..127174958	-	FLJ33365	Novel_Gene	Chr7:127156890..127159883	+	ATP6V1F	Known_Gene	Antisense overlapping
Chr7:128128692..128247317	-	UBE2H	Known_Gene	Chr7:128201097..128204032	+	CATR1	Known_Gene	Structurally overlapping
Chr7:128641943..128664190	+	CPA5	Known_Gene	Chr7:128662808..128665006	-	FLJ40591	Known_Gene	Antisense overlapping

Chromosome Position	Orientation	Gene Symbol	Gene Class	Chromosome Position	Orientation	Gene Symbol	Gene Class	Category
Chr7:128781654..128801771	+	<b>MEST_isoform_2*</b>	Known_Gene	Chr7:128782536..128786652	-	<b>MESTIT1*</b>	Known_Gene	Structurally overlapping
Chr7:128787575..128801775	+	<b>MEST*</b>	Known_Gene	Chr7:128801723..128959479	-	<b>COPG2</b>	Known_Gene	Antisense overlapping
Chr7:128801723..128959479	-	<b>COPG2</b>	Known_Gene	Chr7:128835019..128836908	+	<b>COPG2IT1*</b>	Known_Gene	Structurally overlapping
Chr7:130419839..130520324	-	KIAA1550	Known_Gene	Chr7:130419869..130451489	-	FLJ38287	Novel_Gene	Sense overlapping
Chr7:133451169..133468997	+	FLJ11000	Known_Gene	Chr7:133468883..133472160	-	MGC5242	Known_Gene	Antisense overlapping
Chr7:133488876..133514610	-	HSPC049	Known_Gene	Chr7:133501507..133503280	+	FLJ32860	Novel_Gene	Structurally overlapping
Chr7:135171761..135323346	+	CHRM2	Known_Gene	Chr7:135201859..135467125	-	FLJ40151	Partial_Gene	Structurally overlapping
Chr7:137009206..137101126	-	ATP6V0A4_variant_1	Known_Gene	Chr7:137100938..137108170	+	FLJ31279	Partial_Gene	Antisense overlapping
Chr7:137663554..137726735	+	CGI-74	Known_Gene	Chr7:137692467..137696262	+	pp12708	Novel_Gene	Structurally overlapping
Chr7:137864851..137876148	-	THC1179197	Known_Gene	Chr7:137875909..138079218	-	HIPK2	Known_Gene	Sense overlapping
Chr7:140056098..140089334	-	FLJ40852	Known_Gene	Chr7:140060525..140082251	+	THC1158528	Novel_Gene	Antisense overlapping
Chr7:141804882..141822641	-	EPHA1	Known_Gene	Chr7:141821559..141937177	+	DKFZp686O0656	Known_Gene	Structurally overlapping
Chr7:141821559..141937177	+	DKFZp686O0656	Known_Gene	Chr7:141891588..141892510	+	TAS2R41	Novel_Gene	Structurally overlapping
Chr7:147379600..147411928	+	TCAG_hCT1964447	Known_Gene	Chr7:147411583..147414687	+	DKFZp667J212	Known_Gene	Sense overlapping
Chr7:147664566..147742112	-	FLJ12700	Known_Gene	Chr7:147704260..147717316	+	FLJ32307	Novel_Gene	Structurally overlapping
Chr7:147891625..147945454	+	TCAG_hCT7210	Known_Gene	Chr7:147941961..147951378	+	FLJ36112	Known_Gene	Sense overlapping
Chr7:148900024..148910003	-	LR8	Known_Gene	Chr7:148909480..148913846	+	HCA112	Known_Gene	Structurally overlapping
Chr7:149102476..149123254	+	NOS3	Known_Gene	Chr7:149121610..149125803	-	FLJ14885	Novel_Gene	Antisense overlapping
Chr7:149195709..149253389	+	CENTG3	Known_Gene	Chr7:149230276..149232602	+	FLJ34452	Known_Gene	Structurally overlapping
Chr7:149450714..149487249	+	NUB1	Known_Gene	Chr7:149480687..149483411	-	FLJ32062	Known_Gene	Antisense overlapping
Chr7:149490417..149521030	-	TCAG_hCT10662	Partial_Gene	Chr7:149491797..149493723	-	LOC285975	Known_Gene	Structurally overlapping
Chr7:153025713..153047094	+	FLJ32047	Novel_Gene	Chr7:153040875..153073350	-	PAXIP1L	Known_Gene	Antisense overlapping
Chr7:153164556..153169018	-	IMAGE:5287138	Partial_Gene	Chr7:153168387..153181993	+	HTR5A	Known_Gene	Antisense overlapping
Chr7:155645692..156663374	-	PTPRN2_variant_1	Known_Gene	Chr7:155956878..155967157	+	IMAGE:4475530	Novel_Gene	Structurally overlapping
Chr7:157102829..157218534	-	VIPR2	Known_Gene	Chr7:157106016..157108168	+	IMAGE:4903629	Novel_Gene	Antisense overlapping

**Table S7.** Twenty regions on human chromosome 7 and the syntenic intervals in mouse devoid of genes (gene deserts) (Tables S7a and S7b). For this study a gene desert was defined as a region with no known, novel, or partial genes in a 500-kb region. The control regions examined are shown (Tables S7c, S7d, S7e). In addition to the randomly selected genomic regions (described in the published text), as an additional control, we examined the genomic intervals encompassed by large genes (>500 kb) on chromosome 7 (Table S7f). In all cases, the only correlation was with low CpG density. The orthologous genes in mouse and syntenic anchor points were used to identify the equivalent regions in the murine genome. Table S7b shows the orthologous murine gene that flanks the region. If an orthologous gene was not yet defined, the nearest known gene flanking the region was selected (the closest EST or syntenic anchor marker to the boundary is also listed in brackets). When syntenic anchors are used, the corresponding region in the public mouse assembly (UCSC) can be identified by retrieving the syntenic anchor sequence from the chromosome 7 Genome Browser (<http://www.chr7.org>), and searching the mouse assembly. The best sequence alignment (using BLAT) will represent the border of the mouse region equivalent to the human gene desert. For our analysis of mouse, we examined both the UCSC (results shown in Table S7b) and the Celera assemblies separately. In 19 of 20 cases, the results were equivalent. In one instance (human desert #7), a break in synteny occurred, and UCSC placed both mouse segments on chromosome 5 (Table S7b), whereas Celera positioned them on chromosomes 12 and 5. Notwithstanding, using our criteria for all 20 regions would still characterize them as deserts in mouse. The location of the human gene deserts along chromosome 7 can also be observed in the "Structural Feature" track in the Genome Browser at <http://www.chr7.org>.

(Table S7a) Putative Gene Deserts on Chromosome 7

Gene Desert	Size (kb)	Location	Flanking Reference Genes		# of predicted, putative, and pseudogenes	CpG Island	CpG Island/Mb	% Syntenic (>75%)	Repetitive Content	
									LINES	SINES
1	1850	7q11.22-q11.23	FLJ13195	AUTS2	2 predicted	1	0.5	4.2%	10.8%	25.8%
2	1740	7q31.1	THC1201470	IMMP2L	3 predicted; 1 pseudogene	0	0.0	4.6%	32.9%	5.2%
3	1700	7p12.2-p12.1	KIAA0633	FLJ40449	2 predicted; 1 putative; 1 pseudogene	4	2.4	2.9%	26.0%	8.1%
4	1690	7p22.1-p21.3	NXPPI	IMAGE:3605453	2 predicted; 1 putative; 1 pseudogene	3	1.8	4.7%	31.2%	5.3%
5	1640	7q31.31-q31.32	ANKRD7	hCT1816883	1 pseudogene	2	1.2	2.1%	32.2%	4.2%
6	1190	7p21.3	ARL4	ETV1	1 pseudogene	2	1.7	5.6%	24.0%	8.7%
7	1040	7q21.11-q21.13	IMAGE:5272175	GRM3	1 predicted	1	1.0	2.1%	33.8%	4.9%
8	990	7p12.3-p12.2	MGC26484	ZBPB	none	1	1.0	2.6%	31.0%	6.0%
9	950	7q31.33-q31.2	THC1079110	GRM8	1 predicted	0	0.0	3.1%	28.7%	5.1%
10	900	7q36.1-q36.2	ARP3BETA	DPP6	2 predicted	2	2.2	1.9%	24.3%	8.4%
11	850	7p13-p12.3	IGFBP3	PRO1866	2 predicted; 1 pseudogene	1	1.2	1.5%	31.4%	5.8%
12	810	7q31.2-q31.31	IMAGE:4276820	TFEC	1 putative	0	0.0	11.8%	19.9%	6.1%
13	770	7q21.2	FLJ32110	IMAGE:5295327	1 predicted	1	1.3	4.8%	33.6%	5.2%
14	730	7q31.2	GPR85	PPP1R3	none	1	1.4	4.1%	31.4%	5.1%
15	720	7q35	TPK1	THC1203597	3 pseudogenes	1	1.4	4.1%	29.7%	6.3%
16	660	7p14.1-p13	GLI3	MGC2821	1 predicted	0	0.0	4.1%	31.1%	6.5%
17	610	7q21.11	AIP1	GNAI1	1 predicted; 1 pseudogene	0	0.0	2.4%	30.0%	6.3%
18	540	7p21.2-p21.1	FERD3L	LOC221830	1 predicted	0	0.0	6.5%	35.4%	5.2%
19	540	7p14.1	BC033981	INHBA	none	0	0.0	6.6%	16.6%	9.7%
20	540	7q21.3-q22.1	DC11	TAC1	2 predicted	0	0.0	3.3%	28.7%	12.2%
Total (kb)						Average	0.8	4.2%	28.1%	7.5%
20460						Standard Deviation	0.8	2.3%	6.3%	4.7%

**(Table S7b) Mouse Regions Syntenic to Gene Deserts on Chromosome 7 (Data shown are from analysis on UCSC mouse sequence)**

Gene Desert	Mouse Size (kb)	Mouse Location	Flanking Reference Genes	# of known genes or spliced EST clusters	CpG Island	CpG Island/Mb	% Syntenic (>75%)	Repetitive LINES	Content SINES	
1	1488	chr5	BC021509	Gats (hmSA93056)	none	1	0.7	6.4%	7.8%	16.2%
2	1951	chr12	Imp2l-pending	Dnajb9 (hmSA72264)	none	1	0.5	4.2%	31.5%	2.9%
3	2128	chr11	U26967	Sec61g (hmSA220916)	none	3	1.4	2.3%	33.8%	2.6%
4	2216	chr6	Ica1 (BB641832)	BC011114 (hmSA167586)	2 spliced EST clusters	4	1.8	3.8%	39.8%	2.3%
5	2019	chr6	AW214405 (hmSA16662)	Kcnd2 (hmSA15769)	none	1	0.5	1.8%	39.3%	2.2%
6	1138	chr12	Etv1	Arl4	none	0	0.0	5.8%	27.5%	4.0%
7	172	chr5	Cdk6 (hmSA322558)	Sema3a (BB451280)	none	0	0.0	4.8%	33.8%	3.4%
8	1117	chr5	Telomere (BB871298)	Png (hmSA322547)	none	3	2.7	1.9%	34.6%	2.6%
9	1239	chr11	Rpo2-3 (hmSA220133)	Zpbp	1 spliced EST cluster	3	2.4	2.1%	35.4%	2.5%
10	972	chr6	AI851169 (hmSA138658)	Gprc1h	none	0	0.0	3.2%	34.9%	2.3%
11	967	chr5	Xrcc2 (BI248187)	AF092507	none	0	0.0	1.8%	15.0%	4.5%
12	740	chr11	Igfbp3	Spin (hmSA193672)	none	1	1.4	1.8%	23.7%	2.9%
13	1003	chr6	Foxp2 (hmSA23596)	Tcfec	none	0	0.0	9.6%	25.1%	3.5%
14	1010	chr5	Tiarp-pending (hmSA318744)	Hspa8	1 known: M36516 (homology to chr19)	0	0.0	4.1%	32.7%	5.1%
15	833	chr6	GPR85	Ppp1r3a	none	0	0.0	4.1%	32.6%	3.2%
16	974	chr6	Tpk1	Rbpsuh (hmSA166085)	1 spliced EST cluster	1	1.0	3.8%	30.5%	4.2%
17	830	chr13	AW209491	Gli3	1 spliced EST cluster	1	1.2	3.4%	35.6%	2.9%
18	841	chr5	Cd36	Acvrp1-pending (hmSA354360)	1 spliced EST cluster	1	1.2	1.9%	32.4%	3.0%
19	493	chr12	BC017546 (BI689342)	Nato3-pending	none	1	2.0	7.4%	30.8%	3.7%
20	628	chr13	Inhba	AF397014 (hmSA45738)	none	1	1.6	6.2%	25.7%	2.4%
20	521	chr6	Dlx5 (BB645073)	Tac1	1 spliced EST cluster	0	0.0	3.5%	31.1%	3.1%
Total (kb)					Average	0.9	4.0%	30.2%	3.8%	
23280					Standard Deviation	0.9	2.1%	7.6%	2.9%	

**(Table S7c) Random 1 Mb Control Regions (20 from Chromosome 7 and 5 from Other Chromosomes)**

Size (kb)	Location	Flanking Genes	Genes and Models in Region	CpG Island	CpG Island/Mb	Syntenic Coverage	Repetitive Content LINES	SINES
1000	7q32.3	N/A	5 known; 5 putative; 3 predicted	6	6	5.00%	23.50%	17.70%
1000	7q22.1	N/A	5 known; 1 novel; 1 putative	11	11	5.30%	7.60%	35.30%
1000	7q21.2	N/A	6 known; 1 novel; 3 putative	10	10	5.70%	28.80%	12.50%
1000	7q11.23	N/A	4 known; 4 novel; 1 predicted	7	7	6.30%	17.30%	19.70%
1000	7q11.23	N/A	2 known	2	2	2.80%	14.30%	24.20%
1000	7p14.3	N/A	5 known	10	10	5.40%	21.00%	12.20%
1000	7p15.1	N/A	10 known; 3 predicted	6	6	8.00%	19.00%	9.10%
1000	7p15.2	N/A	18 known; 1 novel; 1 partial; 1 putative; 9 predicted; 1 pseudogenes	36	36	14.50%	16.80%	14.50%
1000	7p22.1	N/A	14 known; 1 partial; 4 putative; 6 predicted	35	35	4.40%	8.10%	36.00%
1000	7q35	N/A	1 known; 1 putative; 2 pseudogenes	0	0	5.00%	20.60%	8.20%
1000	7p14.2	N/A	1 known; 1 putative; 4 predicted	3	3	6.00%	21.10%	9.30%
1000	7q21.3	N/A	5 known; 2 putative; 4 predicted	3	3	9.60%	25.90%	8.50%
1000	7q11.23	N/A	20 known; 2 predicted	35	35	5.30%	8.90%	42.40%
1000	7q32.1	N/A	8 known; 1 novel; 2 putative; 3 predicted	12	12	14.20%	15.50%	11.20%
1000	7q33	N/A	4 known; 1 partial; 2 putative; 1 predicted	2	2	7.60%	15.50%	11.50%
1000	7p21.3	N/A	1 known; 1 novel	2	2	5.40%	20.70%	6.70%
1000	7q21.11	N/A	2 known; 1 partial; 1 putative; 1 predicted	3	3	5.30%	18.30%	7.10%
1000	7q22.1	N/A	8 known; 3 predicted; 1 pseudogene	6	6	8.60%	22.10%	12.60%
1000	7p13	N/A	20 known; 4 putative; 3 predicted	22	22	7.20%	16.40%	23.10%
1000	7q31.32	N/A	5 known; 1 novel; 1 putative; 5 predicted	7	7	8.90%	20.70%	8.10%
1000	1p32.2	N/A	4 known	2	2	12.00%	16.50%	14.10%
1000	3p22.2	N/A	7 known	4	4	1.50%	31.50%	9.40%
1000	5p12	N/A	2 known	0	0	8.70%	31.80%	5.20%
1000	10q26.12	N/A	3 known	5	5	9.50%	14.60%	11.70%
1000	12q13.12	N/A	15 known	17	17	8.00%	12.50%	30.40%
Total				Average	9.8	7.20%	18.80%	16.00%
25000				Standard Deviation	10.9	3.10%	6.40%	10.20%

**(Table S7d) Putative Gene Deserts on Other Chromosomes**

Size (kb)	Location	Flanking Genes		# of predicted, putative, and pseudogenes	CpG Island	CpG Island/Mb	Syntenic Coverage	Repetitive Content	
								LINES	SINEs
1245	14q31.2	SEL1L	FLRT2	1 putative	12	9.6	4.00%	19.50%	6.90%
687	20q12	DDX35	MAFB	N/A	0	0	8.50%	20.60%	9.50%
2594	21q21.1	PRSS7	PRED14	1 predicted; 5 pseudogenes	0	0	2.80%	21.80%	6.40%
3269	21q21.2	PRED16	C21ORF42	1 predicted; 6 pseudogenes	0	0	1.90%	21.10%	7.80%
Total					Average	2.4	4.30%	20.80%	7.70%
7804					Standard Deviation	4.8	2.90%	1.00%	1.40%

**(Table S7e) FRA7H Region on Chromosome 7**

Size (kb)	Location	Flanking Genes		Genes and Models in Region	CpG Island	CpG Island/Mb	Syntenic Coverage	Repetitive Content	
								LINES	SINEs
375	7q32.3	BTEB5	MKLN1	5 putative; 1 predicted	1	2.7	2.40%	24.90%	15.50%

**(Table S7f) Large Genes (>500 kb) on Chromosome 7**

Size (kb)	Location	Large Gene	Genes and Models in Region	CpG Island	CpG Island/Mb	Syntenic Coverage	Repetitive Content	
							LINES	SINEs
2305	7q35	CNTNAP2	1 known; 1 putative	2	0	5.30%	18.40%	9.50%
1436	7q11.23	AIP1	1 known; 1 novel; 1 partial	0	0	6.50%	20.40%	8.40%
1194	7q11.23	AUTS2	1 known	3	0	11.70%	11.50%	13.70%
1018	7q36.3	PTPRN2	1 known; 1 novel; 2 predicted	35	0	2.40%	10.80%	4.30%
901	7q31.1	IMMP2L	2 known; 1 partial; 1 putative	1	0	8.60%	20.40%	7.40%
830	7q36.2	DPP6	1 known; 1 predicted	3	0	2.30%	11.10%	9.70%
812	7q33	SEC8	1 known; 1 putative	0	0	10.40%	17.50%	10.90%
805	7q32.1	GRM8	1 known	1	0	10.00%	21.10%	6.80%
727	7p14.1	C7orf10	1 known; 1 putative	0	0	7.00%	19.20%	15.30%
694	7p21.3	DGKB	1 known	0	0	7.60%	20.80%	6.30%
629	7q11.23	CALN1	1 known; 1 putative; 1 predicted	1	0	1.80%	17.20%	24.90%
611	7q21.2	PFTK1	1 known	1	0	7.50%	21.90%	7.90%
594	7p14.2	ELMO1	1 known; 1 putative	2	0	6.80%	15.80%	10.20%
581	7q11.23	WBSCR17	1 known; 1 novel	0	0	3.50%	13.70%	24.30%
546	7p15.1	PDE1C	1 known	2	0	3.40%	22.90%	8.20%
517	7q22.1	RELN	1 known; 1 predicted	1	0	10.90%	18.40%	9.70%
Total				Average	0	6.60%	17.60%	11.10%
14201				Standard Deviation	0	3.20%	3.90%	5.90%

**Table S8.** Correlation of segmental duplications on human chromosome 7 with human-mouse syntenic breakpoint regions. All segmental duplications on chromosome 7 greater than 50 kb in size are listed in order from 7pter-7qter by their identifier name (column 1; note that iDup= intrachromosomal, xDup= transchromosomal duplication). The 35 human regions that encompass mouse synteny breakpoints are also listed from 7pter-qter (column 3) and the flanking hmSA-mouse syntenic anchor markers are provided. Comparison of the coordinates of the segmental duplications and the human regions where mouse synteny breaks occurred, identified 12 regions of overlap (shaded in grey). In five cases, there was more than one segmental duplication mapped within the region of overlap. The DNA sequences of the segmental duplications and the mouse syntenic anchor loci are available at <http://www.chr7.org>. The precise position and relationship of the two features along the chromosome are shown in the Genome Browser at the same Web site.

All Duplications > 50 kb	Duplication Size (bp)	All Syntenic Breakpoint Regions	Region Size (bp)	Syntenic Change
iDup_7_1	75148	hmSA196973/hmSA196974 hmSA197826/hmSA197827	9873 36161	Chr5 to Chr5 Chr5 to Chr5
iDup_7_6	67773			
iDup_7_7	127699	hmSA385722/hmSA258176	99284	Chr5 to Chr5
iDup_7_7	124171	hmSA13838/hmSA193154	354411	Chr5 to Chr6
xDup_7_46	68134			
		hmSA165224/hmSA165219	77844	Chr6 to Chr12
iDup_7_13	88594	hmSA103494/hmSA103488	127254	Chr12 to Chr5
		hmSA103436/hmSA103433 hmSA103265/hmSA103263	91109 30219	Chr5 to Chr5 Chr5 to Chr6
iDup_7_15	53696			
iDup_7_18	78246	hmSA16512/hmSA50080	558401	Chr6 to Chr9
iDup_7_15	52698			
iDup_7_25	156231			
iDup_7_25	140603			
iDup_7_26	57685			
iDup_7_27	55196			
		hmSA21555/hmSA21557 hmSA21689/hmSA21691 hmSA49953/hmSA49974	27915 55275 303408	Chr9 to Chr9 Chr9 to Chr13 Chr13 to Chr11
iDup_7_30	79903			
iDup_7_33	130719			
xDup_7_291	50995			
		hmSA220922/hmSA220923 hmSA221260/hmSA221261	80482 120854	Chr11 to Chr11 Chr11 to Chr11
iDup_7_23	83162	hmSA221963/hmSA222024	837747	Chr11 to Chr5
xDup_7_278	62771			
iDup_7_53	105421			
iDup_7_54	247337			
iDup_7_55	136827	hmSA353539/hmSA218883	2320	Chr5 to Chr5
xDup_7_291	93655			
xDup_7_293	145626			
xDup_7_58	72790			
xDup_7_293	169406			
xDup_7_294	104155			
iDup_7_66	117832			
iDup_7_68	53045			
iDup_7_68	53129			
iDup_7_55	137251			
iDup_7_54	303106			
iDup_7_83	206483			
iDup_7_84	233976			
xDup_7_154	79817			
iDup_7_90	52629			
iDup_7_98	72470			

iDup_7_99	69801			
iDup_7_100	82459			
iDup_7_102	85313	hmSA183574/hmSA93056	176434	Chr5 to Chr5
iDup_7_99	82485			
iDup_7_98	72992	hmSA183580/hmSA222083	124009	Chr5 to Chr5
iDup_7_105	297178			
iDup_7_102	61888			
iDup_7_108	100329			
iDup_7_109	95615			
iDup_7_105	310103	hmSA50171/hmSA222081	297629	Chr5 to Chr5
iDup_7_110	92980			
iDup_7_112	70803			
iDup_7_102	63407	hmSA93051/hmSA144376	596792	Chr5 to Chr5
xDup_7_24	56213			
iDup_7_118	456044			
iDup_7_13	88569	hmSA349369/hmSA349355	37036	Chr5 to Chr12
		hmSA322553/hmSA322551	73869	Chr12 to Chr5
		hmSA294688/hmSA294685	46347	Chr5 to Chr5
		hmSA294362/hmSA294360	35995	Chr5 to Chr5
xDup_7_101	120393	hmSA292493/hmSA292491	168222	Chr5 to Chr6
		hmSA258217/hmSA258179	160722	Chr6 to Chr5
		hmSA255156/hmSA255098	160304	Chr5 to Chr5
iDup_7_30	82395	hmSA107230/hmSA144371	288592	Chr5 to Chr5
iDup_7_30	93055			
xDup_7_44	57137			
xDup_7_44	92657	hmSA75979/hmSA75978	31795	Chr5 to Chr12
		hmSA73217/hmSA73216	33907	Chr12 to Chr12
		hmSA45576/hmSA45574	12273	Chr12 to Chr6
xDup_7_267	106776			
xDup_7_113	99917			
xDup_7_114	78802			
iDup_7_139	60920			
iDup_7_141	171405			
		hmSA190953/hmSA190955	62387	Chr6 to Chr6
iDup_7_145	333092			
iDup_7_147	71478			
		hmSA192437/hmSA192439	62521	Chr6 to Chr5
iDup_7_145	79571			
iDup_7_147	80546			
		hmSA219791/hmSA219793	11577	Chr5 to Chr5
iDup_7_145	77794			
iDup_7_145	209095			
		hmSA273843/hmSA273876	91623	Chr5 to Chr12

**Table S9.** Summary of clinical cases described in this research article. For each of these regions the chromosomal breakpoints are located in genomic intervals that were previously established in other studies to be critical regions for the disease phenotype in the patient. Therefore, genes near the breakpoints should be considered candidates to be involved in the disease. The complete list of over 1400 patients described in this study can be found at <http://www.chr7.org>. This list will be updated regularly with new information coming from the continuation of this study as well as contributions from the community.

Patient ID	Karyotype	Phenotype	Breakpoint #1	Breakpoint #2
8279479_IV-17	46, XX, t(2;7)(q21.1;q22.1)	unilateral split hand, normal feet, hearing loss		D7S527/ D7S1812
Unpublished_14282_B02253	46, XY, t(6;7)(p21.3;q22)	split hand/split foot		AZ757831/ AZ757826
Unpublished_11231-T10	47, XY, t(4;7)(p14;q21.2)	split hand/split foot		D7S2401/ D7S1812
Unpublished_14330	46, XY, t(6;7)(q36.3;q22)	phenotypically normal		D7S821/ D7S2324
1877619_T2.2_2	46, XY, t(5;9;7)(5pter-> 5q11.2::5q34-> 5qter;9pter-> 9q22.1::7q31.3-> 7q21.2::5q34-> 5q11.2::7q31.3-> 7qter;7pter-> 7q21.2::9q22.1-> 9qter)	bilateral split hand/split foot, mildly dysmorphic, low-set ears, normal cognitive development, ectrodactyly		stSG48566/ stsG13314
1773535_1	46, XX, t(7;9)(q21.3;p12)	syndactyly of right hand, bilateral split foot, light coloured sparse hair, high arched palate, abnormal ears		D7S479/ D7S1848
802340_T1	46, XX, t(7;12)(q22.1;q24.2)	bilateral split hand/split foot, mild speech delay		D7S491/ D7S624
11685205_15441_1	46, XX, inv(7)(q11.23q21.3) AND WBS INV-1	ectrodactyly, WBS facies, developmental delay, strabismus, WBS-like behavior profile, lordosis, chronic otitis media, normal growth, inattention	AC067941/ AC005074	AZ757826/ AZ757825
7987313_T6	46, XY, inv(1)(q2q3), t(4;7)(q2;q21.3), inv(11)(p15q23)	bilateral split hand/split foot, submucous cleft palate, deafness, mental retardation, microcephaly, abnormal ears		stSG48566/ stsG13314
7616545_T8	46, XY, inv(7)(p22q21.3)	bilateral split hand/split foot, haemangioma		D7S527/ D7S1812
8322806_T4	46, XY, inv ins (3;7)(q21;q34q22)	unilateral split hand, bilateral split foot, high arched palate, bifid uvula, normal cognitive development	stSG48566/ stSG13314	
Unpublished_17430	46, XY, WBS INV-II	undiagnosed anxiety disorder; daughter with Williams syndrome	D7S613/ D7S1870	D7S2490/ D7S1440
Unpublished_17495	46, XX, WBS INV-II	Williams syndrome-like, dysmorphic features, growth retardation, developmental delay, hypersensitivity to sound	D7S613/ D7S1870	D7S2490/ D7S1440
Unpublished_16724	46, XY, t(6;7)(p11.2;q22)	autism	AQ373444/ AQ373441	
Unpublished_18667	46, XX, t(7;11)(q31.2;q25)	autism	AC092062	
Unpublished_11550	46, XY, t(5;7)(q15;q31.32)	autism, developmental delay	AQ497751/ AQ409364	
Unpublished_14298	46, XY, inv(7)(p15q36)	cerebral infarct	AC004549	

Patient ID	Karyotype	Phenotype	Breakpoint #1	Breakpoint #2
Unpublished_RP 9834205_9	46, X?, t(7;11)(q33;q?) 46, XY [3]/ 46, XY, der(7)t(1;7)(q11;q31)	splenic lymphoma acute myeloid leukemia	D7S640/ D7S2452 D7S681/ ALDR1	
12461750_110/01	46, XX, del(7)(q21q31)	splenic lymphoma with villous lymphocytes	D7S2317/ D7S522	MEST/ D7S2452
8618441_17_1	47, XX, del(7)(q32q36), t(7;12)(q36;p13), +19	acute myeloid leukemia -M5	D7S681/ ALDR1	AC006357/ D7S68
Unpublished_CM Unpublished_OM Unpublished_00KM881_2	46, X?, del(7)(q32q34) 46, X?, del(7)(q32q34)	splenic lymphoma splenic lymphoma acute myeloid leukemia	D7S2487/ D7S2093E D7S500/ CHRM2	D7S761/ D7S2029 BPGM/ CALD1 sWSS2058/ D7S68
Unpublished_SH 9447826_3	43, XX, 2x add(4)(p12), del(5)(q13q33), del(7)(q22), del(8)(p21), - 12, -13, -17, -22, +mar1, +mar2 [25]	splenic lymphoma refractoy anemia with excess blasts	sWSS4854/ D7S2776 D7S672/ D7S660	D7S761/ D7S1566 D7S631/ D7S676
9447826_4_1	45, XX, -5,del(7)(q32), del(12)(p12)/55, XX, +3, +5, +7, +9, +11, +15, +15, +19, +21	refractoy anemia with excess blasts	D7S687/ D7S677	D7S631/ D7S676
9834205_10	46, XY,t(8;21)(q22;q22)/ 46,XY,del(7)(q32),t(8;21) (q22,q22)	acute myeloid leukemia	D7S650/ D7S2452	D7S2303/ D7S68
9834205_2	43-45, XX, add(7)(q33- 36), inc	myelodysplastic syndrome	D7S650/ D7S2452	
8822909_10 9834205_6_1 9834205_1	4?, X?, del(7)(q22q34) 45, XY, del(7)(q31),-17 46XX,del(7)(q35)/ 47,XX,del(7)(q35),+8	acute myeloid leukemia acute myeloid leukemia myelodysplastic syndrome	D7S500/ CHRM2	D7S500/ sWSS2056 D7S1763/ D7S2452 D7S2447/ D7S68
9834205_8	46, XY, del (7)(q?22), inc/ 46,XY, del(7)(q?22),-7, +mar, inc	acute myeloid leukemia	D7S500/ CHRM2	D7S1815/ D7S1491
Unpublished_99KM1001	45, XX, -14, der(7)t(7;14)(q32;q22) [5]/ 46,XX [5]	acute myeloid leukemia	D7S2536/ APS	
Unpublished_02KM694	46, XY, del(7)(q22), del(20)(q11) [13]	acute myeloid leukemia	RELN/ D7S2494	PRKAR2B/ DRA
Unpublished_96KM707	45, XY, del(5)(q13q33), - 6, del(7)(q?31), add(9)(p13), der(9)t(?;9;? ) , add (19)(p13), -20, +mar	myelodysplastic syndrome	D7S2453/ D7S2783	
Unpublished_02PB347_1	45, XY, -1, del(4)(q?31), -5, del(7)(q?22q?36), -12, -13, -20, -22, +5mar, inc [20]	acute myeloid leukemia	D7S2453/ D7S2305	
Unpublished_97PB471	42, XX, inv(1), -5, del(7)(q?), +8, +8, +der(8), -10, -11, der (12), -19, -20, -21/44, XX, del(7)(q?), +8, +8, +der(8), -10, -11, der(12), -20, - 20/ 47, XX, del(4), del(7)(q?), +8, +8, -10, der(12)/ 43, XX, t(3;19), del(7)(?), -10, -11, der(12), -16, -19, -20, +2mar	acute myeloid leukemia	APS/ D7S2509	

## References

- S1. H. H. Heng, L. C. Tsui, *Chromosoma* **102**, 325-332 (1993).
- S2. J. Kunz *et al.*, *Genomics* **22**, 439 (1994).
- S3. A. K. Hudek, J. Cheung, A. P. Boright, S. W. Scherer, *Bioinformatics*, in press.
- S4. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
- S5. R. Heilig *et al.*, *Nature* **421**, 601-607(2003).
- S6. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
- S7. K. Nakabayashi *et al.*, *Hum. Mol. Genet.* **15**, 1743 (2003).
- S8. K. Hannula *et al.*, *Genomics* **1**, 1 (2001).
- S9. A. Hellman *et al.*, *Cancer Cell.* **1**, 89 (2002).
- S10. D. Mishmar *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 8141 (1998).
- S11. M. Ciullo *et al.*, *Hum. Mol. Genet.* **11**, 2887 (2002).
- S12. C. A. Semple, S. W. Morris, D. J. Porteus, K. L. Evans, *Genome Res.* **12**, 424 (2002).
- S13. L. R. Osborne *et al.*, *Genomics* **36**, 328 (1996).
- S14. G. Glockner *et al.*, *Genome Res.* **8**, 1060 (1998).
- S15. M. D. Wilson *et al.*, *Nucl. Acids Res.* **6**, 1352 (2001).
- S16. L. R. Osborne *et al.*, *Genomics* **45**, 402 (1997).
- S17. L. R. Osborne *et al.*, *Nature Genet.* **29**, 321 (2001).
- S18. B. Ma, J. Tromp, M. Li, *Bioinformatics* **18**, 440 (2002).
- S19. A. K. Ewart *et al.*, *Nature Genet.* **5**, 11 (1993).
- S20. L. A. Perez-Jurado *et al.*, *Hum. Mol. Genetics* **7**, 325 (1998).