

# The DNA sequence of human chromosome 21

## The chromosome 21 mapping and sequencing consortium

M. Hattori\*, A. Fujiyama\*, T. D. Taylor\*, H. Watanabe\*, T. Yada\*, H.-S. Park\*, A. Toyoda\*, K. Ishii\*, Y. Totoki\*, D.-K. Choi\*, E. Soeda†, M. Ohki‡, T. Takagi§, Y. Sakaki\*§; S. Taudien||, K. Blechschmidt||, A. Polley||, U. Menzel||, J. Delabar¶, K. Kumpf||, R. Lehmann||, D. Patterson#, K. Reichwald||, A. Rump||, M. Schillhabel||, A. Schudy||, W. Zimmermann||, A. Rosenthal||; J. Kudoh\*, K. Shibuya\*, K. Kawasaki\*, S. Asakawa\*, A. Shintani\*, T. Sasaki\*, K. Nagamine\*, S. Mitsuyama\*, S. E. Antonarakis\*\*, S. Minoshima\*, N. Shimizu\*, G. Nordsiek††, K. Hornischer††, P. Brandt††, M. Scharfe††, O. Schön††, A. Desario‡‡, J. Reichelt††, G. Kauer††, H. Blöcker††; J. Ramser§§, A. Beck§§, S. Klages§§, S. Hennig§§, L. Riesselmann§§, E. Dagand§§, T. Haaf§§, S. Wehrmeyer§§, K. Borzym§§, K. Gardiner#, D. Nizetic|||, F. Francis§§, H. Lehrach§§, R. Reinhardt§§ & M.-L. Yaspo§§

### Consortium Institutions:

\* RIKEN, Genomic Sciences Center, Sagamihara 228-8555, Japan

|| Institut für Molekulare Biotechnologie, Genomanalyse, D-07745 Jena, Germany

\* Department of Molecular Biology, Keio University School of Medicine, Tokyo 160-8582, Japan

†† GBF (German Research Centre for Biotechnology), Genome Analysis, D-38124 Braunschweig, Germany

§§ Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin-Dahlem, Germany

### Collaborating Institutions:

† RIKEN, Life Science Tsukuba Research Center, Tsukuba 305-0074, Japan

‡ Cancer Genomics Division, National Cancer Center Research Institute, Tokyo 104-0045, Japan

§ Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan

¶ UMR 8602 CNRS, UFR Necker Enfants-Malades, Paris 75730, France

# Eleanor Roosevelt Institute, Denver, Colorado 80206, USA

\*\* Medical Genetics Division, University of Geneva Medical School, Geneva 1211, Switzerland

‡‡ CNRS UPR 1142, Institut de Biologie, Montpellier, 34060, France

||| School of Pharmacy, University of London, London WC1N 1AX, UK

**Chromosome 21 is the smallest human autosome. An extra copy of chromosome 21 causes Down syndrome, the most frequent genetic cause of significant mental retardation, which affects up to 1 in 700 live births. Several anonymous loci for monogenic disorders and predispositions for common complex disorders have also been mapped to this chromosome, and loss of heterozygosity has been observed in regions associated with solid tumours. Here we report the sequence and gene catalogue of the long arm of chromosome 21. We have sequenced 33,546,361 base pairs (bp) of DNA with very high accuracy, the largest contig being 25,491,867 bp. Only three small clone gaps and seven sequencing gaps remain, comprising about 100 kilobases. Thus, we achieved 99.7% coverage of 21q. We also sequenced 281,116 bp from the short arm. The structural features identified include duplications that are probably involved in chromosomal abnormalities and repeat structures in the telomeric and pericentromeric regions. Analysis of the chromosome revealed 127 known genes, 98 predicted genes and 59 pseudogenes.**

Chromosome 21 represents around 1–1.5% of the human genome. Since the discovery in 1959 that Down syndrome occurs when there are three copies of chromosome 21 (ref. 1), about twenty disease loci have been mapped to its long arm, and the chromosome's structure and gene content have been intensively studied. Consequently, chromosome 21 was the first autosome for which a dense linkage map<sup>2</sup>, yeast artificial chromosome (YAC) physical maps<sup>3–6</sup> and a *NotI* restriction map<sup>7</sup> were developed. The size of the long arm of the chromosome (21q) was estimated to be around 38 megabases (Mb), based on pulsed-field gel electrophoresis (PFGE) studies using *NotI* restriction fragments<sup>7</sup>. By 1995, when the sequencing effort was initiated, around 60 messenger RNAs specific to chromosome 21 had been characterized. Here we report and discuss the sequence and gene catalogue of the long arm of chromosome 21.

## Chromosome geography

**Mapping.** We converted the euchromatic part of chromosome 21 into a minimum tiling path of 518 large-insert bacterial clones. This collection comprises 192 bacterial artificial chromosomes (BACs), 111 P1 artificial chromosomes (PACs), 101 P1, 81 cosmids, 33 fosmids and 5 polymerase chain reaction (PCR) products (Fig. 1). We used clones originating from four whole-genome libraries and nine chromosome-21-specific libraries. The latter were particularly useful for mapping the centromeric and telomeric repeat-contain-

ing regions and sequences showing homology with other human chromosomes.

We used two strategies to construct the sequence-ready map of chromosome 21. In the first, we isolated clones from arrayed genomic libraries by large-scale non-isotopic hybridization<sup>8</sup>. We built primary contigs from hybridization data assembled by simulated annealing, and refined clone overlaps by restriction digest fingerprinting. Contigs were anchored onto PFGE maps of *NotI* restriction fragments and ordered using known sequence tag site (STS) framework markers. We used metaphase fluorescent *in situ* hybridization (FISH) to check the locations of more than 250 clones. The integrity of the contigs was confirmed by FISH, and gaps were sized by a combination of fibre FISH and interphase nuclei mapping. Gaps were filled by multipoint clone walking. In the second strategy, we isolated seed clones using selected STS markers and then either end-sequenced or partially sequenced them at fivefold redundancy. Seed clones were extended in both directions with new genomic clones, which were identified either by PCR using amplimers derived from parental clone ends or by sequence searches of the BAC end sequence database (<http://www.tigr.org>). Nascent contigs were confirmed by sequence comparison.

The final map is shown in Fig. 1. It comprises 518 bacterial clones forming four large contigs. Three small clone gaps remain despite screening of all available libraries. The estimated sizes of

these gaps are 40, 30 and 30 kilobases (kb), respectively, as indicated by fibre FISH (see supporting data set, last section (<http://chr21.r2-berlin.mpg.de>)).

**Sequencing.** We used two sequencing strategies. In the first, large-insert clones were shotgun cloned into M13 or plasmid vectors. DNA of subclones was prepared or amplified, and then sequenced using dye terminator and dye primer chemistry. On average, clones were sequenced at 8–10-fold redundancy. In the second approach, we sequenced large-insert clones using a nested deletion method<sup>9</sup>. The redundancy of the nested deletion method was about fourfold. Gaps were closed by a combination of nested deletions, long reads, reverse reads, sequence walks on shotgun clones and large insert clones using custom primers. Some gaps were also closed by sequencing PCR products.

The total length of the sequenced parts of the long arm of chromosome 21 is 33,546,361 bp. The sequence extends from a 25-kb stretch of  $\alpha$ -satellite repeats near the centromere to the telomeric repeat array. Seven sequencing gaps remain, totalling less than 3 kb. The largest contig spans 25.5 Mb on 21q. The total length of 21q, including the three clone gaps, is about 33.65 Mb. Thus, we achieved 99.7% coverage of the chromosome. We also sequenced a small contig of 281,116 bp on the p arm of chromosome 21.

We estimated the accuracy of the final sequence by comparing 18 overlapping sequence portions spanning 1.2 Mb. We estimate from this external checking exercise that the accuracy of the entire sequence exceeds 99.995%.

**Sequence variations.** Twenty-two overlapping sequence portions comprising 1.36 Mb and spread over the entire chromosome were compared for sequence variations and small deletions or insertions. We detected 1,415 nucleotide variations and 310 small deletions or insertions and confirmed them by inspecting trace files. There was an average of one sequence difference for each 787 bp, but the observed sequence variations were not evenly distributed along 21q. In the telomeric portion (21q22.3–qter) the average was one difference for each 500 bp. The highest sequence variation (one difference in 400 bp) was found in a 98-kb segment from this region. In the proximal portion (21q11–q22.3) we found on average one

difference per 1,000 bp; the lowest level was 1 in 3,600 bp in a 61-kb segment of 21q22.1.

**Interspersed repeats.** Table 1 summarizes the repeat content of chromosome 21. Chromosome 21 contains 9.48% Alu sequences and 12.93% LINE1 elements, in contrast with chromosome 22 which contains 16.8% Alu and 9.73% LINE1 sequences<sup>10</sup>.

### Gene catalogue

The gene catalogue of chromosome 21 contains known genes, novel putative genes predicted *in silico* from genomic sequence analysis

**Figure 1** The sequence map of human chromosome 21. Sequence positions are indicated in Mb. Annotated features are shown by coloured boxes and lines. The chromosome is oriented with the short p-arm to the left and the long q-arm to the right. Vertical grey box, centromere. The three small clone gaps are indicated by narrow grey vertical boxes (in proportion to estimated size) on the right of the q-arm. The cytogenetic map was drawn by simple linear stretching of the ISCN 850-band, Giemsa-stained ideogram to match the length of the sequence: the boundaries are only indicative and are not supported by experimental evidence. In the mapping phase, information on STS markers was collected from publicly available resources. The progress of mapping and sequencing was monitored using a sequence data repository in which sequences of each clone were aligned according to their map positions. A unified map of these markers was automatically generated (<http://hgp.gsc.riken.go.jp/marker/>) and enabled us to carry out simultaneous sequencing and library screening among centres. Vertical lines: markers, according to sequence position, from GDB (black; <http://www.gdb.org/>), the GB4 radiation hybrid map (blue; Whitehead Institute, Massachusetts Institute of Technology)<sup>43</sup>, the G3 radiation hybrid map (dark green; Stanford Human Genome Centre, California)<sup>44</sup> and two linkage maps (red; Genethon; CHLC)<sup>45,46</sup>. Only marker distribution is presented here: additional details, such as marker names and positions, can be found on our web sites. The *NotI* physical map of chromosome 21 was also used<sup>7</sup> (*NotI* sites, light green). Genes are indicated as boxes or lines according to strand along the upper scale in three categories: known genes (category 1, red), predicted genes (categories 2 and 3, light green; category 4, light blue) and pseudogenes (category 5, violet). For genes of categories 1, 2, 3 and 5, the approved symbols from the HUGO nomenclature committee are used. CpG islands are olive (they were identified when they exceeded 400 bp in length, contained more than 55% GC, showed an observed over expected CpG frequency of >0.6 and had no match to repetitive sequences). The G+C content is shown as a graph in the middle of the Figure. It was calculated on the basis of the number of G and C nucleotides in a 100-kb sliding window in 1-kb steps across the sequence. The clone contig consists of all clones that were sequenced to 'finished' quality from all five centres in the consortium. Clones are indicated as coloured boxes by centre: red, RIKEN; dark blue, IMB; light blue, Keio; yellow, GBF; and green, MPIMG. Clones that were only partially sequenced have grey boxes on either end to show the actual or estimated clone end position. Four whole-genome libraries (RPC1-11 BAC, Keio BAC, Caltech BAC and RPC1, 3-5 PAC) and nine chromosome-specific libraries (CMB21-BAC, Roizes-BAC, CMP21-P1, CMC21-cosmid, LLNCO21, KU21D, ICRFc102 and ICRFc103 cosmid, and CMF21-fosmid) were used to isolate clones (see <http://hgp.gsc.riken.go.jp> or <http://chr21.rz-berlin.mpg.de> for library information). Breakpoints from chromosomal rearrangements are shown as coloured boxes according to their classification: natural (green), spontaneously occurring in cell lines (yellow), radiation induced (purple) and combinations of the above (black). Blue boxes, intra-chromosomal duplications; green boxes, inter-chromosomal duplications (see text). Alu (red) and LINE1 (blue) interspersed repeat element densities are shown in the bottom graph as the percentage of the sequence using the same method of calculation as for G+C content. The final non-redundant sequence was divided into 340-kb segments (grey boxes), with 1-kb overlaps (to avoid splitting of most exons in both segments), and has been registered, along with biological annotations, in the DDBJ/EMBL/GenBank databases under accession numbers AP001656–AP001761 (DDBJ) and AL163201–AL163306 (EMBL). Segments for the three clone gaps (accession numbers AP001742/AL163287, AP001744/AL163289 and AP001750/AL163295) have also been deposited in the databases with a number of Ns corresponding to the estimated gap lengths. The sequences and additional information can be found from the home pages of the participating centres of the chromosome 21 sequencing consortium (RIKEN, <http://hgp.gsc.riken.go.jp/>; IMB, <http://genome.imb-jena.de/>; Keio, <http://dmb.med.keio.ac.jp/>; GBF, <http://www.genome.gbf.de/>; MPI, <http://chr21.rz-berlin.mpg.de/>).

**Table 1 The content of interspersed repeats in human chromosome 21**

Repeat type	Total number of elements	Coverage (bp)	Coverage (%)
SINEs	15,748	3,667,752	10.84%
ALUs	12,341	3,208,437	9.48%
MIRs	3,407	459,315	1.36%
LINEs	12,723	5,245,516	15.51%
LINE1	8,982	4,372,851	12.93%
LINE2	3,741	872,665	2.58%
LTR elements	9,598	3,116,881	9.21%
MaLRs	5,379	1,646,297	4.87%
Retroviral	2,115	760,119	2.25%
MER4 group	1,396	479,451	1.42%
Other LTR	708	231,014	0.68%
DNA elements	3,950	812,031	2.40%
MER1 type	2,553	460,769	1.36%
MER2 type	851	257,653	0.76%
Mariners	168	26,235	0.08%
Other DNA elements	378	67,374	0.20%
Unclassified	64	15,234	0.05%
Total interspersed repeats	42,083	12,857,414	38.01%
Simple repeats	5,987	427,755	1.26%
Low complexity	5,868	249,449	0.74%
Total	54,045	13,551,271	40.06%
Total sequence length	33,827,477		
G+C%	40.89%		

and pseudogenes. The catalogue was arbitrarily divided into five main hierarchical categories (see below) to distinguish known genes from pure gene predictions, and also anonymous complementary DNA sequences from those exhibiting similarities to known proteins or modular domains.

The criteria governing the gene classification were based on the results of the integrated results of computational analysis using exon prediction programs and sequence similarity searches. We applied the following parameters: (1) Putative coding exons were predicted using GRAIL, GENSCAN and MZEF programs. Consistent exons were defined as those that were predicted by at least two programs. (2) Nucleotide sequence identities to expressed sequence tags (ESTs) (as identified by using BlastN with default parameters) were considered as a hallmark for gene prediction only if these ESTs were spliced into two or more exons in genomic DNA, and showed greater than 95% identity over the matched region. These criteria are conservative and were chosen to discard spurious matches arising from either cDNAs primed from intronic sites or repetitive elements frequently found in 5' or 3' untranslated regions. (3) Amino-acid similarities to known proteins or modular functional domains were considered to be significant when an overall identity of greater than 25% over more than 50 amino-acid residues was observed (as detected using BlastX with Blossum 62 matrix against the non-redundant database).

**Gene categories.** The results of sequence analysis were visually inspected to locate known genes, to identify new genes and to unravel novel putative transcription units after assembling consistent predicted exons into so-called *in silico* gene models. These gene predictions were also evaluated by incorporating information provided by EST and protein matches. Each gene was assigned to one of the following sub-categories:

**Category 1:** Known human genes (from the literature or public databases). *Subcategory 1.1:* Genes with 100% identity over a complete cDNA with defined functional association (for example, transcription factor, kinase). *Subcategory 1.2:* Genes with 100% identity over a complete cDNA corresponding to a gene of unknown function (for example, some of the KIAA series of large cDNAs).

**Category 2:** Novel genes with similarities over essentially their total length to a cDNA or open reading frame (ORF) of any organism. *Subcategory 2.1:* Genes showing similarity or homology to a characterized cDNA from any organism (25–100% amino-acid identity). This class defines new members of human gene families, as well as new human homologues or orthologues of genes from yeast, *Caenorhabditis elegans*, *Drosophila*, mouse and so on. *Subcategory 2.2:* Genes with similarity to a putative ORF predicted *in silico* from the genomic sequence of any organism but which currently lacks experimental verification.

**Category 3:** Novel genes with regional similarities to confined protein regions. *Subcategory 3.1:* Genes with amino-acid similarity confined to a protein region specifying a functional domain (for example, zinc fingers, immunoglobulin domains). *Subcategory 3.2:* Genes with amino-acid similarity confined to regions of a known protein without known functional association.

**Category 4:** Novel anonymous genes defined solely by gene prediction. These are putative genes lacking any detectable similarity to known proteins or protein motifs. These models are based solely on spliced EST matches, consistent exon prediction or both.

*Subcategory 4.1:* Predicted genes composed of a pattern of two or more consistent exons (located within <20 kb) and supported by spliced EST match(es). *Subcategory 4.2:* Predicted genes corresponding to spliced EST(s) but which failed to be recognized by exon prediction programs. *Subcategory 4.3:* Predicted genes composed only of a pattern of consistent exons without any matches to EST(s) or cDNA. Intuitively, predicted genes from subcategory 4.1 are considered to have stronger coding potential than those of subcategory 4.3.

**Category 5:** Pseudogenes may be regarded as gene-derived DNA sequences that are no longer capable of being expressed as protein products. They were defined as predicted polypeptides with strong similarity to a known gene, but showing at least one of the following features: lack of introns when the source gene is known to have an intron/exon structure, occurrence of in-frame stop codons, insertions and/or deletions that disrupt the ORF or truncated matches. Generally, this was an unambiguous classification.

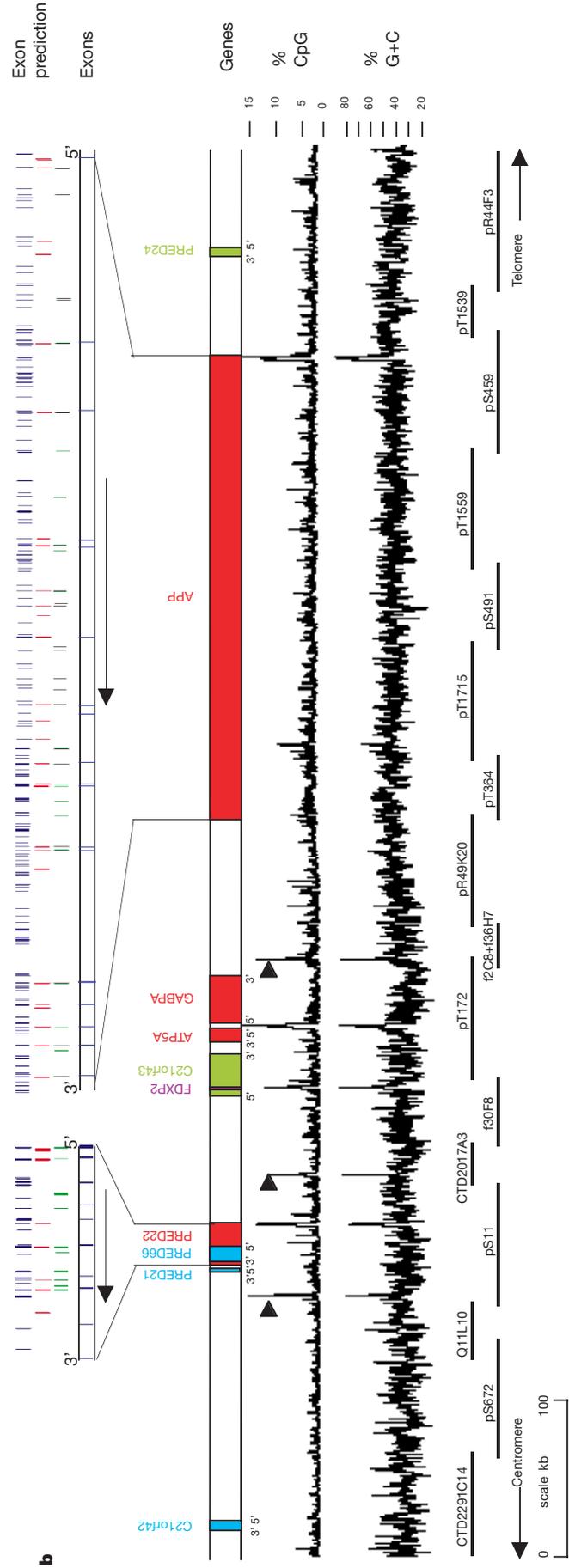
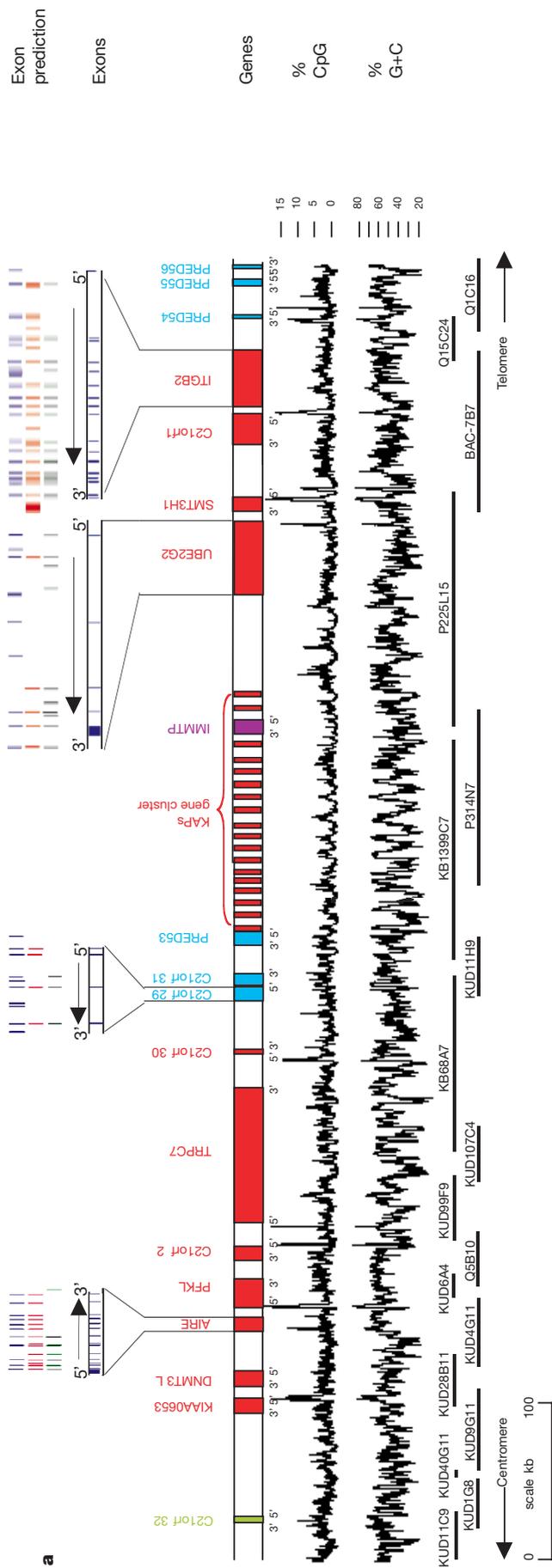
When a gene could fulfil more than one of these criteria, it was placed into the higher possible category (for example, gene prediction with spliced EST exhibiting a significant match to a known protein was placed in subcategory 2.2 rather than 4.2).

**The gene content of chromosome 21.** For the gene catalogue of chromosome 21, see Table 2. The chromosome contains 225 genes and 59 pseudogenes. Of these, 127 correspond to known genes (subcategories 1.1 and 1.2) and 98 represent putative novel genes predicted *in silico* (categories 2, 3 and 4). Of the novel genes, 13 are similar to known proteins (subcategories 2.1 and 2.2), 17 are anonymous ORFs featuring modular domains (subcategories 3.1 and 3.2), and most (68 genes) are anonymous transcription units with no similarity to known proteins (subcategories 4.1, 4.2 and 4.3). Our data show that about 41% of the genes that were identified on chromosome 21 have no functional attributes.

In a rough generic description, the gene catalogue of chromosome 21 contains at least 10 kinases (PRED1, PRSS7, C21orf7, PRED33, PRKCBP2, DYRKA1, ANKDR3, SNF1LK, PDXK and PFKL), five genes involved in ubiquitination pathways (USP25, USP16, UBASH, UBE2G2 and SMT3H1), five cell adhesion molecules (NCAM2, IGSF5, C21orf43, DSCAM and ITGB2), a number of transcription factors and seven ion channels (C21orf34, KCNE2, KCNE1, CILC1L, KCNJ6, KCNJ15 and TRPC7). Several clusters of functionally related genes are arranged in tandem arrays on 21q, indicating the likelihood of ancient sequential rounds of gene duplication. These clusters include the five members of the interferon receptor family that spans 250 kb on 21q (positions 20,179,027–20,428,899), the trefoil peptide cluster (TFF1, TFF2 and TFF3) spanning 54 kb on 21q22.3 (positions 29,279,519–29,333,970) and the keratin-associated protein (KAP) cluster spanning 164 kb on 21q22.3 (positions 31,468,577–31,632,094) (Table 2). The last contains 18 units of this highly repetitive gene family featuring genes and different pseudogene fragments and revealing inverted duplications within the gene cluster (described below). Finally, the p arm of chromosome 21 contains at least one gene (TPTE) encoding a putative tyrosine phosphatase. This is the first description of a protein-coding gene mapping to the p arm of an acrocentric chromosome. However, the functional activity of this gene remains to be demonstrated.

Chromosome 21 contains a very low number of identified genes (225) compared with the 545 genes reported for chromosome 22 (ref. 10). Figure 1 shows the overall distribution of the 225 genes and 59 pseudogenes on chromosome 21 in relation to compositional features such as G+C content, CpG islands, Alu and L1 repeats and the positions of selected STSs, polymorphic markers and chromosomal breakpoints. Earlier reports indicated that gene-rich regions are Alu rich and LINE1 poor, whereas gene-poor regions contain more LINE1 elements at the expense of Alu sequences<sup>11</sup>. Our data, and the comparison with chromosome 22, support these findings (see Tables 1 and 2, Fig. 1 and ref. 10). There is a large 7-Mb region

◀ **Table 2** Gene catalogue of chromosome 21. The table displays the gene symbol, accession number, gene description, gene category, orientation, gene start position, gene end position, genomic size and corresponding genomic clone name. The gene categories are colour coded as follows: known genes (category 1) in red, novel genes with similarities to characterized cDNAs from any organism and novel genes with similarities to protein domains (categories 2 and 3) in green, novel gene prediction (category 4) in blue, and pseudogenes (category 5) in purple. Coordinates are given in base pairs.



(between 5 and 12 Mb on Fig. 1) with low G+C content (35% compared with 43% for the rest of the chromosome) that correlates with a paucity of both Alu sequences and genes. Only two known genes (PRSS7 and NCAM2) and five predicted genes can be found in this region. Further reinforcing the concept that compositional features correlate with gene density, Fig. 2 compares the genomic organization and gene density in a 831-kb G+C-rich DNA region (53%; Fig. 2a) with that of a 915-kb DNA stretch representative of a G+C-poor region (39.5%; Fig. 2b). Figure 2a shows eleven known genes, seven predicted genes, one pseudogene and the KAP cluster. Figure 2b shows four known genes, five predicted genes and one pseudogene. Figure 2 also displays examples of exon/intron structures as defined by the exon prediction programs in parallel with the real gene structure that was obtained by sequence alignment using the cognate mRNA. Most exons were predicted by the combination of the three programs. However, MZEF tends to overpredict exons compared with GRAIL and GENSCAN, in particular for the large APP gene. In addition, CpG islands correlate well as indicators of the 5' end of genes in both of these regions.

**Structural features of known and predicted genes.** Among the 127 known genes, 22 genes are larger than 100 kb, the largest being DSCAM (840 kb). Seven of the largest known genes cover 1.95 Mb and lie within a region of 4.5 Mb (positions 23.7 Mb–28.2 Mb) that contains only four predicted genes and two pseudogenes. The average size of the genes is 39 kb, but there is a bias in favour of the category 1 genes. Known genes have a mean size of 57 kb, whereas predicted genes (categories 2, 3 and 4) have a mean size of 27 kb. This is not unexpected, because of the inherent difficulties in extending exon prediction to full-length gene identification. For instance, exon prediction and EST findings are usually not exhaustive. This would also explain the fact that 69% of the predicted genes have no similarity to known proteins.

Despite the shortcomings of current gene prediction methods, all known genes previously shown to map on chromosome 21 (ref. 12) were identified independently by *in silico* methods. Patterns of consistent exon prediction alone were sufficient to locate at least partial gene structures for more than 95% of these. This was true even for large A+T-rich genes, such as NCAM2, APP (Fig. 2b) and GRIK1. These three genes are several hundred kilobases long with a G+C content of 38–40%, but most exons were well predicted and enough introns were sufficiently small that a clear pattern of consistent exons was seen. In addition, more than 95% of the known genes were independently identified from spliced ESTs. Characteristics of genes that could be missed using our detection methods include those with poor exon prediction and long 3' untranslated regions (>2 kb); those with poor exon prediction and very restricted expression pattern; and those with very large introns (>30 kb).

We designed our gene identification criteria to extract most of the coding potential of the chromosome and to minimize false positive predictions. Errors to be expected in the predictions include false positive exons, incorrect splice sites, false negative exons, fusion of multiple genes into one transcription unit and separation of a single gene into two or more transcription units. We believe that our method is sufficiently robust to pinpoint real genes, but our models still require experimental validation. In a pilot experiment on 14 predicted category 4 genes we performed RT-PCR (PCR with reverse transcription) in 12 tissues. We could confirm 11 genes and connect two gene predictions into a single transcription unit.

Pseudogenes are often overlooked in a gene catalogue aimed at specifying functional proteins, but they may be important in influencing recombination events. The 59 pseudogenes described here are not randomly located in the chromosome (Fig. 1). Twenty-four pseudogenes are distributed in the first 12 Mb of 21q, which is a gene-poor region. In contrast, a cluster of 11 pseudogenes was found within a 1-Mb stretch of DNA that is gene rich and corresponds precisely to the highest density of Alu sequences on the chromosome (positions 22,421,026–23,434,597).

**Base composition and gene density.** It is tempting to speculate on possible correlations between the base composition, gene density and molecular architecture of the chromosome bands. Giemsa-dark chromosomal bands are comprised of L isochores (<43% G+C), whereas Giemsa-light bands have variable composition. The latter include L, H1/H2 (43–48% G+C) and H3 isochores (>48% G+C)<sup>15</sup>. In humans, the average gene density is around one gene per 150 kb in L, one per 54 kb in H1/H2 and one per 9 kb in H3 isochores<sup>14</sup>. The proximal half of 21q (from 0.2 to 17.7 Mb of Fig. 1), which corresponds mainly to the large Giemsa dark band, 21q21, comprises a long continuous L isochore, harbouring extensive stretches of 34–37% G+C, and rare segments of more than 40% G+C. Twenty-five category 1 genes and 33 category 2–4 genes were found in this region, giving an average density of one gene per 301 kb.

The distal half of 21q (17.7–33.5 Mb) largely comprises stretches of H1/H2 isochores alternating with L isochores, and H3 isochores localized within the region spanning positions 29–33.5 Mb. The overall gene density in the telomeric half is much higher than that in the proximal half: 101 genes of category 1 and 66 genes of categories 2–4 were found in this region, giving an average of about one gene per 95 kb. The DSCAM gene, found within an L isochore in this region, spans 834 kb. In contrast, the region spanning the H3 isochores contains 46 category 1 genes and 31 category 2–4 genes, averaging one gene per 58 kb.

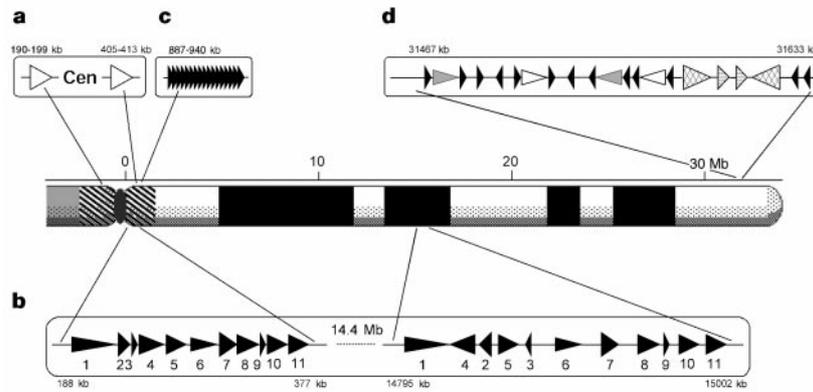
The L isochores have lower gene density than that predicted from whole-genome analysis: one gene per 301 kb compared with one per 150 kb. The H3 isochores are also lower in gene content, averaging one gene per 58 kb compared with one gene per 9 kb estimated for the genome as a whole. This discrepancy may be due to an overestimation of the total number of human genes based on EST data (see below). Alternatively, we may have missed half of the genes on this chromosome. This second possibility is unlikely as more than 95% of the known genes have been predicted using our criteria.

### Chromosomal structural features

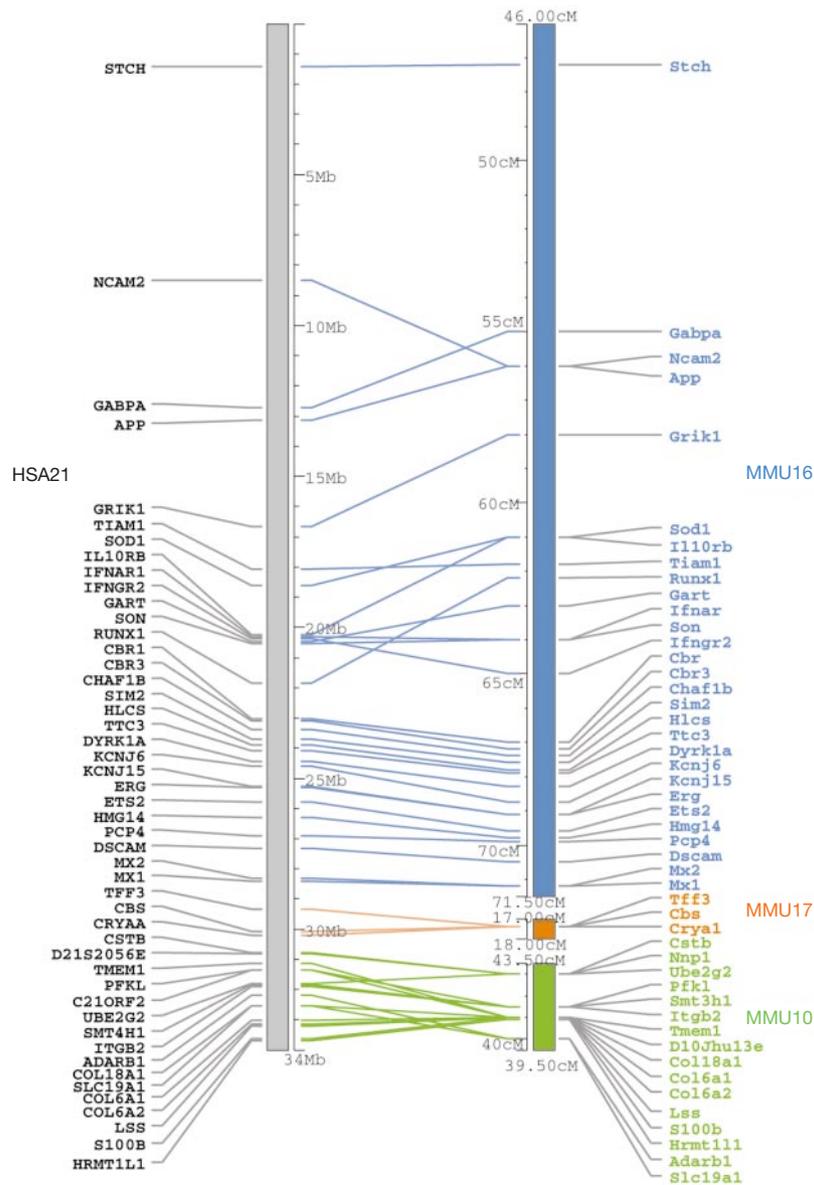
**Duplications within chromosome 21.** The unmasked sequence of the whole chromosome was compared with itself to detect intra-chromosomal duplications. We identified a 10-kb duplication in the pericentromeric regions of the p- and q-arms (Fig. 3a). The p-arm copy extends from 190 to 199 kb of the p-arm contig, and the q-arm copy extends from 405 to 413 kb of the 21q sequence. We identified a CpG island on the centromeric side of the duplication in the p-arm, indicating that there may be an active gene in the vicinity of the duplicated regions. A similar structure was reported for chromosome 10 (ref. 15), so such repeats close to the centromere may have a functional role. The pericentromeric region in the q-arm also contains several duplications, including several clusters of  $\alpha$ -satellite sequences and even telomeric satellites

Another duplication corresponding to a large 200-kb region has been identified in proximal and distal locations on 21q (Fig. 3b). This duplication was previously reported<sup>16</sup> but was not analysed in detail at the sequence level. The proximal copy is located from 188 to 377 kb in 21q11.2, whereas the distal copy lies in 21q22 and extends from 14,795 to 15,002 kb. The two copies are highly conserved and show 96% identity. We detected two large inversions, several other rearrangements and several translocations or duplications within the duplicated units (Fig. 3b), which caused segmentation of the

**Figure 2** Gene organization on chromosome 21. **a**, A G+C-rich region of the telomeric part; **b**, an AT-rich region of the centromeric part. Genes are represented by coloured boxes. Category 1, red; categories 2 and 3, green; category 4, blue; category 5, violet. Predicted exons shown in the enlarged gene areas are represented as: MZEF, blue; GenScan, red; Grail, green. Arrowheads, orphan CpG islands that may indicate the presence of a cryptic gene.



**Figure 3** Schematic view of the duplicated regions in chromosome 21 as described in the text. **a–d**, Duplicated regions. The positions of each repeat structure are shown in kb starting at the centromere. The arrowheads represent the orientation and approximate size of each repetitive unit.



**Figure 4** Schematic view of the syntenic regions between human chromosome 21 (HSA21) and mouse chromosomes 16 (MMU16), 17 (MMU17) and 10 (MMU10). Left: sequence map of human chromosome 21. Right: corresponding mouse chromosomes. Each pair of syntenic markers is joined with a line.

units into at least 11 pieces. The distal copy is 207 kb long and the proximal copy is 189 kb; the 18-kb size difference between the two duplicated segments is due to insertions in the distal copy, deletions in the proximal copy or both.

In the region on 21q between 887 and 940 kb a block of sequence is repeated 17 times (Fig. 3c). The similarity of these repetitive units indicates that they were formed by a recent triplication event of a region of six repeat unit blocks, which had in turn been generated by duplication of a three-block unit.

Another repeat sequence lies between the TRPC7 and UBE2G2 genes on 21q22.3 (31,467–31,633 kb). This feature corresponds to the 166-kb KAP gene and pseudogene cluster described above (Fig. 2a). A 0.5–1-kb segment is repeated at least 13 times, with 5–10-kb spacer intervals (Fig. 3d). The repeat units share more than 91% identity with each other.

**Comparison of chromosome 21 with chromosome 22.** The two chromosomes are similar in size, and both are acrocentric. The gene density, however, is much higher on chromosome 22 (ref. 10). We detected sequence similarity in the pericentromeric and sub-telomeric regions of both chromosomes. For example, two different regions in the 21p contig (42–84 kb; 239–263 kb) are duplicated in 22q (1043–1067 kb; 1539–1564 kb). These duplications are located within the pericentromeric regions of both chromosomes<sup>17</sup>. Half of the first region is further duplicated at the position 22,223–22,248 kb in chromosome 22. In addition, two inverted duplications in 21q at 88–156 kb and 646–751 kb have also been observed on 22q at positions 572–637 kb and 45–230 kb. Large clusters of  $\alpha$ -satellite sequences (10 kb for chromosome 21 and 119 kb for chromosome 22) are located on 21q (88–156 kb) and 22q (572–637 kb).

The most telomeric clone, F50F5, isolated from the chromosome-specific CMF21 fosmid library, contains a telomeric repeat array that represents the hallmark of the telomeric end of a chromosome. This array was missing in the chromosome 22q sequence<sup>10</sup>. However, the 22q sequence ends very near to the telomere, considering that it shows strong homology with a 2.5–10-kb stretch of telomeric sequence present in F50F5.

**Comparison of chromosome 21 with other autosomes.** In the most telomeric region of chromosome 21 we also identified a novel repeat structure featuring a non-identical 93-bp unit that is repeated 10 times. This block of 93-bp repeats is located 7.5 kb from the start point of the telomeric array. Similar 93-bp repeat sequences were also detected by BLAST analysis in chromosomes 22, 10 and 19. FISH analysis data suggest that this 93-bp repeat unit is also located on 5qter, 7pter, 17qter, 19pter, 19qter, 20pter, 21qter and 22qter, as well as on other chromosomal ends. Thus, this 93-bp repeat may be a common structural feature shared by many human telomeres.

We have found some paralogous regions between chromosome 21 and other human chromosomes, which were also pointed out by metaphase FISH analysis of the corresponding genomic clones. For example, a 100-kb region of clone B15L0C0 located on 21p is shared with chromosomes 4, 7, 20 and 22. A second homologous region of 50 kb on 21q between 15,530 and 15,580 kb is shared with a segment on chromosome 16 between the genes 44M2.1 and 44M2.2. More details on these regions can be found at <http://hgp.gsc.riken.go.jp/>.

**Synteny with mouse.** Human chromosome 21 shows conserved syntenies to mouse chromosomes 16, 17 and 10 (<http://www.informatics.jax.org/>). Figure 4 shows a comparative map of human chromosome-21-specific genes with their mouse orthologues. A number of inversions can be seen. These changes in gene order may be due to rearrangements during genome evolution. Alternatively, they may reflect the fact that the mouse gene map is still inaccurate because it is based on linkage and physical mapping.

**Breakpoints.** Figure 1 shows the locations of 39 breakpoints on the physical map. Here we describe several classes of breakpoint, all of which either occurred naturally in the human population before hybrid construction or were induced by irradiation. The natural

breakpoints arose mainly from reciprocal translocations of chromosome 21 with other human chromosomes (6;21, 4;21, 3;21, 1;21, 8;21, 10;21, 11;21 and 21;22). A second class of naturally occurring breakpoints derived from intrachromosomal rearrangements of chromosome 21 (ACEM, 6918, MRC2, R210 and DEL21). A third class of breakpoints, designated 3x1, 3x2, 1x4D, 1x4F and 1x18, were generated experimentally by irradiation of hybrids containing intact chromosome 21q arms<sup>18</sup>. Hybrids 2Fur, 750 and 511 represent rearrangements of chromosome 21 that occurred spontaneously in somatic cell hybrids. All of these chromosome derivatives were isolated in Chinese hamster ovary (CHO)  $\times$  human somatic cell hybrids.

Fine mapping revealed an uneven distribution of breakpoints that fell roughly in two clusters on chromosome 21. Nine breakpoints occur within the pericentromeric region (0–2.2 Mb) and another nine are located within a 2.4-Mb region in 21q22 (20.1–22.5 Mb) (Fig. 1). In contrast, large regions are totally devoid of breakpoints. For instance, only two translocation breakpoints are located in the 10-Mb region between 4.95 and 14.4 Mb of the q arm.

Several breakpoints occur within or near the duplicated regions described above. For instance, three breakpoints (1x4D, 1x18 and 2Fur) occur between positions 100 and 400 kb on 21q. This region corresponds to the proximal copy of the large duplicated region described in Fig. 3b. Another breakpoint (ACEM) occurs between positions 14,400 and 14,525 kb, close to the distal copy of this duplicated region. We also found a naturally occurring 21;22 translocation breakpoint (position 31,350–31,380 kb) in the KAP cluster.

Duplicated regions may mediate certain mechanisms involved in chromosomal rearrangement. It is likely that similar sequence features may be important for duplication, genetic recombination and chromosomal rearrangement. Further sequence analysis will help to unravel the underlying molecular mechanisms of chromosome breakage and recombination.

**Recombination.** The distribution of the recombination frequency on chromosome 21 is different in males and females<sup>12</sup>. In Fig. 5 genetic distances of known polymorphic markers from male, female and sex-average maps are compared with the distances in nucleotides on 21q. The recombination frequency is relatively higher near the centromere in females and near the telomere in males. This confirms earlier analysis based on physical maps<sup>11</sup>. Unlike chromosome 22, chromosome 21 does not appear to contain particular regions with a steep increase in recombination frequency in the middle of the chromosome.

## Medical implications

**Down syndrome.** Besides the constant feature of mental retardation, individuals with Down syndrome also frequently exhibit congenital heart disease, developmental abnormalities, dysmorphic features, early-onset Alzheimer's disease, increased risk for specific leukaemias, immunological deficiencies and other health problems<sup>19</sup>. Ultimately, all these phenotypes are the result of the presence of three copies of genes on chromosome 21 instead of two. Data from transgenic mice indicate that only a subset of the genes on chromosome 21 may be involved in the phenotypes of Down syndrome<sup>20</sup>. Although it is difficult to select candidate genes for these phenotypes, some gene products may be more sensitive to gene dosage imbalance than others. These may include morphogens, cell adhesion molecules, components of multi-subunit proteins, ligands and their receptors, transcription regulators and transporters. The gene catalogue now allows the hypothesis-driven selection of different sets of candidates, which can then be used to study the molecular pathophysiology of the gene dosage effects. The complete catalogue will also provide the opportunity to search systematically for candidate genes without pre-existing hypotheses.

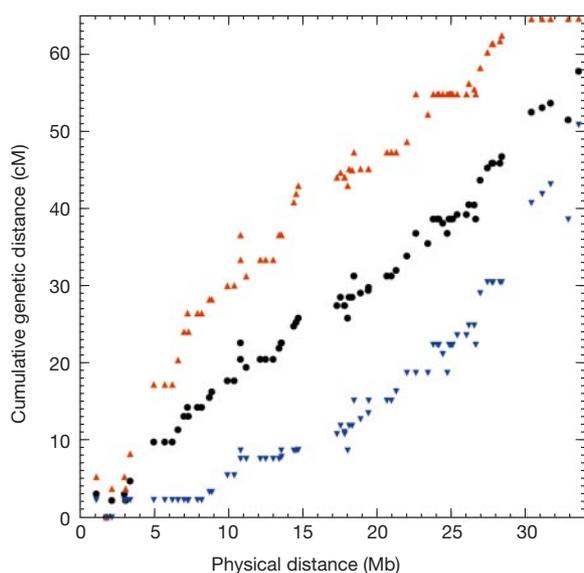
**Monogenic disorders.** Mutations in 14 known genes on chromo-

some 21 have been identified as the causes of monogenic disorders including one form of Alzheimer's disease (APP), amyotrophic lateral sclerosis (SOD1), autoimmune polyglandular disease (AIRE), homocystinuria (CBS) and progressive myoclonus epilepsy (CSTB); in addition, a locus for predisposition to leukaemia (AML1) has been mapped to 21q (for details of each of these disorders, see <http://www.ncbi.nlm.nih.gov/omim/>). The cloning of some of these genes, including the AIRE gene<sup>21,22</sup>, was facilitated by the sequencing effort. Loci for the following monogenic disorders have not yet been cloned: recessive nonsyndromic deafness (DFNB10 (ref. 23) and DFNB8 (ref. 24)), Usher syndrome type 1E<sup>25</sup>, Knobloch syndrome<sup>26</sup> and holoprocencephaly type 1 (HPE1 (ref. 27)). The gene catalogue and mapping coordinates will help in their identification. Mutation analysis of candidate genes in patients will lead to the cloning of the responsible genes.

**Complex phenotypes.** Two loci conferring susceptibility to complex diseases have been mapped to chromosome 21 (one for bipolar affective disorder<sup>28</sup> and one for familial combined hyperlipidaemia<sup>29</sup>) but the genes involved remain elusive.

**Neoplasias.** Loss of heterozygosity has been observed for specific regions of chromosome 21 in several solid tumours<sup>30–36</sup> including cancers of the head and neck, breast, pancreas, mouth, stomach, oesophagus and lung. The observed loss of heterozygosity indicates that there may be at least one tumour suppressor gene on this chromosome. The decreased incidence of solid tumours in individuals with Down syndrome indicates that increased dosage of some chromosome 21 genes may protect such individuals from these tumours<sup>37–39</sup>. On the other hand, Down syndrome patients have a markedly increased risk of childhood leukaemia<sup>19</sup>, and trisomy of chromosome 21 in blast cells is one of the most common chromosomal aneuploidies seen in childhood leukaemias<sup>40</sup>.

**Chromosome abnormalities.** Chromosome 21 is also involved in chromosomal aberrations including monosomies, translocations and other rearrangements. The availability of the mapped and sequenced clones now provides the necessary reagents for the accurate diagnosis and molecular characterization of constitutional and somatic chromosomal abnormalities associated with various phenotypes. This, in turn, will aid in identifying genes involved in mechanisms of disease development.



**Figure 5** Comparison of the genetic map and the sequence map of chromosome 21 aligned from centromere to telomere. Genetic distance in cM; physical distance in Mb. Each spot reflects the position of a particular genetic marker retrieved from <http://www.marshmed.org>. Black circles, sex-average; orange upwards triangles, female; blue downwards triangles, male.

The analysis of the genetic variation of many of the genes on chromosome 21 is of particular importance in the search for associations of polymorphisms with complex diseases and traits. Single nucleotide polymorphism (SNP) genotyping may also aid in the identification of modifier genes for numerous pathologies. Similarly, SNPs are useful tools in the development of diagnostic and predictive tests, which may eventually lead to individualized treatments. Chromosome-21-specific nucleotide polymorphisms will also facilitate evolutionary studies.

## Discussion

Our sequencing effort provided evidence for 225 genes embedded within the 33.8 Mb of genomic DNA of chromosome 21. Five hundred and forty-five genes have been identified in the 33.4 Mb of chromosome 22 (ref. 10). These data support the conclusion that chromosome 22 is gene-rich, whereas chromosome 21 is gene-poor. This finding is in agreement with data from the mapping of 30,181 randomly selected Unigene ESTs<sup>41</sup>. These two chromosomes together represent about 2% of the human genome and collectively contain 770 genes. Assuming that both chromosomes combined reflect an average gene content of the genome, we estimate that the total number of human genes may be close to 40,000. This figure is considerably lower than previous estimates, which range from 70,000 to 140,000 (ref. 42), and which were mainly based on EST clustering. It is possible that not all of the genes on chromosomes 21 and 22 have been identified. Alternatively, our assumption that the two chromosomes represent good models may be incorrect.

Our analysis of the chromosomal architecture revealed repeat units, duplications and breakpoints. A 93-bp repeat in the telomeric region, which was also found in other chromosomes, should provide a basis for studying the structural and functional organization and evolution of the telomere. One striking feature of chromosome 21 is that there is a 7-Mb region (positions 5.5–12.5 Mb) that contains only one gene. This region is much larger than the whole genome of *Escherichia coli*, but the evolutionary process permitted the existence of such a gene-poor DNA segment. Three other 1-Mb regions on 21q are also devoid of genes. Together, these gene-poor regions comprise almost 10 Mb, which is one-third of chromosome 21. Chromosome 22 also has a 2.5-Mb region near the telomeric end, as well as two other regions, each of 1 Mb, which are devoid of genes. We propose that similar large gene-less or gene-poor regions exist in other mammalian chromosomes. These regions may have a functional or architectural significance that has yet to be discovered.

Having the complete contiguous sequence of human chromosomes will change the methodology for finding disease-related genes. Disease genes will be identified by combining genetic mapping with mutation analysis in positional candidate genes. The laborious intermediate steps of physical mapping and sequencing are no longer necessary. Therefore, any individual investigator will be able to participate in disease gene identification.

The complete sequence analysis of human chromosome 21 will have profound implications for understanding the pathogenesis of diseases and the development of new therapeutic approaches. The clone collection represents a useful resource for the development of new diagnostic tests. The challenge now is to unravel the function of all the genes on chromosome 21. RNA expression profiling with all chromosome-21-specific genes may allow the identification of up- and downregulated genes in normal and disease samples. This approach will be particularly important for studying expression differences in trisomy and monosomy 21. Furthermore, chromosome-21-homologous genes can be systematically studied by overexpression and deletion in model organisms and mammalian cells.

The relatively low gene density on chromosome 21 is consistent with the observation that trisomy 21 is one of the only viable human autosomal trisomies. The chromosome 21 gene catalogue will open

new avenues for deciphering the molecular bases of Down syndrome and of aneuploidies in general. □

## Methods

Details of the protocols used by the five sequencing centres are available from our web sites (see below), including methods for the construction of sequence-ready maps and for sequencing large insert clones by shotgun cloning and nested deletion. Many software programs were used by the five groups for data processing, sequence analysis, gene prediction, homology searches, protein annotation and searches for motifs using pfam and SMART. Most of these programs are in the public domain. Software suites have been developed by the consortium members to allow efficient analysis. All information is available from the following web pages: RIKEN: <http://hgp.gsc.riken.go.jp>; Institut für Molekulare Biotechnologie, Jena: <http://genome.imb-jena.de>; Keio University: <http://www.dmb.med.keio.ac.jp>; GBF-Braunschweig: <http://genome.gbf.de>; Max-Planck-Institut für Molekulare Genetik (MPIMG), Berlin: <http://chr21.rz-berlin.mpg.de>.

Received 17 April; accepted 3 May 2000.

1. Lejeune, J., Gautier, M. & Turpin, R. Etude des chromosomes somatique des neuf enfants mongoliens. *CR Acad. Sci. Paris* **248**, 1721–1722 (1959).
2. McInnis, M. G. *et al.* A linkage map of human chromosome 21: 43 PCR markers at average intervals of 2.5 cM. *Genomics* **16**, 562–571 (1993).
3. Chumakov, I. *et al.* Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359**, 380–387 (1992).
4. Nizetic, D. *et al.* An integrated YAC-overlap and “cosmid-pocket” map of the human chromosome 21. *Hum. Mol. Genet.* **3**, 759–770 (1994).
5. Gardiner, K. *et al.* YAC analysis and minimal tiling path construction for chromosome 21q. *Somat. Cell Mol. Genet.* **21**, 399–414 (1995).
6. Korenberg, J. R. *et al.* A high-fidelity physical map of human chromosome 21q in yeast artificial chromosomes. *Genome Res.* **5**, 427–443 (1995).
7. Ichikawa, H. *et al.* A *NotI* restriction map of the entire long arm of human chromosome 21. *Nature Genet.* **4**, 361–366 (1993).
8. Hildmann, T. *et al.* A contiguous 3-Mb sequence-ready map in the S3-MX region on 21q22.2 based on high-throughput nonisotopic library screenings. *Genome Res.* **9**, 360–372 (1999).
9. Hattori, M. *et al.* A novel method for making nested deletions and its application for sequencing of a 300 kb region of human APP locus. *Nucleic Acids Res.* **25**, 1802–1808 (1997).
10. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
11. Korenberg, J. R. & Rykowski, M. C. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**, 391–400 (1988).
12. Antonarakis, S. E. 10 years of Genomics, chromosome 21, and Down syndrome. *Genomics* **51**, 1–16 (1998).
13. Saccone, S. *et al.* Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl Acad. Sci. USA* **90**, 11929–11933 (1993).
14. Zoubak, S., Clay, O. & Bernardi, G. The gene distribution of the human genome. *Gene* **174**, 95–102 (1996).
15. Jackson, M. S. *et al.* Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications and unstable sequences with homologies to telomeric and other centromeric locations. *Hum. Mol. Genet.* **8**, 205–215 (1999).
16. Dutriaux, A. *et al.* Cloning and characterization of a 135- to 500-kb region of homology on the long arm of human chromosome 21. *Genomics* **22**, 472–477 (1994).
17. Ruault, M. Juxta-centromeric region of human chromosome 21 is enriched for pseudogenes and gene fragments. *Gene* **239**, 55–64 (1999).
18. Graw, S. L. *et al.* Molecular analysis and breakpoint definition of a set of human chromosome 21 somatic cell hybrids. *Somat. Cell Mol. Genet.* **21**, 415–428 (1995).
19. Epstein, C. J. in *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C. R. *et al.*) 749–794 (McGraw-Hill, New York, 1995).
20. Kola, I. & Hertzog, P. J. Animal models in the study of the biological function of genes on human chromosome 21 and their role in the pathophysiology of Down syndrome. *Hum. Mol. Genet.* **6**, 1713–1727 (1997).
21. Nagamine, K. *et al.* Positional cloning of the APECED gene. *Nature Genet.* **17**, 393–398 (1997).
22. The Finnish-German APECED Consortium. An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. Autoimmune Polyendocrinopathy-Candidiasis-Ectodermal Dystrophy. *Nature Genet.* **17**, 399–403 (1997).
23. Bonné-Tamir, B. *et al.* Linkage of congenital recessive deafness (Gene DFNB10) to chromosome 21q22.3. *Am. J. Hum. Genet.* **58**, 1254–1259 (1996).
24. Veske, A. *et al.* Autosomal recessive non-syndromic deafness locus (DFNB8) maps to chromosome 21q22 in a large consanguineous kindred from Pakistan. *Hum. Mol. Genet.* **5**, 165–168 (1996).
25. Chaib, H. *et al.* A newly identified locus for Usher syndrome type I, USH1E, maps to chromosome 21q21. *Hum. Mol. Genet.* **6**, 27–31 (1997).
26. Sertie, A. L. *et al.* A gene which causes severe ocular alterations and occipital encephalocle (Knobloch syndrome) is mapped to 21q22.3. *Hum. Mol. Genet.* **5**, 843–847 (1996).
27. Estabrooks, L. L., Rao, K. W., Donahue, R. P., & Aylsworth, A. S. Holoprosencephaly in an infant with a minute deletion of chromosome 21(q22.3). *Am. J. Med. Genet.* **36**, 306–309 (1990).
28. Straub, R. E. *et al.* A possible vulnerability locus for bipolar affective disorder on chromosome 21q22.3. *Nature Genet.* **8**, 291–296 (1994).
29. Pajukanta, P. *et al.* Genomewide scan for familial combined hyperlipidemia genes in Finnish families, suggesting multiple susceptibility loci influencing triglyceride, cholesterol, and apolipoprotein B levels. *Am. J. Hum. Genet.* **64**, 1453–1463 (1999).
30. Sakata, K. *et al.* Commonly deleted regions on the long arm of chromosome 21 in differentiated adenocarcinoma of the stomach. *Genes Chromosomes Cancer* **18**, 318–321 (1997).

31. Kohno, T. *et al.* Homozygous deletion and frequent allelic loss of the 21q11.1–q21.1 region including the ANA gene in human lung carcinoma. *Genes Chromosomes Cancer* **21**, 236–243 (1998).
32. Ohgaki, K. *et al.* Mapping of a new target region of allelic loss to a 6-cM interval at 21q21 in primary breast cancers. *Genes Chromosomes Cancer* **23**, 244–247 (1998).
33. Yamamoto, N. *et al.* Frequent allelic loss/imbalance on the long arm of chromosome 21 in oral cancer: evidence for three discrete tumor suppressor gene loci. *Oncol. Rep.* **6**, 1223–1227 (1999).
34. Ghadimi, B. M. *et al.* Specific chromosomal aberrations and amplification of the AIB1 nuclear receptor coactivator gene in pancreatic carcinomas. *Am. J. Pathol.* **154**, 525–536 (1999).
35. Bockmuhl, U. *et al.* Genomic alterations associated with malignancy in head and neck cancer. *Head Neck* **20**, 145–151 (1998).
36. Schwendel, A. *et al.* Chromosome alterations in breast carcinomas: frequent involvement of DNA losses including chromosomes 4q and 21q. *Br. J. Cancer* **78**, 806–811 (1998).
37. Satge, D. *et al.* M. A tumor profile in Down syndrome. *Am. J. Med. Genet.* **78**, 207–216 (1998).
38. Hasle, H., Clemmensen, I. H., & Mikkelsen, M. Risks of leukaemia and solid tumours in individuals with Down's syndrome. *Lancet* **355**, 165–169 (2000).
39. Satge, D. *et al.* A lack of neuroblastoma in Down syndrome: a study from 11 European countries. *Cancer Res.* **58**, 448–452 (1998).
40. Wan, T. S., Au, W. Y., Chan, J. C., Chan, L. C. & Ma, S. K. Trisomy 21 as the sole acquired karyotypic abnormality in acute myeloid leukemia and myelodysplastic syndrome. *Leuk. Res.* **23**, 1079–1083 (1999).
41. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
42. Fields, C., Adams M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nature Genet.* **7**, 345–346 (1998).
43. Gyapay, G. *et al.* A radiation hybrid map of the human genome. *Hum. Mol. Genet.* **5**, 339–346 (1996).
44. Stewart, E. A. *et al.* An STS-based radiation hybrid map of the human genome. *Genome Res.* **7**, 422–433 (1997).
45. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
46. Murray, J. C. *et al.* A comprehensive human linkage map with centimorgan density. *Science* **265**, 2049–2054 (1994).

## Acknowledgements

The RIKEN group thank T. Itoh and C. Kawagoe for support of computational data management, M. Ohira and R. Ohki for clones and the members listed on <http://hgp.gsc.riken.go.jp> for technical support. The Jena group thank C. Baumgart, M. Dette, B. Drescher, G. Glöckner, S. Kluge, G. Nyakatura, M. Platzer, H.-P. Pöhle, R. Schattevoi, M. Schilling, J. Weber and all present and past members of the sequencing teams. The Keio group thank E. Nakato, M. Asahina, A. Shimizu, I. Abe, J. Wang, N. Sawada, M. Tatsuyama, M. Takahashi, M. Sasaki, H. Harigai and all members of the sequencing team, past and present. The MPIMG group thank M. Klein, C. Steffens, S. Arndt, K. Heitmann, I. Langer, D. Buczek, J. O'Brien, M. Christensen, T. Hildmann, I. Szulzewsky, E. Hunt and G. Teltow for technical support, and T. Haaf and A. Palotie for help with FISH. The German groups (IMB, GBF and MPIMG) thank the Resource Center of the German Human Genome Project (RZPD) and its group members for support and for clones and resources (<http://www.rzpd.de/>). We also thank J. Aaltonen, J. Buard, N. Creau, J. Gröet, R. Orti, J. Korenberg, M.C. Potier and G. Roizes for bacterial clones; D. Cox for discussions; A. Fortna, H.S. Scott, D. Slavov and G. Vacano for contributions; and N. Weizenbaum for editorial assistance. The RIKEN group is mainly supported by a Special Fund for the Human Genome Sequencing Project from the Science and Technology Agency (STA) Japan, and also by a Fund for Human Genome Sequencing from the Japan Society and Technology Corporation (JST) and a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sport and Culture, Japan. The Jena group was supported by the Federal German Ministry of Education, Research and Technology (BMBF) through Projektträger DLR, in the framework of the German Human Genome Project, and by the Ministry of Science, Research and Art of the Freestate of Thuringia (TMWFK). The Keio group was supported in part by the Fund for Human Genome Sequencing Project from the JST, Grants-in-Aid for Scientific Research, and the Fund for “Research for the Future” Program from the Japan Society for the Promotion of Science (JSPS); they also received support from Grants-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture of Japan. The Braunschweig group was supported by BMBF through Projektträger DLR, in the framework of the German Human Genome Project. The MPIMG-Berlin group acknowledge grants from BMBF through Projektträger DLR in the framework of the German Human Genome Project and from the EU. Support also came from the Boettcher Foundation, NIH, Swiss National Science Foundation, EU and MRC. We also thank E. Wain and the Human Gene Nomenclature Committee for working out the chromosome 21 gene symbols, and Y. Groner for cloning and sequencing the first gene on chromosome 21 (SOD).

Correspondence and requests for materials should be addressed to Y.S. (e-mail: sakaki@gsc.riken.go.jp), A.R. (e-mail: andrexlx@aol.com), N.S. (e-mail: shimizu@dmb.med.keio.ac.jp), H.B. (e-mail: bloecker@gbf.de) or M.L.Y. (e-mail: yasp@molgen.mpg.de). Genomic clones can be requested from any of the five groups. Detailed clone information, maps, FISH data, annotated gene catalogue, gene name alias and supporting data sets are available from the RIKEN and MPIMG web sites (see Methods). Interactive chromosome 21 databases (HSA21DB) are maintained at MPIMG and RIKEN. All sequence data can be obtained from Genbank, EMBL and DDBJ. They are also available from the individual web pages.