16. Siatecka, M., Rozek, M., Barciszewski, J. & Mirande, M. Modular evolution of the Glx-tRNA synthetase family: rooting of the evolutionary tree between the bacteria and archaea/eukarya branches. *Eur. J. Biochem.* **256,** 80–87 (1998).
17. Kim, S.-I. *et al.* A nuclear genetic lesion affecting *Saccharomyces cerevisiae* mitochondrial translation is complemented by a homologous *Bacillus* gene. *J. Bacteriol.* **179,** 5625–5627 (1997).
18. Nabholz, C. E., Hauser, R. & Schneider, A. *Leishmania tarentolae* contains distinct cytosolic and mitochondrial glutaminyl-tRNA synthetase activities. *Proc. Natl Acad. Sci. USA* **94,** 7903–7908 (1997).
19. Gupta, R. *Halobacterium volcanii* tRNAs: identification of 41 tRNAs covering all amino acids, and the sequences of 33 class I tRNAs. *J. Biol. Chem.* **259,** 9461–9471 (1984).
20. Ibba, M., Hong, K. W., Sherman, J. M., Sever, S. & Söll, D. Interactions between tRNA identity nucleotides and their recognition sites in glutaminyl-tRNA synthetase determine the cognate amino acid affinity of the enzyme. *Proc. Natl Acad. Sci. USA* **93,** 6953–6958 (1996).
21. Hayase, Y. *et al.* Recognition of bases in *Escherichia coli* $tRNA^{Gln}$ by glutaminyl-tRNA synthetase: a complete identity set. *EMBO J.* **11,** 4159–4165 (1992).
22. Ludmerer, S. W. & Schimmel, P. Gene for yeast glutamine tRNA synthetase encodes a large amino-terminal extension and provides a strong confirmation of the signature sequence for a group of the aminoacyl-tRNA synthetases. *J. Biol. Chem.* **262,** 10801–10806 (1987).
23. Fournand, D., Bigey, F. & Arnaud, A. Acyl transfer activity of an amidase from *Rhodococcus* sp. strain R312: formation of a wide range of hydroaxamic acids. *Appl. Environ. Microbiol.* **64,** 2844–2852 (1998).
24. Wong, J. T.-F. A co-evolution theory of the genetic code. *Proc. Natl Acad. Sci. USA* **72,** 1909–1912 (1975).
25. Sissler, M. *et al.* An aminoacyl-tRNA synthetase paralog with a catalytic role in histidine biosynthesis. *Proc. Natl Acad. Sci. USA* **96,** 8985–8990 (1999).
26. Weiner, A. M. Molecular evolution: aminoacyl-tRNA synthetases on the loose. *Curr. Biol.* **9,** R842–R844 (1999).
27. Di Giulio, M. The RNA world, the genetic code and the tRNA molecule. *Trends Genet.* **16,** 17–19 (2000).
28. Ibba, M., Bono, J. L., Rosa, P. A. & Söll, D. Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. *Proc. Natl Acad. Sci. USA* **94,** 14383–14388 (1997).
29. Stathopoulos, C. *et al.* One polypeptide with two aminoacyl-tRNA synthetase activities. *Science* **287,** 479–482 (2000).
30. Kim, R. *et al.* Overexpression of archaeal proteins in *Escherichia coli*. *Biotechnol. Lett.* **20,** 207–210 (1998).

Acknowledgements

We thank R. Hedderich for *M. thermoautotrophicum* Marburg cells, J. Reeve for *M. thermoautotrophicum* ΔH DNA, K. O. Stetter for *Pyrococcus* cells, and H. Kobayashi for *E. coli* GlnRS and *E. coli* $tRNA_2^{Gln}$ transcript. We also thank M. Ibba for critically reading the manuscript and S. Fitz-Gibbon, T. Hartsch, A. Johann, D. Oesterhelt, A. Ruepp and S. Schuster for sharing unpublished sequence data. D.L.T. and H.D.B. are postdoctoral fellows of the National Institute of General Medical Sciences and the EMBO, respectively. This work was supported by grants from the National Institute of General Medical Sciences.

Correspondence and requests for materials should be addressed to D.S. (e-mail: soll@trna.chem.yale.edu).

erratum

Atomically defined mechanism for proton transfer to a buried redox centre in a protein

Kaisheng Chen, Judy Hirst, Raul Camba, Christopher A. Bonagura, C. David Stout, Barbara K. Burgess & Fraser A. Armstrong

Nature **405,** 814–817 (2000).

The title to Box 1 in this paper was printed incorrectly. The correct title is 'Sequential electron and proton transfers at the [3Fe-4S] cluster in ferredoxin I'. □

corrections

The duration of antigen receptor signalling determines $CD4^+$ versus $CD8^+$ T-cell lineage fate

Koji Yasutomo, Carolyn Doyle, Lucio Miele, Chana Fuchs & Ronald N. Germain

Nature **404,** 506–510 (2000).

Chana Fuchs was inadvertantly omitted from the list of authors. Dr Fuchs is affiliated with the Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, Maryland 20892, USA. □

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, Y. Groner, 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. Schibuya, K. Kawasaki, S. Asakawa, A. Shintani, T. Sasaki, K. Nagamine, S. Mitsuyama, S. E. Antonarakis, S. Minoshima, N. Shimizu; G. Nordsiek, K. Hornischer, P. Brant, 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

Nature **405,** 311–319 (2000).

The name of one author, Y. Groner, was inadvertantly omitted from the author list but is included above. He is affiliated with the Department of Molecular Genetics and Human Genome Center, The Weizmann Institute of Science, Rehovot, Israel. Also, in the second paragraph of the section headed 'Comparison of chromosome 21 with other autosomes', the second sequence should read 'For example, a 100-kb region of a contig (AP001656) on 21p is shared with chromosomes 4, 7, 20 and 22.' □

Atomically defined mechanism for proton transfer to a buried redox centre in a protein

Kaisheng Chen*†, Judy Hirst†‡§, Raul Camba†‡, Christopher A. Bonagura*, C. David Stout‖, Barbara. K. Burgess* & Fraser A. Armstrong‡

** Department of Molecular Biology and Biochemistry, University of California, Irvine, California 92612, USA*
‡ Department of Chemistry, Oxford University, Oxford OX1 3QR, UK
‖ Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037-1083, USA
† These authors contributed equally to this work
§ Present address: Medical Research Council Dunn Human Nutrition Unit, Hills Road, Cambridge CB2 2XY, UK

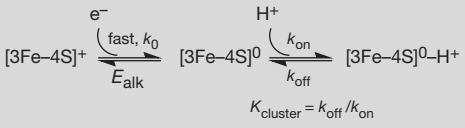
The basis of the chemiosmotic theory is that energy from light or respiration is used to generate a trans-membrane proton gradient[1]. This is largely achieved by membrane-spanning enzymes known as 'proton pumps'[2–5]. There is intense interest in experiments which reveal, at the molecular level, how protons are drawn through proteins[6–13].Here we report the mechanism, at atomic resolution, for a single long-range electron-coupled proton transfer. In *Azotobacter vinelandii* ferredoxin I, reduction of a buried iron–sulphur cluster draws in a solvent proton, whereas re-oxidation is 'gated' by proton release to the solvent. Studies of this 'proton-transferring module' by fast-scan protein film voltammetry, high-resolution crystallography, site-directed mutagenesis and molecular dynamics, reveal that proton transfer is exquisitely sensitive to the position and pK of a single amino acid. The proton is delivered through the protein matrix by rapid penetrative excursions of the side-chain carboxylate of a surface residue (Asp 15), whose pK shifts in response to the electrostatic charge on the iron–sulphur cluster. Our analysis defines the structural, dynamic and energetic requirements for proton courier groups in redox-driven proton-pumping enzymes.

Current studies of proton-pumping enzymes such as bacteriorhodopsin and cytochrome *c* oxidase are revealing the existence of 'proton wires' comprising chains of water molecules and protonatable amino-acid side-chains[2–12]. Apart from water channels, which provide natural proton conductors[14], transfer of protons is heavily restricted by their short tunnelling distance; thus, whereas electrons may easily tunnel 10 Å, a proton with comparable energy is limited to hops of less than 0.25 Å[15–17]. This restriction can be used, along with barriers imposed by pK differences between proton donors and acceptors[18], to control proton flow in response to electron-transfer events at nearby redox sites[2–12]. However, proton pumps are complex membrane-bound enzymes and mechanistic details are difficult to analyse coherently. We now report conclusive studies of the mechanism of redox-driven proton transfer between solvent and a buried iron–sulphur cluster ([3Fe–4S]) in ferredoxin I (FdI) from *A. vinelandii*. This small protein provides an electron/proton-transfer 'module' with which to understand the redox-linked proton-transferring components of the larger enzymes[13,19]. A major advantage of the FdI system is that native and mutant structures are known to high structural resolution in all

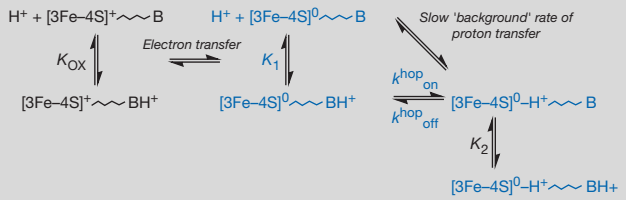
Box 1

Sequential electron and proton transfers at the [3fe–45] cluster in ferredoxin I

Scheme I, Sequence of electron and proton transfers defining the redox chemistry of the [3Fe–4S] cluster in Ferredoxin I. Electron transfers (standard first-order electrochemical rate constant, k_0) are fast, and E_{alk} is the reduction potential if pH ≫ $pK_{cluster}$.



Scheme II, Sequence by which proton transfer to the cluster is catalysed by Asp 15 (B). Fast proton transfer (species highlighted in blue) is pH dependent, and protonation constants of Asp 15 are sensitive to cluster charge. At low pH, Asp 15 re-protonates (K_2), thus inhibiting proton transfer off the cluster. For native *A. vinelandii* FdI, pK_{OX} = 5.4. See Table I for rate expressions.



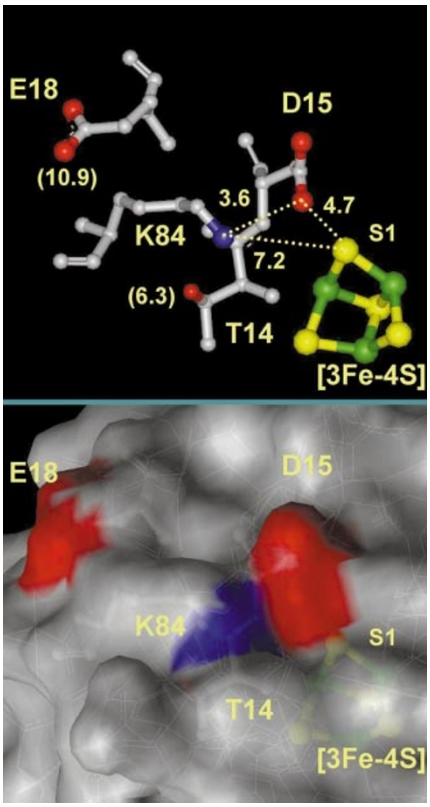


Figure 1 The residues in the vicinity of the [3Fe–4S] cluster in oxidized native *Azotobacter vinelandii* Ferredoxin I, taken from coordinates 7FD1. The upper frame shows the side chains of residues that were varied in this study, and indicates the separation (in Å) between D15 Oδ$_1$ and S1, the least buried μ_2 sulphide of [3Fe–4S]. Only the major conformation of the K84 side chain is shown (68% occupancy). The lower frame shows the same view, onto which is superimposed the solvent-accessible surface (calculated with a 1.4 Å probe). The carboxylate Oδ atoms of D15 and E18, and Nζ of K84 are solvent exposed. The Lys 84 side chain adopts two conformations in the oxidized native protein, as shown by the pH 8.5 structure (Protein data bank (PDB) code 7FD1, ref. 22) where these have 68% and 32% relative occupancies and (K84)Nζ–(D15)Oδ$_1$ distances of 3.6 Å and 3.1 Å respectively. The Glu 18 side chain is disordered (anomalously high B values).

relevant oxidation and protonation states[20–22]. Importantly, FdI can be studied by fast-scan protein film voltammetry, a technique able to reveal, in detail, the bi-directional kinetics of coupled proton–electron-transfer reactions and their relation to thermodynamics[13,23] (See Supplementary Information).

Scheme I (Box 1) shows the bi-directional proton transfer that accompanies fast electron transfer to the $[3Fe–4S]^{1+/0}$ cluster of FdI. Reduction drives a proton onto the cluster (k_{on}), while its reoxidation is 'gated' by proton release (k_{off}). Scheme I stems from numerous lines of evidence. The $[3Fe–4S]^{1+/0}$ reduction potential is pH-dependent, and one proton is taken up in the reduced state ($pK_{cluster} = 7.8$)[13,23]. This uptake can be observed spectroscopically: thus, circular dichroism and magnetic circular dichroism spectra of oxidized FdI (spin-state (S) = 1/2) are independent of pH, whereas the one-electron reduced FdI ($[3Fe–4S]^0$, S = 2) shows an acid–base transition[24]. The X-ray structures of oxidized and reduced FdI at high and low pH eliminate the possibility that the pH-dependent spectral changes are due to ligand exchange and/or structural rearrangements[20]. Site-directed mutagenesis has established that the spectral changes are not due to protonation of the nearest ionizable residue (Asp 15)[19], while Mössbauer spectroscopy has revealed that protonation perturbs the electronic structure of the cluster[25]. The obvious protonation site is a μ_2 sulphide (one of the three cluster sulphur atoms with a free coordination site).

Figure 1 shows the region above the buried [3Fe–4S] cluster, both within the protein and looking down on the surface. Even at highest resolution (1.4 Å), the crystal structures, and NMR on a related protein[26], reveal no internal water molecules (or accommodating spaces) to act as proton-transfer agents[14,20–22]. The starting point for this study is the mutant (D15N) in which Asp 15, the closest ionizable residue, which moves upon cluster reduction[20], is changed to Asn[19]. Protein film voltammetry showed that proton coupling is severely retarded in D15N FdI, indicating that the carboxylate acts as a proton relay group[13,19]. However, in the native protein, Asp 15 is salt-bridged to Lys 84, raising the possibility of proton migration across the bridge. In addition, the highly mobile side chain of a second surface carboxylate residue, Glu 18, lies close by.

With the ability to make extensive kinetic, thermodynamic and structural measurements, we designed several new mutants. Figure 2 shows the arrangements of cluster-region amino acids for all variants, with distances between relevant atoms and the μ_2 sulphide (S1) situated closest to the protein surface. Circular dichroism spectroscopy confirmed that protonation of $[3Fe—4S]^0$ occurs in all cases. Table 1 displays the corresponding kinetic and

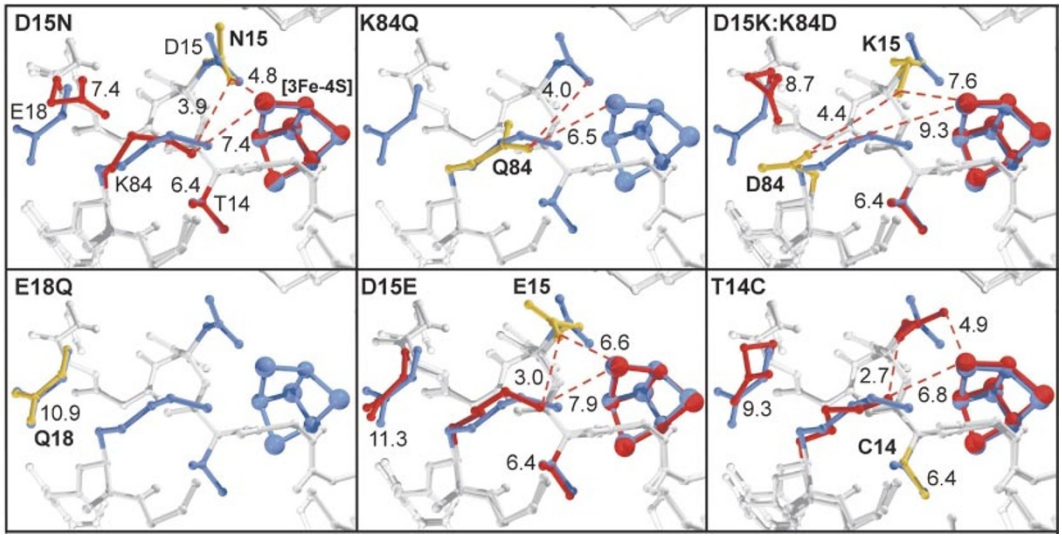


Figure 2 Structures of the various mutant forms of FdI in the region of interest, compared with the structure of oxidized native FdI. Side chains for the native protein are shown in blue and those altered by mutagenesis in yellow. Where X-ray structures are available, the positions of the other residues in the mutant proteins are shown in red. All distances are given in ångstroms and where lines are not shown the number refers to the distance between the indicated residue and S1. Structures have been solved for D15N (PDB code 1FDD, ref. 19), D15K/K84D (PDB code 1B0T), D15E (PDB code 1D3W) and T14C (PDB code 1A6L, ref. 27). Crystals have not been obtained for K84Q or E18Q: these structures were modelled using the Insight II/Biopolymer program from Molecular Simulations for which the native oxidized structure (1.35 Å, PDB code 7FD1) was substituted with the desired mutant residue at the program-assigned minimum energy orientation.

Table 1 Kinetic and thermodynamic parameters for 'on' and 'off' proton transfers

Fast	E_{alk} (V)	$pK_{cluster}$ (high pH)	$pK_{cluster}$ (low pH)	$k_{on}(M^{-1}s^{-1})$* (pH 7.0)	$k_{off}(s^{-1})$* (pH 7.0)	pK_1	pK_2	$k_{on}^{hop}(s^{-1})$	$k_{off}^{hop}(s^{-1})$
Native	–0.443	7.8 ± 0.1	6.5 ± 0.1	7.9×10^9	308	7.2 ± 0.1	5.9 ± 0.1	1,294 ± 100	332 ± 25
E18Q	–0.453	7.7 ± 0.1	6.7 ± 0.1	4.8×10^9	207	7.1 ± 0.1	6.1 ± 0.1	930 ± 90	230 ± 25
T14C	–0.464	8.4 ± 0.1	7.1 ± 0.1	6.6×10^9	207	8.0 ± 0.1	6.7 ± 0.1	720 ± 70	310 ± 25
K84Q	–0.476	8.1 ± 0.1	6.6 ± 0.1	9.0×10^9	232	7.4 ± 0.1	5.9 ± 0.1	1,252 ± 100	250 ± 25
Native (D_2O)	–0.443	7.8 ± 0.1	6.5 ± 0.1	6.0×10^9	222	7.2 ± 0.1	5.9 ± 0.1	970 ± 100	240 ± 25
Slow	E_{alk} (V)	$pK_{cluster}$		$k_{on}(M^{-1}s^{-1})$	$k_{off}(s^{-1})$				
D15N	–0.408	6.9 ± 0.1		2.0×10^7	2.5 ± 0.1				
D15K–K84D	–0.397	6.6 ± 0.1		1.2×10^7	3.0 ± 0.1				
D15E	–0.388	6.7 ± 0.1		2.0×10^7	4.5 ± 0.2				

All terms are as defined in Schemes I and II.
*For the fast reactions, interpreted in terms of Scheme II, $k_{on} = k_{on}^{hop}/([H^+] + K_1)$ and $k_{off} = k_{off}^{hop}K_2/([H^+] + K_2)$. Agreement with the simple bimolecular rate law of Scheme I requires consideration of the fact that interaction with Asp 15 causes the pK of the cluster to differ at high and low pH. In all cases k_0, the standard first-order electrochemical rate constant for electron exchange at the reduction potential, is $> 200\,s^{-1}$; the exponential increase in rate that occurs as a driving force is applied means that electron transfer is never rate limiting.

thermodynamic parameters for 'on' and 'off' proton transfers, defined according to Schemes I and II (Box 1), and measured by protein film cyclic voltammetry. In all cases, reduction involves electron transfer followed by proton transfer, whereas oxidation requires that proton transfer precede electron transfer, that is, the two processes do not occur in concert. Electron transfer is very fast and is clearly separated from proton transfer.

The mutants fall sharply into two categories with respect to proton-transfer kinetics—'slow' and 'fast'—with rates differing by three orders of magnitude. The kinetics of the 'slow' mutants are simple and pH-independent (they are described adequately by Scheme I) and rates are similar in each case. By contrast, the 'fast' mutants show a complex pH dependence (Scheme II) with second-order 'on' rate constants approaching diffusion control at pH 7.

Referring to Fig. 2, the following facts are established. The D15N mutation replaces the carboxylate by carbamide and the slow proton-transfer rate is consistent with loss of a protonatable group (base B) which acts as a proton relay (Scheme II)[13,19]. However, the mutation also disrupts the salt bridge between the Asp 15 carboxylate and a surface lysine (Lys 84)[19]. In addition, the side chain of Glu 18 shifts, although the change is less-defined due to disorder. In the mutant K84Q, Lys 84 is changed to a residue (Gln) that cannot form a salt-bridge to Asp 15. The fast proton-transfer kinetics demonstrate that it is the presence of the carboxylate and not the salt-bridge that is important. The double mutant D15K/K84D was designed to invert the salt-bridge orientation; however, the salt-bridge does not form. Proton transfer is slow, as it is in D15N, showing that the introduced Asp 84 side chain is incorrectly placed to function as a proton-transfer group and that lysine cannot substitute for aspartate at position 15. The possibility that Glu 18 facilitates proton transfer was eliminated by the mutant E18Q, which has very similar rates to the native protein. Experiments with D15E finally established the critical nature of the distance between the cluster and the carboxylate at position 15. Insertion of a single CH_2 group in the side chain (which increases the distance by approximately 2 Å) retards proton transfer as much as deletion of the carboxylate altogether (Table I). The T14C mutant has a polarizable S atom within the sphere of influence of the cluster and raises the cluster pK while the position of the carboxylate is not significantly changed[27]. The native proton-transfer kinetics are retained, confirming that they are dependent solely on the position of the carboxylate.

The exacting requirement for Asp 15 is thus demonstrated, leading to a detailed model for the mechanism of proton transfer between water at the protein surface and the buried redox centre. The data in Table 1 show that the H_2O/D_2O isotope effect for the native protein is small (approximately 1.3), and reveal (from the interdependence of respective pK values) that there are significant electrostatic interactions between the cluster and Asp 15. As shown in Fig. 3a, electron transfer drives rotation of the Asp 15 carboxylate about the Cβ–Cγ bond and increases the Oδ to S1 distance from 4.7 Å to 4.9 Å, in accordance with the less favourable electrostatics[22]. More significantly, the pK of the carboxylate increases to 7.2 compared with 5.4 in the oxidized ($[3Fe–4S]^{1+}$) state (pK_{OX} in Scheme II), thus promoting proton capture from solvent water[18]. It is important to note that a sizeable, transient shift in pK value is essential for fitting the data. After the proton has transferred to the cluster, the pK drops back to 5.9, and the X-ray structure of the reduced protein at pH 6.1 reveals that the Asp 15 carboxylate reverts towards its position in the oxidized form[20]. This 'relaxation' is expected because the $[3Fe–4S]^0$-H^+ cluster has the same electrical charge as the oxidized $[3Fe–4S]^{1+}$ cluster.

To explain how Asp 15 mediates such fast proton transfer, it is necessary to ascertain whether a proton that has transferred from solvent to a carboxylate O atom can then be carried easily to within hydrogen-bonding distance of the cluster S1 atom[28]. We therefore carried out molecular dynamics calculations to study the mobility of the state in which the reduced cluster is unprotonated while Asp 15 is protonated and thus no longer repelled by the increased negative charge on the cluster. The results reveal that the Asp 15 side chain is very mobile, executing a high frequency of short-range encounters between Oδ and S1 atoms (Fig. 3b). Thus, within 80 ps, an Oδ atom makes one excursion to 3.05 Å (Fig. 3c), well within the sum of van der Waals radii, and five other excursions to distances less than 4 Å. The motions of the side-chain carboxyl/carboxylate group thus convey H^+ between solvent and reduced cluster, accomplishing 'atom-to-atom' transfer across this hydrophobic barrier. Cluster deprotonation occurs by reversal of these steps, noting that at pH < 5.9, protonation of Asp 15 inhibits proton transfer off the cluster.

In conclusion, fast proton transfer in FdI requires the presence of Asp 15 and rapid penetrative excursions of its side-chain carboxyl/carboxylate to within hydrogen-bonding distance of the cluster. Our experiments, which combine detailed kinetic and thermodynamic data for a series of mutants with molecular dynamics based on high-resolution protein crystal structures, highlight the exacting specifications that must be met for proton-pumping motifs in enzymes[2–5]. The pK values of the proton-relaying carboxylate and the buried active site are tightly coupled and adjust to facilitate sequential transfers of an electron and a proton. Re-oxidation is gated by

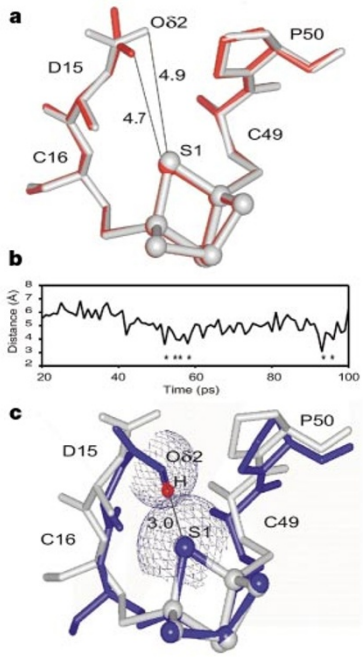


Figure 3 Movement of the Asp15 side chain during redox-driven proton transfers. **a**, Comparison of oxidized (red; PDB code 7FD1) and reduced (grey; PDB code 7FDR) high-pH structures of native FdI in the region of the cluster and Asp 15. Upon reduction, the Lys 84 side chain adopts a single conformation with a Nζ–$Oδ_1$ distance of 3.3 Å, and the carboxylate side chain has rotated by 90°. (In accordance with previous nomenclature, the closest O atom to the cluster in the reduced form is denoted $Oδ_2$, whereas it is $Oδ_1$ in the oxidized form[22].) **b**, Distances of $Oδ_2$ (from Asp 15) to S1 (from the cluster) during the molecular dynamics simulation; asterisks indicate the points where the distance is below 4 Å. **c**, Superposition of the reduced native FdI (grey) structure (also shown in **a**) and one of the molecular-dynamics-generated snapshot structures (blue) around the region of Asp 15, showing one of the approaches of $Oδ_2$ to the S1 atom of $[3Fe–4S]^0$ within the period of simulation. The distance between $Oδ_2$ and S1 is 4.9 Å for the native structure and 3.0 Å for the molecular-dynamics-generated structure. The H atom is positioned collinear with $Oδ_2$–S1 at a normal bond distance (1.0 Å) from $Oδ_2$. Van der Waals surfaces of $Oδ_2$ (radius 1.4 Å) and S1 (1.85 Å) are shown. For this closest approach, Oδ and S1 are well within the sum (1.4 + 1.85 = 3.25 Å) of their van der Waals radii and, in the limit of a collinear O–H···S arrangement, the resulting H–S distance is approximately 2.25 Å.

proton transfer, demonstrating how biology can control long-range electron transfer by linking it to a far more discriminating process. Mechanistic requirements for proton transfer are so stringent that even minor structural changes produce decreases in rate over orders of magnitude. □

Methods

Protein film voltammetry

Protein film voltammetry probes electron-transfer processes that are induced by perturbing the electrochemical potential applied to a mono-/submono-layer of protein molecules bound at an electrode surface (See Supplementary Information). Electron transfers in and out of the active site are observed as electrical current signals, which are altered in specific ways if there is coupling to processes such as proton transfer. Thus, by analysing signals over a range of voltage scan rate and pH, a detailed and integrated picture of the coupling kinetics and energetics is obtained. All measurements were carried out as described previously[13] and 'trumpet plots' of peak positions versus scan rate were analysed in terms of the kinetics and thermodynamics described by Scheme II.

Site directed mutagenesis and protein crystallography

Site directed mutagenesis and protein purification were carried out as previously reported[19,27]. Crystal structures were determined accordingly to procedures recently described[22].

Molecular dynamics

Molecular dynamics simulations were carried out using the DISCOVER program from Molecular Simulations. The AMBER force field used for the protein and Fe–S cluster was as described[29] except that cluster atomic charges were derived from Xα density functional calculations[30]. The structure of reduced FdI at pH 8.5 (1.35 Å resolution, protein data bank code 7FDR, ref. 21) was used as the starting model except that the $O\delta_2$ atom of Asp 15 was protonated. Besides the crystallographic water molecules, a 9 Å layer of water molecules was added around the protein and those occupying the outermost 5 Å were constrained to prevent their evaporation. After initial minimization, the system was heated for 5 ps at 100, 200, 273 K, and then the dynamics were run at 273 K for 100 ps with a time step of 1.5 fs: the first 20 ps were discarded. The mean structure over the 80 ps considered had r.m.s. deviation values from the starting structure of 1.20 Å for the backbone atoms and 1.45 Å for all heavy atoms. The individual structures generated had average r.m.s. deviation values (from the mean structure) of 0.68 Å for the backbone atoms and 0.86 Å for all heavy atoms, within the range expected[26].

Received 26 November 1999; accepted 12 April 2000.

1. Nicholls, D. G. & Ferguson, S. J. Bioenergetics 2 (Academic, San Diego, 1992).
2. Wikström, M. Proton translocation by bacteriorhodopsin and heme-copper oxidases. *Curr. Opin. Struct. Biol.* **8,** 480–488 (1998).
3. Gennis, R. B. How does cytochrome oxidase pump protons? *Proc. Natl Acad. Sci. USA* **95,** 12747–12749 (1998).
4. Malmström, B. G. Cytochrome oxidase: pathways for electron tunneling and proton transfer. *J. Biol. Inorg. Chem.* **3,** 339–343 (1998).
5. Michel, H. The mechanism of proton pumping by cytochrome *c* oxidase. *Proc. Natl Acad. Sci. USA* **95,** 12819–12824 (1998).
6. Rammelsberg, R., Huhn, G., Lübben, M. & Gerwert, K. Bacteriorhodopsin's intramolecular proton-release pathway consists of a hydrogen-bonded network. *Biochemistry* **37,** 5001–5009 (1998).
7. Luecke, H., Schobert, B., Richter, H.-T., Cartailler, J. P. & Lanyi, J. Structural changes in bacteriorhodopsin during ion transport at 2 Å resolution. *Science* **286,** 255–260 (1999).
8. Nabedryk, E., Breton, J., Okamura, M. Y. & Paddock, M. L. Proton uptake by carboxylic acid groups upon photoreduction of the secondary quinone (Q_B) in bacterial reaction centres from Rhodobacter sphaeroides: FTIR studies on the effects of replacing Glu H173. *Biochemistry* **37,** 14457–14462 (1998).
9. Yoshikawa, S. *et al.* Redox-coupled crystal structural changes in bovine heart cytochrome *c* oxidase. *Science* **280,** 1723–1729 (1998).
10. Konstantinov, A. A., Siletsky, S., Mitchell, D., Kaulen, A. & Gennis, R. B. The roles of two proton input channels in cytochrome c oxidase from *Rhodobacter sphaeroides* probed by the effects of site-directed mutations on time-resolved electrogenic intraprotein proton transfer. *Proc. Natl Acad. Sci. USA* **94,** 9085–9090 (1997).
11. Lübben, M., Prutsch, A., Mamat, B. & Gerwert, K. Electron transfer induces side-chain conformational changes of glutamate-286 from cytochrome bo_3. *Biochemistry* **38,** 2048–2056 (1999).
12. Junemann, S., Meunier, B., Fisher, N. & Rich, P. R. Effects of mutation of the conserved glutamic acid-286 in subunit I of cytochrome *c* oxidase from *Rhodobacter sphaeroides*. *Biochemistry* **38,** 5248–5255 (1999).
13. Hirst, J. *et al.* Kinetics and mechanism of redox-coupled, long-range proton transfer in an iron-sulfur protein. Investigation by fast-scan protein-film voltammetry. *J. Am. Chem. Soc.* **120,** 7085–7094 (1998).
14. Meyer, E. Internal water molecules and H-bonding in biological macromolecules: a review of structural features with functional implications. *Protein Sci.* **1,** 1543–1562 (1992).
15. Beratan, D. N., Onuchic, J. N., Winkler, J. R. & Gray, H. B. Electron-tunneling pathways in proteins. *Science* **258,** 1740–1741 (1992).
16. Page, C. G., Moser, C. C., Chen, X. & Dutton, P. L. Natural engineering principles of electron tunnelling in biological oxidation-reduction. *Nature* **402,** 47–52 (1999).
17. Klinman, J. P. Quantum mechanical effects in enzyme-catalyzed hydrogen transfer reactions. *Trends Biochem. Sci.* **14,** 368–373 (1989).
18. Gutman, M. & Nachliel, E. The dynamics of proton exchange between bulk and surface groups. *Biochim. Biophys. Acta* **1231,** 123–138 (1995).
19. Shen, B. *et al. Azotobacter vinelandii* ferredoxin I. Aspartate 15 facilitates proton transfer to the reduced [3Fe–4S] cluster. *J. Biol. Chem.* **268,** 25928–25939 (1993).
20. Stout, C. D. Crystal structures of oxidized and reduced *Azotobacter vinelandii* ferredoxin at pH 8 and 6. *J. Biol. Chem.* **268,** 25920–25927 (1993).
21. Stout, C. D., Stura, E. A. & McRee, D. E. Structure of Azotobacter vinelandii 7Fe ferredoxin at 1. 35 Å resolution and determination of the [Fe–S] bonds with 0. 01 Å accuracy. *J. Mol. Biol.* **278,** 629–639 (1998).
22. Schipke, C. G., Goodin, D. B., McRee, D. E. & Stout, C. D. Oxidized and reduced *Azotobacter vinelandii* ferredoxin I at 1. 4 Å resolution: conformational change of surface residues without significant change in the $[3Fe–4S]^{+/0}$ cluster. *Biochemistry* **38,** 8228–8239 (1999).
23. Armstrong, F. A., Heering, H. A. & Hirst, J. Reactions of complex metalloproteins studied by protein-film voltammetry. *Chem. Soc. Rev.* **26,** 169–179 (1997).
24. Stephens, P. J. *et al.* Circular-dichroism and magnetic circular-dichroism of *Azotobacter vinelandii* ferredoxin I. *Biochemistry* **30,** 3200–3209 (1991).
25. Hu, Z. G., Jollie, D., Burgess, B. K., Stephens, P. J. & Münck, E. Mössbauer and EPR studies of *Azotobacter vinelandii* ferredoxin I. *Biochemistry* **33,** 14475–14485 (1994).
26. Aono, S. *et al.* Solution structure of oxidized Fe_7S_8 ferredoxin from the thermophillic bacterium *Bacillus schlegelii* by ^1H NMR spectroscopy. *Biochemistry* **37,** 9812–9826 (1998).
27. Gao-Sheridan, A T14C variant of *Azotobacter vinelandii* ferredoxin I undergoes facile $[3Fe–4S]^0$ to $[4Fe–4S]^{2+}$ conversion in vitro but not *in vivo*. *J. Chem. Biol.* **273,** 33692–33701 (1998).
28. Guthrie, J. P. Intrinsic barriers for protons transfer reactions involving electronegative atoms, and the water mediated proton switch: an analysis in terms of Marcus theory. *J. Am. Chem. Soc.* **118,** 12886–12890 (1996).
29. Banci, L., Bertini, I., Carloni, P., Luchinat, C. & Orioli, P. L. Molecular-dynamics on HIPIP from *Chromatium vinosum* and comparison with NMR data. *J. Am. Chem. Soc.* **114,** 10683–10689 (1992).
30. Noodleman, L., Norman, J. G. Jr, Osborne, J. H., Aizman, A. & Case, D. A. Models for ferredoxins—electronic structures of iron sulfur clusters with one, two and four iron atoms. *J. Am. Chem. Soc.* **107,** 3418–3426 (1985).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank T. Poulos and J. Lanyi for comments on the manuscript. This research was supported by grants from the NIH, EPSRC, and BBRSC. B.B.K. thanks The Fulbright Commission for a Senior Scholarship, and the John Simon Guggenheim Foundation for a Travelling Fellowship. R.C. is grateful to The National Council of Science and Technology of Mexico (CONACYT) for their support.

Correspondence and requests for materials should be addressed to F.A.A. (e-mail: fraser.armstrong@chem.ox.ac.uk).

29. Bateman, A. *et al.* Pfam 3.1: 1,313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27,** 260–262 (1999).
30. Higgins, D. G., Bleasby, A. J. & Fuchs, R. CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8,** 189–191 (1992).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

This research was funded by The Wellcome Trust.

Correspondence and requests for materials should be addressed to J.P. (e-mail: parkhill@sanger.ac.uk). The complete sequence and annotation can be obtained from the EMBL database with the ID NMAZ2491 (accession number AL157959), and from our web pages (http://www.sanger.ac.uk/Projects/N_meningitidis).

The duration of antigen receptor signalling determines CD4$^+$ versus CD8$^+$ T-cell lineage fate

Koji Yasutomo*, Carolyn Doyle†, Lucio Miele‡ & Ronald N. Germain*

** Lymphocyte Biology Section, Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA*
† Department of Immunology, Duke University Medical Centre, Durham, North Carolina 27710, USA
‡ Cancer Immunology Program, Cardinal Bernardin Cancer Centre, Loyola University Medical Centre, Maywood, Illinois 60153, USA

Signals elicited by binding of the T-cell antigen receptor and the CD4/CD8 co-receptor to major histocompatibility complex (MHC) molecules control the generation of CD4$^+$ (helper) or CD8$^+$ (cytotoxic) T cells from thymic precursors that initially express both co-receptor proteins[1]. These precursors have unique, clonally distributed T-cell receptors with unpredictable specificity for the self-MHC molecules involved in this differentiation process[2]. However, the mature T cells that emerge express only the CD4 (MHC class II-binding) or CD8 (MHC class I-binding) co-receptor that complements the MHC class-specificity of the T-cell receptor. How this matching of co-receptor-defined lineage and T-cell-receptor specificity is achieved remains unknown[1,3,4], as does whether signalling by the T-cell receptors, co-receptors and/or general cell-fate regulators such as Notch-1 (refs 5, 6) contributes to initial lineage choice, to subsequent differentiation processes or to both. Here we show that the CD4 versus CD8 lineage fate of immature thymocytes is controlled by the co-receptor-influenced duration of initial T-cell receptor-dependent signalling. Notch-1 does not appear to be essential for this fate determination, but it is selectively required for CD8$^+$ T-cell maturation after commitment directed by T-cell receptors. This indicates that the signals constraining CD4 versus CD8 lineage decisions are distinct from those that support subsequent differentiation events such as silencing of co-receptor loci.

The AND T-cell receptor (TCR) is specific for a pigeon cytochrome *c* peptide bound to the MHC class II molecule I-E^k (ref. 7). Transgenic thymocytes expressing this TCR are efficiently selected into the CD4$^+$ lineage in mice expressing wild-type I-A^b MHC class II molecules. In contrast, AND TCR transgenic mice expressing mutant I-A^b molecules that are defective in interaction with CD4 but not the TCR[8] generate CD8$^+$ but not CD4$^+$ mature cells[9]. To investigate how altering CD4 co-receptor binding controls the lineage fate of AND thymocytes, we took advantage of a modified two-stage reaggregate culture system[10] that allows controlled delivery of MHC-dependent and independent signals to thymocytes at distinct stages of maturation (Yasutomo *et al.*, manuscript in preparation). Immature CD69^{lo}CD4$^+$CD8$^+$ TCR transgenic thymocytes (double positive; DP) are incubated in dispersed culture with cells expressing the desired MHC molecule ligands, to initiate selection events (culture 1). This first TCR stimulation does not lead to silencing of the CD4 or CD8 locus in dispersed culture for up to three days, but it does rapidly induce upregulation of the activation marker CD69 that characterizes thymocytes undergoing maturation *in vivo*[1,11]. When the CD69$^{med/hi}$ thymocytes arising in culture 1 after 20 h are purified and reaggregated with either MHC-positive thymic stromal cells (TSC) or MHC-negative TSC in the presence of dendritic cells of the appropriate MHC type (culture 2), mature functional T cells expressing only a single co-receptor develop over the next 60 h. Experiments using this model showed that TCR–MHC molecule interactions are required not only in culture 1 to generate the CD69^{hi} thymocytes, but also in culture 2 to generate mature T cells. Using this approach, we asked whether the first, the second or both sets of TCR/co-receptor–MHC interaction events determined the differentiation of AND TCR transgenic T cells along the CD4 versus CD8 pathways.

Thymocytes from mice expressing only the AND TCR in the absence of MHC class II molecules (AND TCR transgenic RAG-2$^{-/-}$ MHC A$\beta^{-/-}$, referred to as AND throughout) were stimulated in culture 1 using dendritic cells from wild-type mice (WT-DC) or mutant (Mu-DC) I-A^b transgenic mice. When purified cells of the CD69$^+$CD4$^+$CD8$^+$ phenotype (CD69$^+$DP) generated by this stimulation were reaggregated with TSC from MHC class II$^{-/-}$ mice, no CD4$^+$ or CD8$^+$ mature T cells developed (Fig. 1a, b). Inclusion of

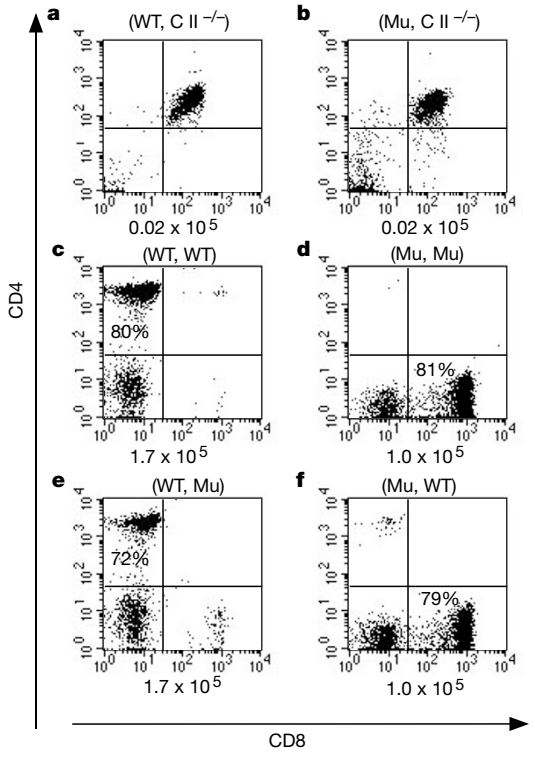


Figure 1 Role of co-receptor in CD4/CD8 lineage choice. Sorted CD69$^+$CD4$^+$CD8$^+$ cells obtained after stimulation by wild-type dendritic cells (WT-DC; **a**, **c**, **e**) or mutant dentritic cells (Mu-DC; **b**, **d**, **f**) were cultured with MHC class II$^{-/-}$ TSC (**a**, **b**) or MHC$^{-/-}$ TSC and wild-type dendritic cells (WT; **c**, **f**), or mutant dentritic cells (Mu; **d**, **e**) for 72 h in thymic reaggregates. The recovered cells from all cultures were stained with anti-CD4 or CD8 monoclonal antibodies and examined by flow cytometry. The percentage of cells with a CD4$^+$ or CD8$^+$ phenotype is given in the upper left or lower right quadrant of each panel and the absolute number of recovered viable thymocytes is given below each panel.

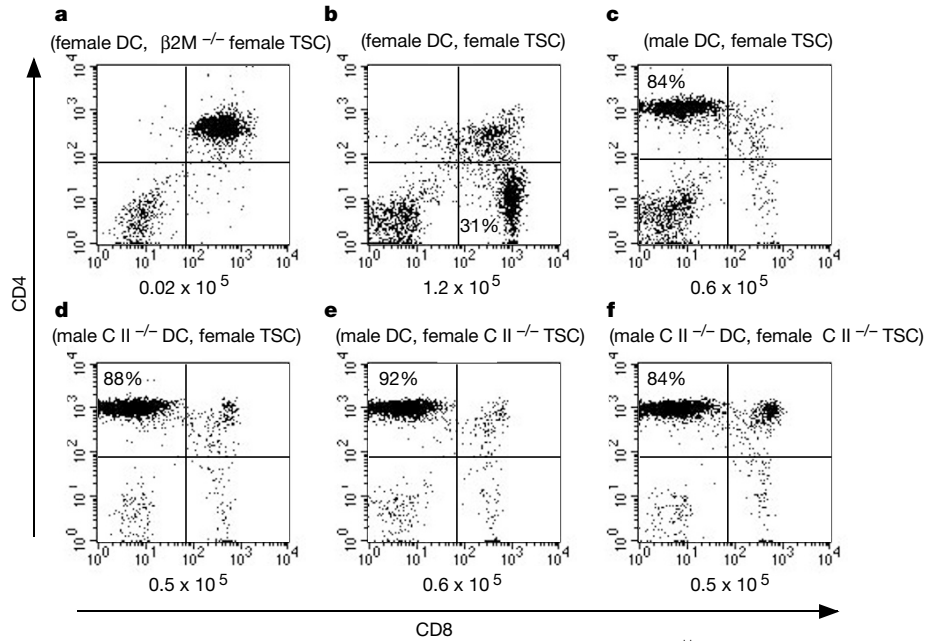


Figure 2 Potent TCR ligands can promote thymocytes with an MHC class-I-restricted TCR to become CD4$^+$ T cells. DP thymocytes from female HY TCR transgenic H-2^d RAG-2$^{-/-}$ mice were stimulated for 20 h with splenic dendritic cells (DC) from female (**a**, **b**) or male (**c**, **e**) C57BL/6 mice or splenic DC from male C57BL/6 crossed with MHC class II$^{-/-}$ mice (**d**, **f**) (culture 1). The CD69^{hi} cells generated in culture 1 were placed in reaggregate culture for 72 h with TSC from female β2M$^{-/-}$ (**a**), female C57BL/6 WT (**b–d**) or female MHC class II$^{-/-}$ (**e**, **f**) mice (culture 2). Cells were analysed as in Fig. 1.

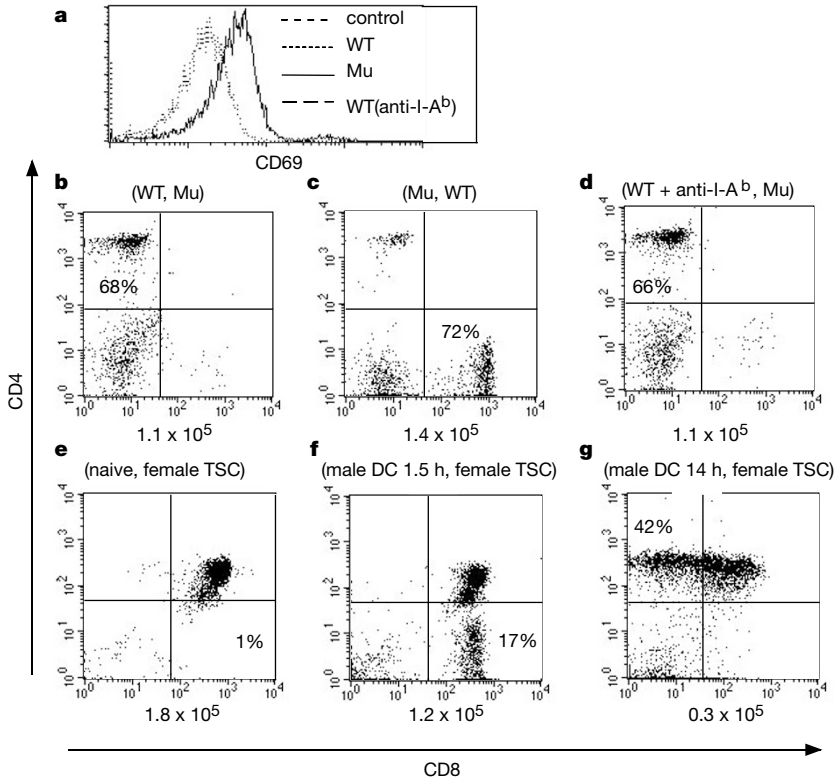


Figure 3 The duration, not the magnitude, of TCR signalling dictates lineage choice. **a**, DP thymocytes from AND mice were stimulated for 20 h with Mu-DC or WT-DC in the absence or presence of anti-I-A^b monoclonal antibodies (3JP). Purified CD69^{hi}CD4^{lo}CD8^{lo} cells after stimulation by Mu-DC (**c**), or WT-DC in the absence (**b**) or presence (**d**) of anti-I-A^b monoclonal antibodies, were then cultured with TSC from MHC$^{-/-}$ mice and WT-DC (**c**) or Mu-DC (**b**, **d**) for 60 h. DP thymocytes from female HY TCR transgenic H-2^d RAG-2$^{-/-}$ mice were stimulated with splenic dendritic cells from male C57BL/6 mice for 0 (**e**), 1.5 (**f**) or 14 (**g**) hours. The CD69^{hi} cells generated by culture were placed in reaggregate culture with TSC from female C57BL/6 mice for 60 (**e**), 58 (**f**) or 46 (**g**) hours. Cells were analysed as in Fig. 1.

either WT-DC or Mu-DC in culture 2 yielded mature T cells (Fig. 1c–f). However, the $CD69^+$DP generated using WT-DC differentiated exclusively into $CD4^+$ cells, whereas $CD69^+$DP produced using Mu-DC developed only into $CD8^+$ cells, irrespective of which dendritic cells were used in culture 2.

These findings are inconsistent with previously described stochastic/selection models for CD4/8 lineage commitment and mature T-cell development[4,12]. Such models propose that co-receptor and TCR co-recognition of MHC molecules is required during late stages of development to maintain cell viability, but our results show that CD4 function is not required for $CD4^+$ T-cell development in culture 2 (Fig. 1e). We do, however, find that MHC class I–CD8 interactions are required for generation of mature $CD8^+$ cells in culture 2 (data not shown). In addition, both wild-type and mutant MHC class II molecules can support either $CD4^+$ or $CD8^+$ T-cell development in culture 2 (Fig. 1d, f). If the initial TCR signal produced a stochastic pattern of commitment, then, unless stringent selection for survival takes place in culture 1, both $CD4^+$ and $CD8^+$ T-cell differentiation should have been seen in culture 2. This was not the case (Fig. 1e, f). Finally, selection during culture 1 based on inadequate signalling resulting from mismatches between co-receptor expression and TCR binding specificities for MHC molecules cannot explain these findings, because neither CD4 nor CD8 expression is extinguished at this time[11] (Yasutomo *et al.*, manuscript in preparation).

DP cells have a higher fraction of CD4 as compared to CD8 linked to the key src-family kinase Lck, which is involved in TCR-dependent signalling[13,14]. This asymmetry could explain why most cells with TCR that bind MHC class II ligands co-recognized by CD4 were restricted to the CD4 pathway, whereas those with TCR engaging MHC class I ligands along with CD8 were constrained to the CD8 lineage. To test this presumption, we investigated whether exposure to a potent TCR agonist that efficiently recruits CD8 (ref. 15) could overcome this biased Lck distribution and generate $CD69^+$DP expressing MHC class I-restricted TCR that would develop along the CD4 pathway. The deletion that accompanies continuous exposure to such strong ligands in typical thymic organ cultures[16] can be avoided in our model by excluding agonist ligands from culture 2 (Yasutomo *et al.*, manuscript in preparation).

HY-TCR transgenic thymocytes are deleted in male mice, owing to expression of the agonist HY male antigen presented by $H\text{-}2D^b$, whereas they are positively selected into the CD8 lineage in female $H\text{-}2^b$ mice[17]. When $CD69^+$DP generated in culture 1 by the stimulation of HY-TCR transgenic thymocytes with female $H\text{-}2^b$ dendritic cells were reaggregated in culture 2 with female $H\text{-}2^b$ TSC (Fig. 2b), only $CD8^+$ T cells were produced. Like AND thymocytes stimulated with Mu-DC, $CD8^+$ T cells did not develop from these $CD69^+$DP when TSC from female β2m-deficient mice were used (Fig. 2a). Remarkably, when we placed $CD69^{hi}$ HY-TCR transgenic cells generated by stimulation with male $H\text{-}2^b$ dendritic cells in culture 2 with TSC from female C57BL/6 mice, $CD4^+$ and not $CD8^+$ T cells developed (Fig. 2c). This development of $CD4^+$ T cells from HY-TCR transgenic thymocytes did not require recognition of MHC class II molecules in either culture 1 or 2 (Fig. 2d–f). Maturation into $CD4^+$ T cells can be completed even though CD8 expression is lost during this process, providing additional evidence against selection schemes postulating that survival of maturing thymocytes demands continuous signalling from ligand co-recognition by TCR and co-receptor.

Published models postulating that DP lineage decisions are controlled by a unique biochemical signal from Lck-associated co-receptors did not predict this generation of $CD4^+$ T cells from MHC class I-restricted TCR transgenic thymocytes using a potent agonist[18], re-opening the question of how co-receptor function influences lineage fate. We first investigated whether a change in ligand density, and hence in peak signalling intensity, controls

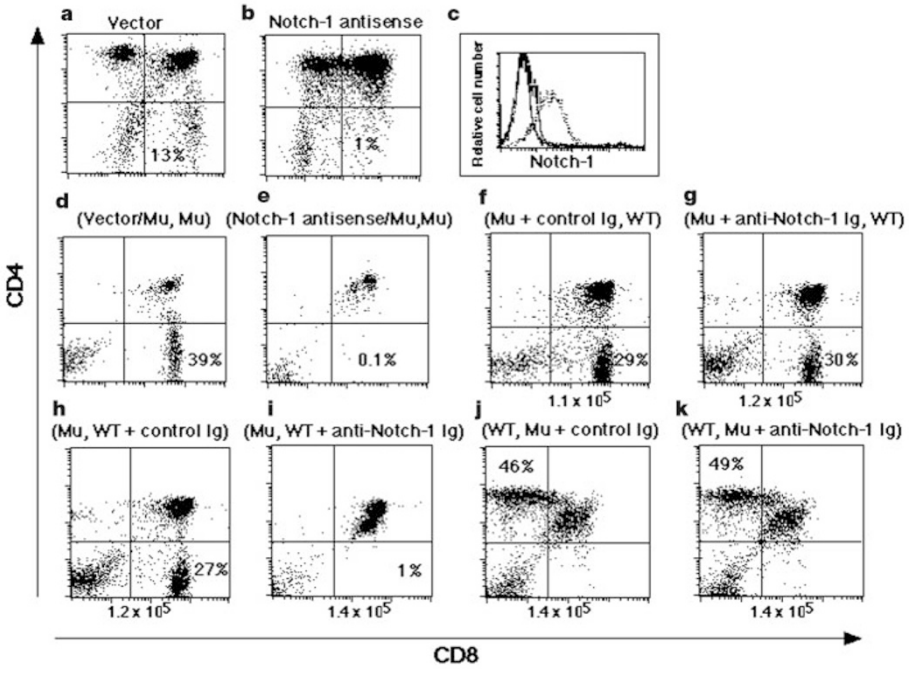


Figure 4 Notch-1 is involved in post-commitment maturation of CD8 lineage T cells. Day 14 fetal thymocytes from C57BL/6 mice (**a–c**) or AND mice (**d**, **e**) were infected with retrovirus encoding GFP (**a**, **c**, **d**) or GFP and anti-sense mouse Notch-1 (64–1164) (**b**, **c**, **e**) and cultured as described in the Methods. Cells infected by vector alone or vector encoding antisense mouse Notch-1 were stained with a combination of anti-Notch-1, anti-CD4 and anti-CD8 monoclonal antibodies (**c**) or by anti-CD4 and anti-CD8 monoclonal antibodies (**a**, **b**, **d**, **e**). Notch-1 expression in (**c**) is gated on $CD4^+CD8^-$ cells (sparse dot, vector; dense dot, antisense Notch-1; solid, control). The cells in **a**, **b**, **d**, **e** are gated on GFP^+ cells. DP thymocytes from AND mice were also stimulated for 20 h with Mu-DC in the presence of mouse IgG2b (**f**) or anti Notch-1 mAb (**g**), Mu-DC (**h**, **i**), or WT-DC (**j**, **k**). The sorted $CD69^+CD4^+CD8^+$ cells obtained after stimulation were cultured for 60 h with $MHC^{-/-}$ TSC and WT-DC (**f**, **g**), WT-DC in the presence of mouse IgG2b (**h**) or anti-Notch-1 antibodies (**i**), or Mu-DC in the presence of mouse IgG2b (**j**) or anti-Notch-1 antibodies (**k**).

lineage fixation. The interaction between the AND TCR and I-A^b was blocked with an anti-MHC class II monoclonal antibody, using a concentration of antibody that limits CD69 expression induced by WT-DC to the same level as that induced by Mu-DC without antibody (Fig. 3a). The thymocytes thus serve as their own control that this is an appropriate level of blocking, because the level of CD69 upregulation on DP has been shown to track the density of offered ligand and hence to reflect TCR occupancy[19] (K. Yasutomo, unpublished observations). Cells exposed to less wild-type signal with CD4 function intact still entered the CD4 lineage (Fig. 3d). These data indicate that the CD4 co-receptor does more than simply increase the absolute level of MHC-dependent signalling.

A previous study using pharmacologic agents indicated that the duration of active signal generation influenced the CD4/CD8 decision process[20]. We therefore examined the effects of varying the length of time for which the TCR could engage ligand. When CD69$^+$ cells generated by the stimulation of HY TCR transgenic thymocytes using male H-2^b dendritic cells for 1.5 h were placed in culture 2 with female H-2^b TSC, only CD8$^+$ T cells developed (Fig. 3f). Under the same culture conditions, CD8$^+$ T cells do not mature from naive DP thymocytes (Fig. 3e). In contrast, after 14 h of initial stimulation, only CD4$^+$ T cells appeared (Fig. 3g). This indicates that effective co-receptor activity may constrain DP development to the CD4 lineage by helping to ensure signalling over an adequate time interval, consistent with recent observations on the role of co-receptor activity in forming an 'immunological synapse' with an antigen-presenting cell that is necessary for sustained TCR signalling[21].

Signals from receptors other than those recognizing MHC molecules have been proposed to influence CD4/CD8 commitment. Notch is central in cell-fate determination in many tissues and some data indicate that activated Notch-1 may promote CD8 lineage choice[5]. Others have suggested that Notch-1 signalling simply prolongs survival of all TCR-signalled DP thymocytes[6]. Gene targeting of Notch-1 interferes with T-cell development before the DP stage, so such mutant animals or their cells cannot be used to resolve this issue[22]. The preceding evidence showing a distinction between lineage commitment and lineage progression allowed us to re-examine this issue. First, we used a retrovirus encoding Notch-1 in the antisense orientation to infect fetal thymocytes. The infected cells produced normal numbers of CD4$^+$ T cells in organ culture but CD8$^+$ T cell development was severely impaired (Fig. 4a, b). The antisense vector was highly effective, reducing Notch-1 expression on the emerging CD4$^+$ cells by more than 90% (Fig. 4c). Likewise, when the same retrovirus was used to infect AND thymocytes and these cells were stimulated by Mu-DC, very few DP became CD8$^+$ T cells, with no compensatory appearance of CD4$^+$ cells (Fig. 4d, e).

The second set of experiments employed an anti-Notch-1 monoclonal antibody. When AND CD69$^+$DP generated by culture 1 stimulation with Mu-DC were reaggregated in culture 2 with MHC class II$^{-/-}$ TSC and WT-DC, CD8$^+$ T cells developed (Fig. 4f). If anti-Notch-1 monoclonal antibodies were added to culture 1, thymocytes still became CD8$^+$ in similar numbers (Fig. 4g). However, when anti-Notch-1 monoclonal antibodies were added to culture 2, differentiation to CD8$^+$ SP (single positive) cells was very limited (Fig. 4h, i). Anti-Notch monoclonal antibodies did not interfere with the development of CD4$^+$ T cells when present at either culture step (Fig. 4j, k). Together, the antisense and antibody blocking data show that Notch-1 activity is selectively involved in CD8 but not CD4 lineage progression after lineage fate is fixed by TCR/co-receptor signalling.

These results indicate that the duration of initial TCR signalling determined by co-receptor-influenced self-MHC recognition rapidly fixes thymocyte lineage fate; lineage choice does not appear to be stochastic or cell autonomous. The simplest model that fits our data is an instructive one, with signal duration controlling which set of genes is activated or silenced so that a DP thymocyte adopts either of two possible fates. Such a model is consistent with data showing that signal duration can influence gene expression and cellular differentiation in non-lymphoid systems[23]. Alternatively, DP could consist of two subpopulations, each committed to either the CD4 or CD8 lineage. Long-lasting signals would be necessary to induce the CD4 program and, to fit our data, would also need to kill pre-CD8 cells. Short signals would activate only the pre-CD8 cells, without affecting pre-CD4 thymocytes. Without direct evidence for bipotency of immature thymocytes, this alternative cannot be dismissed.

In either case, these observations provide a specific mechanism for previous results showing that a 'strength of signal' parameter affects CD4/CD8 choice[24] and that antibody crosslinking of CD3ε with either CD4 or CD8 drives CD4 lineage development, whereas CD3ε crosslinking alone produces CD8$^+$ T cells[18]. The data also connect our demonstration that mitogen-activated protein kinase (MAPK)-mediated positive feedback enhances the duration of TCR signalling (Stefanova *et al.*, manuscript in preparation) to evidence that interference with the MAPK pathway promotes CD8 lineage choice[25,26]. The lack of a unique fate-determining biochemical signal arising from CD4 versus CD8 engagement implies that errors will be made in matching TCR class-specificity with lineage choice. Errors in the direction of CD4 commitment will probably involve very potent MHC class I ligands for the TCR that will induce late-stage deletion. Errors in the converse direction should be less frequent and cells with MHC class II-specific TCR that traverse the CD8-dependent progression steps are in any case unlikely to be dangerous to the host.

A final finding of our study is the clear distinction between lineage commitment itself and subsequent lineage-specific differentiation events such as selective silencing of co-receptor expression. This new insight allowed us to show that Notch-1 is critical only for supporting CD8 lineage progression after lineage fate is fixed. Because several Notch family members are expressed by T-cell precursors and both they and their putative ligands show complex patterns of regulation[27] and functional interactions[28], other Notch receptors may have a complementary function in the CD4 pathway to that shown here for Notch-1 in CD8 differentiation. The methods we have described provide tools for direct analysis of this possibility. □

Methods

Mice

Mice with targeted inactivation of both the β2 microglobulin gene and/or Aβ^b gene loci and HY TCR transgenic RAG-2$^{-/-}$ mice were obtained from Taconic. B10.BR and C57BL/10 mice were obtained from the Jackson Laboratories. The AND TCR transgenic[7] RAG-2$^{-/-}$ Aβ$^{-/-}$ mice were provided by B. J. Fowlkes and are referred to as AND throughout. Transgenic mice expressing a mutant I-Aβ^b chain with the E137A and V142A mutations that interfere with CD4 binding (mutant MHC class II) or expressing the corresponding WT I-Aβ^b chain, each bred to Aβ$^{-/-}$ mice, have been described[9]. All mice were bred and maintained in accordance with established guidelines.

Flow cytometry

Thymocytes were stained with PE-conjugated anti-CD4 and FITC-conjugated anti-CD8 monoclonal antibodies. All antibodies were purchased from Pharmingen. Flow cytometry was performed on a FACScan or a FACStarPlus (Becton Dickinson). List-mode data files were analysed using Cell Quest software (Becton Dickinson).

Thymocyte stimulation and preparation of CD69^{hi} cells

CD4$^+$CD8$^+$ thymocytes (1×10^6) from 1–2 week old AND mice or HY TCR transgenic RAG-2$^{-/-}$ nonselecting B10.D2 mice were mixed with dendritic cells (2×10^5) in the absence of added antigen, then cultured for 1.5 or 20 h at 37 °C. Viable CD69^{hi} cells for transfer into reaggregate cultures were obtained by fluorescence-activated cell sorting after stimulation of CD4$^+$CD8$^+$ thymocytes using the described culture conditions. In some experiments purified mouse IgG2b (20 μg ml^{-1}) (Pharmingen), 3JP monoclonal antibodies against I-A^b (0.12 μg ml^{-1}), or anti-Notch-1 monoclonal antibodies (20 μg ml^{-1}) (L.M., manuscript in preparation) were added during stimulation.

Thymic reaggregate culture

Thymic stromal cells (TSC) were prepared by disaggregating fetal thymic lobes previously

cultured for 5 days in 1.35 mM deoxyguanosine (Sigma) using 0.05% trypsin (GIBCO) and 0.02% EDTA. Reaggregates were formed by mixing together the desired TSC and thymocytes at a 1:1 cell ratio (absolute number, 5×10^5 of each cell type) or TSC, thymocytes and dendritic cells at a cell ratio of 10:10:1 (absolute number of dendritic cells, 5×10^4). After pelleting the cells by centrifugation, the cell mixture was placed as a standing drop on the upper membrane surface of a Transwell culture well containing RPMI-1640 supplemented by 10% FCS and cultured for 72 h at 37 °C. In some experiments purified mouse IgG2b (20 μg ml^{-1}) (Pharmingen) or anti-Notch-1 monoclonal antibodies (20 μg ml^{-1}) was incubated with $CD69^{hi}$ thymocytes, TSC and dendritic cells for 30 min at 4 °C before reaggregate culture, then cell mixtures were cultured in the presence of the same antibody. All results shown were repeated 2–8 times with similar findings.

Preparation of dendritic cells

Dendritic cells were purified from spleen cells from C57BL/6 or B10.BR mice as described[29]. Briefly, low-density spleen cells recovered from a BSA gradient were incubated for 30 min with anti-CD4, anti-CD8 and anti-B220 monoclonal antibodies (Pharmingen). After incubation with this antibody cocktail, the cells were washed and antibody-coated cells were removed by sheep-anti-rat IgG-coupled magnetic beads (Dynabeads, Dynal). The resultant population contained 80–85% CD11c+ cells and is referred to as dendritic cells (DC) in this study.

Retroviral constructs and retroviral transfection

The retroviral vector encoding GFP (GFP–RV)[30] was provided by K. Murphy (Washington Univ., St. Louis). Mouse Notch-1 (+64–+1164) was amplified using Pfu DNA polymerase (Stratagene) and cloned into *Xho*I digested GFP–RV vector. The nucleotide sequence and orientation of the insert were confirmed by dideoxy sequencing. The Phoenix-Eco packaging cell line (gift from G. Nolan, Stanford) was transfected according to Nolan's protocol. Day 14 fetal thymocytes (5×10^6 ml^{-1}) from C57BL/6 or AND mice were infected using 1:1 volume of viral supernatant, polybrene (Sigma) at 8 μg ml^{-1} and mouse IL-7 (20 ng ml^{-1}) (Pharmingen), centrifuged at 1,800*g* for 45 min at room temperature, and incubated at 37 °C for 48 h. The GFP^+ cells purified by fluorescence-activated cell sorting were transferred into deoxyguanosine (1.35 mM; Sigma)-treated (5 days) day 14 fetal thymi from $MHC^{-/-}$ mice in the wells of a Terasaki plate (Applied Scientific). After 1 day, the repopulated thymus lobes were then cultured at 37 °C on the upper membrane surface of a Transwell culture well containing RPMI-1640 supplemented with 10% FCS. After 7 days culture, total cells were stained with Cychrome-conjugated anti-CD4 and PE-conjugated anti-CD8 monoclonal antibodies, then $CD4^+CD8^+$ cells were purified by electronic sorting. These $CD4^+CD8^+$ cells were reaggregated with TSC from C57BL/6 mice and cultured for 72 h. In the case of thymocytes from AND mice, total cells recovered after 7 days culture in repopulated thymic lobes were stimulated with Mu-DC and $CD4^+CD8^+CD69^+$ cells were purified by electronic sorting, which were reaggregated with TSC from C57BL/6 mice and cultured for 60 h.

Received 23 September 1999; accepted 31 January 2000.

1. Robey, E. & Fowlkes, B. J. Selective events in T cell development. *Annu. Rev. Immunol.* **12,** 675–705 (1994).
2. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334,** 395–402 (1988).
3. von Boehmer, H. CD4/CD8 lineage commitment: back to instruction? *J. Exp. Med.* **183,** 713–715 (1996).
4. Chan, S., Correia-Neves, M., Benoist, C. & Mathis, D. CD4/CD8 lineage commitment: matching fate with competence. *Immunol. Rev.* **165,** 195–207 (1998).
5. Robey, E. *et al.* An activated form of Notch influences the choice between CD4 and CD8 T cell lineages. *Cell* **87,** 483–492 (1996).
6. Deftos, M. L., He, Y. W., Ojala, E. W. & Bevan, M. J. Correlating notch signaling with thymocyte maturation. *Immunity* **9,** 777–786 (1998).
7. Kaye, J. *et al.* Selective development of CD4+ T cells in transgenic mice expressing a class II MHC-restricted antigen receptor. *Nature* **341,** 746–749 (1989).
8. Konig, R., Huang, L. Y. & Germain, R. N. MHC class II interaction with CD4 mediated by a region analogous to the MHC class I binding site for CD8. *Nature* **356,** 796–798 (1992).
9. Riberdy, J. M., Mostaghel, E. & Doyle, C. Disruption of the CD4-major histocompatibility complex class II interaction blocks the development of CD4(+) T cells *in vivo*. *Proc. Natl Acad. Sci. USA* **95,** 4493–4498 (1998).
10. Anderson, G., Jenkinson, E. J., Moore, N. C. & Owen, J. J. MHC class II-positive epithelium and mesenchyme cells are both required for T-cell development in the thymus. *Nature* **362,** 70–73 (1993).
11. Lucas, B. & Germain, R. N. Unexpectedly complex regulation of CD4/CD8 coreceptor expression supports a revised model for CD4+CD8+ thymocyte differentiation. *Immunity* **5,** 461–477 (1996).
12. Davis, C. B., Killeen, N., Crooks, M. E., Raulet, D. & Littman, D. R. Evidence for a stochastic mechanism in the differentiation of mature subsets of T lymphocytes. *Cell* **73,** 237–247 (1993).
13. Veillette, A., Zuniga-Pflucker, J. C., Bolen, J. B. & Kruisbeek, A. M. Engagement of CD4 and CD8 expressed on immature thymocytes induces activation of intracellular tyrosine phosphorylation pathways. *J. Exp. Med.* **170,** 1671–1680 (1989).
14. Wiest, D. L. *et al.* Regulation of T cell receptor expression in immature CD4+CD8+ thymocytes by p56lck tyrosine kinase: basis for differential signaling by CD4 and CD8 in immature thymocytes expressing both coreceptor molecules. *J. Exp. Med.* **178,** 1701–1712 (1993).
15. Kisielow, P., Bluthmann, H., Staerz, U. D., Steinmetz, M. & von Boehmer, H. Tolerance in T-cell-receptor transgenic mice involves deletion of nonmature CD4+8+ thymocytes. *Nature* **333,** 742–746 (1988).
16. Hogquist, K. A., Jameson, S. C. & Bevan, M. J. Strong agonist ligands for the T cell receptor do not mediate positive selection of functional CD8+ T cells. *Immunity* **3,** 79–86 (1995).
17. Teh, H. S. *et al.* Thymic major histocompatibility complex antigens and the alpha beta T- cell receptor determine the CD4/CD8 phenotype of T cells. *Nature* **335,** 229–233 (1988).
18. Basson, M. A., Bommhardt, U., Cole, M. S., Tso, J. Y. & Zamoyska, R. CD3 ligation on immature thymocytes generates antagonist-like signals appropriate for CD8 lineage commitment, independently of T cell receptor specificity. *J. Exp. Med.* **187,** 1249–1260 (1998).
19. Merkenschlager, M. *et al.* How many thymocytes audition for selection? *J. Exp. Med.* **186,** 1149–1158 (1997).
20. Iwata, M., Kuwata, T., Mukai, M., Tozawa, Y. & Yokoyama, M. Differential induction of helper and killer T cells from isolated CD4+CD8+ thymocytes in suspension culture. *Eur. J. Immunol.* **26,** 2081–2086 (1996).
21. Grakoui, A. *et al.* The immunological synapse: a molecular machine controlling T cell activation. *Science* **285,** 221–227 (1999).
22. Radtke, F. *et al.* Deficient T cell fate specification in mice with an induced inactivation of Notch1. *Immunity* **10,** 547–558 (1999).
23. Marshall, C. J. Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell* **80,** 179–185 (1995).
24. Matechak, E. O., Killeen, N., Hedrick, S. M. & Fowlkes, B. J. MHC class II-specific T cells can develop in the CD8 lineage when CD4 is absent. *Immunity* **4,** 337–347 (1996).
25. Sharp, L. L., Schwarz, D. A., Bott, C. M., Marshall, C. J. & Hedrick, S. M. The influence of the MAPK pathway on T cell lineage commitment. *Immunity* **7,** 609–618 (1997).
26. Bommhardt, U., Basson, M. A., Krummrei, U. & Zamoyska, R. Activation of the extracellular signal-related kinase/mitogen-activated protein kinase pathway discriminates CD4 versus CD8 lineage commitment in the thymus. *J. Immunol.* **163,** 715–722 (1999).
27. Felli, M. P. *et al.* Expression pattern of notch1, 2 and 3 and jagged1 and 2 in lymphoid and stromal thymus components: distinct ligand-receptor interactions in intrathymic T cell development. *Intl Immunol.* **11,** 1017–1025 (1999).
28. Beatus, P., Lundkvist, J., Oberg, C. & Lendahl, U. The notch 3 intracellular domain represses notch 1-mediated activation through Hairy/Enhancer of split (HES) promoters. *Development* **126,** 3925–3935 (1999).
29. Inaba, M. *et al.* Distinct mechanisms of neonatal tolerance induced by dendritic cells and thymic B cells. *J. Exp. Med.* **173,** 549–559 (1991).
30. Ouyang, W. *et al.* Inhibition of Th1 development mediated by GATA-3 through an IL-4- independent mechanism. *Immunity* **9,** 745–755 (1998).

Acknowledgements

This work was partially support by grants from the Illinois Department of Public Health and from NIH to L.M. and American Cancer Society and the Arthritis Foundation to C.D. K.Y. was supported in part by a fellowship from the Japan Society for the Promotion of Science. We thank B.J. Fowlkes and R. H. Schwartz for their helpful comments on this manuscript, J. Delon for insightful suggestions concerning data interpretation, and M. Verma for generating the antisense Notch-1 construct.

Correspondence and requests for materials should be addressed to R.N.G. (e-mail: Ronald_Germain@nih.gov).

DNA repair protein Ku80 suppresses chromosomal aberrations and malignant transformation

Michael J. Difilippantonio*, Jie Zhu†, Hua Tang Chen†, Eric Meffre‡, Michel C. Nussenzweig‡, Edward E. Max§, Thomas Ried* & André Nussenzweig†

** Genetics Department, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*
† Experimental Immunology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
‡ Laboratory of Molecular Immunology, and Howard Hughes Medical Institute, The Rockefeller Institute, New York, New York 10021, USA
§ Laboratory of Cell Regulation, Center for Biologics Evaluation and Research, Food and Drug Administration, National Institutes of Health, Bethesda, Maryland 20892, USA

Cancer susceptibility genes have been classified into two groups: gatekeepers and caretakers[1]. Gatekeepers are genes that control cell proliferation and death, whereas caretakers are DNA repair genes whose inactivation leads to genetic instability. Abrogation of both caretaker and gatekeeper function markedly increases cancer susceptibility. Although the importance of Ku80 in DNA double-strand break repair is well established, neither Ku80 nor other components of the non-homologous end-joining pathway

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[2], yeast artificial chromosome (YAC) physical maps[3–6] and a *Not*I restriction map[7] 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 *Not*I restriction fragments[7]. 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-containing 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[8]. 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 *Not*I 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[9]. 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 α-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[10].

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%		

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)[43], the G3 radiation hybrid map (dark green; Stanford Human Genome Centre, California)[44] and two linkage maps (red; Genethon; CHLC)[45,46]. Only marker distribution is presented here: additional details, such as marker names and positions, can be found on our web sites. The *Not*I physical map of chromosome 21 was also used[7] (*Not*I 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 (RPCI-11 BAC, Keio BAC, Caltech BAC and RPCI1, 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/).

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 ETS(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, occurence of in-frame stop codons, insertions and/or deletions that disrupt the ORF or truncated matches. Generally, this was an unambiguous classification.

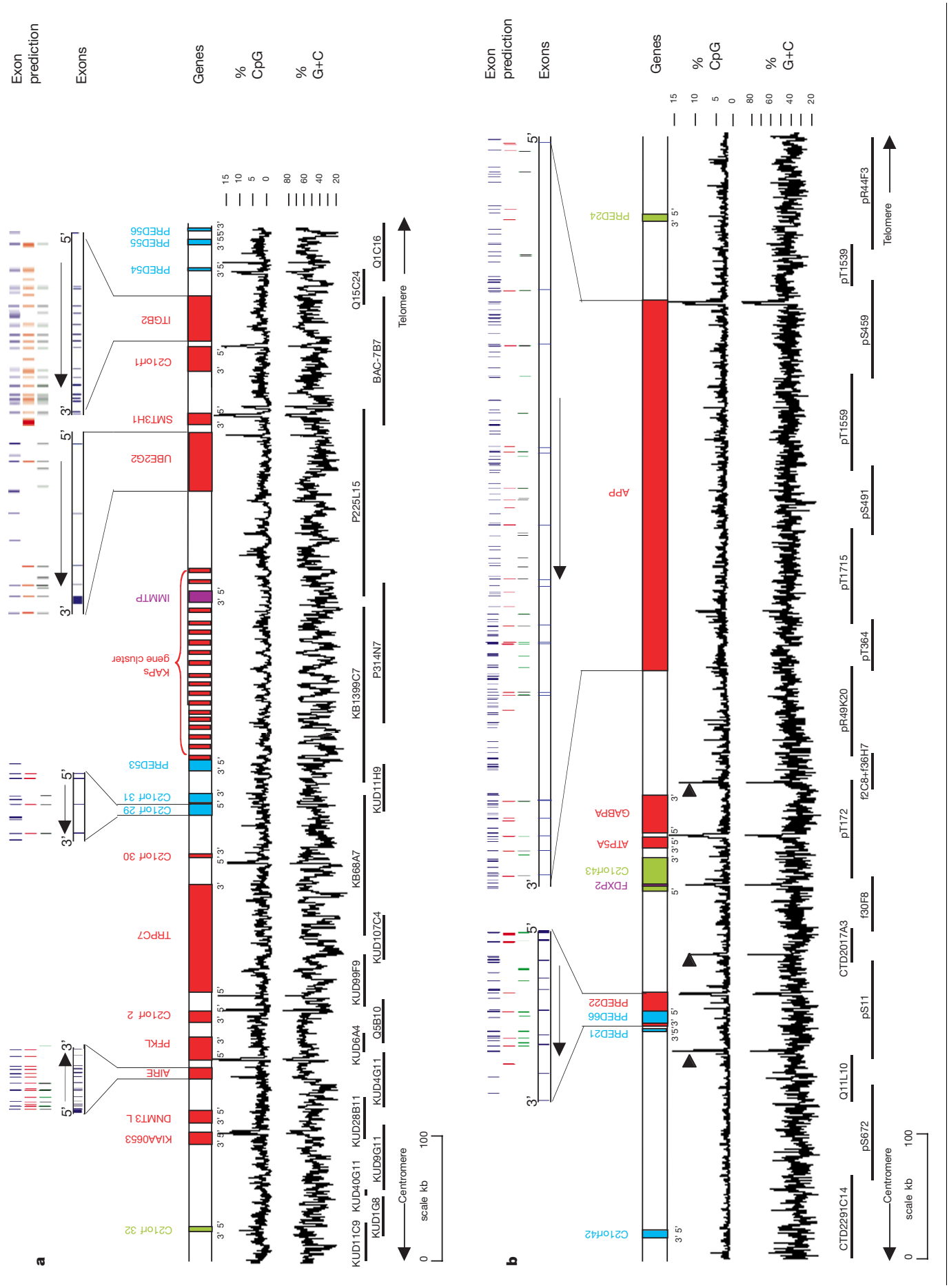
When a gene could fulfil more than one of these criteria, it was placed into the higher possible category (for example, gene prediction with spliced EST exhibiting a significant match to a known protein was placed in subcategory 2.2 rather than 4.2).

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[11]. 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.

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.

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)[13]. 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[14]. 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 intrachromosomal 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 α-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[16] 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

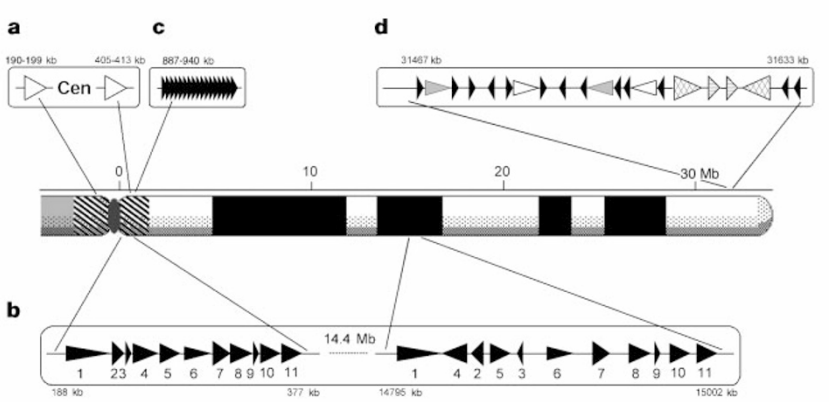


Figure 3 Schematic view of the duplicated regions in chromosome 21 as described in the text. **a**–**d**, Duplicated regions. The positions of each repeat structure are shown in kb starting at the centromere. The arrowheads represent the orientation and approximate size of each repetitive unit.

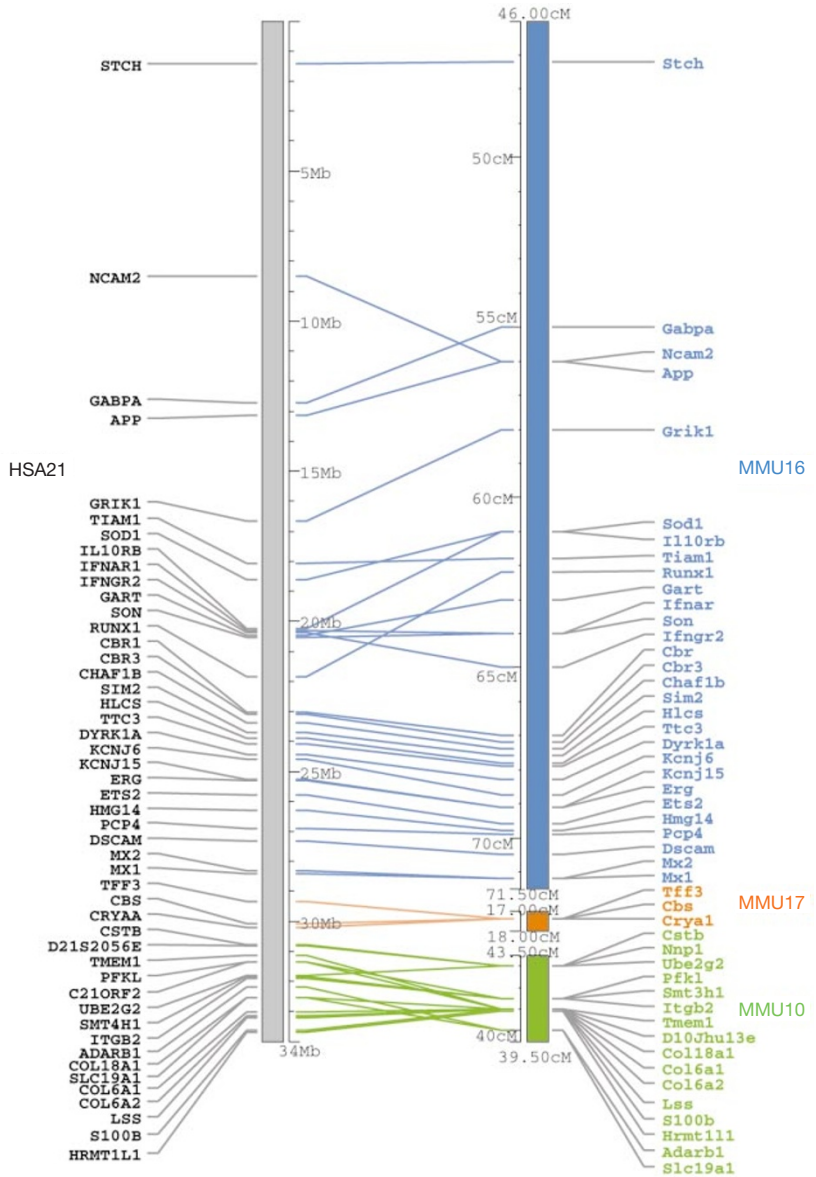


Figure 4 Schematic view of the syntenic regions between human chromosome 21 (HSA21) and mouse chromosomes 16 (MMU16), 17 (MMU17) and 10 (MMU10). Left: sequence map of human chromosome 21. Right: corresponding mouse chromosomes. Each pair of syntenic markers is joined with a line.

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[17]. 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 α-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[10]. 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[18]. 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) × 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[12]. 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[11]. 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[19]. 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[20]. 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[21,22], 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[25], Knobloch syndrome[26] 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[28] and one for familial combined hyperlipidaemia[29]) but the genes involved remain elusive.

Neoplasias. Loss of heterozygosity has been observed for specific regions of chromosome 21 in several solid tumours[30–36] 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[37–39]. On the other hand, Down syndrome patients have a markedly increased risk of childhood leukaemia[19], and trisomy of chromosome 21 in blast cells is one of the most common chromosomal aneuploidies seen in childhood leukaemias[40].

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.

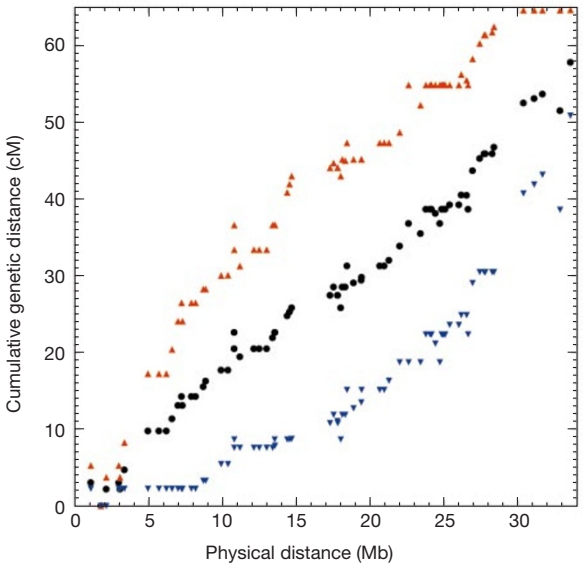


Figure 5 Comparison of the genetic map and the sequence map of chromosome 21 aligned from centromere to telomere. Genetic distance in cM; physical distance in Mb. Each spot reflects the position of a particular genetic marker retrieved from http://www.marshmed.org. Black circles, sex-average; orange 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[41]. 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 neufs 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 *Not*I 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 on 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 encephalocele (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 Chromosome 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., & Mikkolsen, 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 (1994).
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. Pohle, 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 Projekträger DLR, in the framework of the German Human Genome Project, and by the Ministry of Science, Research and Art of the Freestate of Thueringia (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 Projekträger DLR, in the framework of the German Human Genome Project. The MPIMG-Berlin group acknowledge grants from BMBF through Projekträger 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: andrex1x@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: yaspo@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.