lysozyme in the complex were assigned by using a ^{15}N-edited three-dimensional nuclear Overhauser enhancement (NOE) spectroscopy—HSQC experiment. This was collected with eight transients per increment, with 128, 28 and 512 complex points in t_1 (^{1}H), t_2 (^{15}N) and t_3 (^{1}H), and with acquisition times of 14.2, 12.3 and 48.4 ms in t_1, t_2 and t_3. The observed NOE data for bound lysozyme were interpreted by using published assignments for the free protein[29]. The resonances of D67H were assigned by comparison with the assigned spectrum of the unbound protein[11] and that of the wild-type complex.

Received 22 March; accepted 18 June 2003; doi:10.1038/nature01870.

1. Tan, S. Y. & Pepys, M. Amyloidosis. *Histopathology* **25,** 403–414 (1994).
2. Koo, E. H., Lansbury, P. T. Jr & Kelly, J. W. Amyloid diseases: Abnormal protein aggregation in neurodegeneration. *Proc. Natl Acad. Sci. USA* **96,** 9989–9990 (1999).
3. Dobson, C. M. The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. Lond. B* **356,** 133–145 (2001).
4. Pepys, M. B. *et al.* Human lysozyme gene mutations cause hereditary systemic amyloidosis. *Nature* **362,** 553–557 (1993).
5. Valleix, S. *et al.* Hereditary renal amyloidosis caused by a new variant lysozyme W64R in a French family. *Kidney Int.* **61,** 907–912 (2002).
6. Yazaki, M., Farrell, S. A. & Benson, M. D. A novel lysozyme mutation Phe57Ile associated with hereditary renal amyloidosis. *Kidney Int.* **63,** 1652–1657 (2003).
7. Dumoulin, M. *et al.* Single-domain antibody fragments with high conformational stability. *Protein Sci.* **11,** 500–515 (2002).
8. Hamers-Casterman, C. *et al.* Naturally occurring antibodies devoid of light chains. *Nature* **363,** 446–448 (1993).
9. Muyldermans, S. Single domain camel antibodies: current status. *J Biotechnol.* **74,** 277–302 (2001).
10. Booth, D. R. *et al.* Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* **385,** 787–793 (1997).
11. Canet, D. *et al.* Local cooperativity in the unfolding of an amyloidogenic variant of human lysozyme. *Nature Struct. Biol.* **9,** 308–315 (2002).
12. Schenk, D. Amyloid-beta immunotherapy for Alzheimer's disease: the end of the beginning. *Nature Rev. Neurosci.* **3,** 824–828 (2002).
13. Peretz, D. *et al.* Antibodies inhibit prion propagation and clear cell cultures of prion infectivity. *Nature* **412,** 739–743 (2001).
14. White, A. R. *et al.* Monoclonal antibodies inhibit prion replication and delay the development of prion disease. *Nature* **422,** 80–83 (2003).
15. Harper, J. D. & Lansbury, P. T. Jr Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. *Annu. Rev. Biochem.* **66,** 385–407 (1997).
16. Klunk, W. E., Pettegrew, J. W. & Abraham, D. J. Two simple methods for quantifying low-affinity dye-substrate binding. *J. Histochem. Cytochem.* **37,** 1293–1297 (1989).
17. Colon, W. & Kelly, J. W. Partial denaturation of transthyretin is sufficient for amyloid fibril formation in vitro. *Biochemistry* **31,** 8654–8660 (1992).
18. Morozova-Roche, L. A. *et al.* Amyloid fibril formation and seeding by wild-type human lysozyme and its disease-related mutational variants. *J. Struct. Biol.* **130,** 339–351 (2000).
19. Sunde, M. *et al.* Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J. Mol. Biol.* **273,** 729–739 (1997).
20. Aguzzi, A., Glatzel, M., Montrasio, F., Prinz, M. & Heppner, F. L. Interventional strategies against prion diseases. *Nature Rev. Neurosci.* **2,** 745–749 (2001).
21. Wolfe, M. S. Therapeutic strategies for Alzheimer's disease. *Nature Rev. Drug Discov.* **1,** 859–866 (2002).
22. Pepys, M. B. *et al.* Targeted pharmacological depletion of serum amyloid P component for treatment of human amyloidosis. *Nature* **417,** 254–259 (2002).
23. Dobson, C. M. Protein folding and disease: a view from the first Horizon Symposium. *Nature Rev. Drug Discov.* **2,** 154–160 (2003).
24. May, B. C. *et al.* Potent inhibition of scrapie prion replication in cultured cells by bis-acridines. *Proc. Natl Acad. Sci. USA* **100,** 3416–3421 (2003).
25. McCammon, M. G. *et al.* Screening transthyretin amyloid fibril inhibitors. Characterization of novel multiprotein, multiligand complexes by mass spectrometry. *Structure (Cambridge)* **10,** 851–863 (2002).
26. Hammarstrom, P., Wiseman, R. L., Powers, E. T. & Kelly, J. W. Prevention of transthyretin amyloid disease by changing protein misfolding energetics. *Science* **299,** 713–716 (2003).
27. Branden, C. & Tooze, J. *Introduction to Protein Structure* (Garland, New York, 1999).
28. Spencer, A. *et al.* Expression, purification, and characterization of the recombinant calcium-binding equine lysozyme secreted by the filamentous fungus *Aspergillus niger*: Comparisons with the production of hen and human lysozymes. *Protein Expr. Purif.* **16,** 171–180 (1999).
29. Ohkubo, T., Taniyama, Y. & Kikuchi, M. 1H and 15N NMR study of human lysozyme. *J. Biochem. (Tokyo)* **110,** 1022–1029 (1991).
30. Koradi, R., Billeter, M. & Wuthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14,** 51–55 (1996).

Acknowledgements The assistance of J. Zurdo, M. Krebs and B. Luisi for TEM and X-diffraction analysis of fibrils is gratefully acknowledged. We thank J.-M. Frère for many discussions. M.D. and D.C. were supported by a fellowship from the European Community. G.L. was supported by a fellowship from the Wenner-Gren Foundation. C.R. was supported by a BBSRC Advanced Research Fellowship. A.M. is a Research Associate of the FNRS, and was supported in part by a grant from the FRFC. C.V.R. is a Royal Society University Research Fellow. The research of C.M.D. is supported in part by a Programme Grant from the Wellcome Trust. This work was also supported by a BBSRC grant (to C.M.D., C.V.R. and D.B.A.) and by the Belgian Government through the PAI.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.M.D. (cmd44@cam.ac.uk).

Comparative analyses of multi-species sequences from targeted genomic regions

J. W. Thomas[1]*, J. W. Touchman[1,2]*, R. W. Blakesley[1,2], G. G. Bouffard[1,2], S. M. Beckstrom-Sternberg[1,2], E. H. Margulies[1], M. Blanchette[3], A. C. Siepel[3], P. J. Thomas[2], J. C. McDowell[2], B. Maskeri[2], N. F. Hansen[2], M. S. Schwartz[3], R. J. Weber[3], W. J. Kent[3], D. Karolchik[3], T. C. Bruen[3], R. Bevan[3], D. J. Cutler[4], S. Schwartz[5], L. Elnitski[5], J. R. Idol[1], A. B. Prasad[1], S.-Q. Lee-Lin[1], V. V. B. Maduro[1], T. J. Summers[1], M. E. Portnoy[1], N. L. Dietrich[2], N. Akhter[2], K. Ayele[2], B. Benjamin[2], K. Cariaga[2], C. P. Brinkley[2], S. Y. Brooks[2], S. Granite[2], X. Guan[2], J. Gupta[2], P. Haghighi[2], S.-L. Ho[2], M. C. Huang[2], E. Karlins[2], P. L. Laric[2], R. Legaspi[2], M. J. Lim[2], Q. L. Maduro[2], C. A. Masiello[2], S. D. Mastrian[2], J. C. McCloskey[2], R. Pearson[2], S. Stantripop[2], E. E. Tiongson[2], J. T. Tran[2], C. Tsurgeon[2], J. L. Vogt[2], M. A. Walker[2], K. D. Wetherby[2], L. S. Wiggins[2], A. C. Young[2], L.-H. Zhang[2], K. Osoegawa[6], B. Zhu[6], B. Zhao[6], C. L. Shu[6], P. J. De Jong[6], C. E. Lawrence[7], A. F. Smit[8], A. Chakravarti[4], D. Haussler[3,9], P. Green[10], W. Miller[5] & E. D. Green[1,2]

[1]*Genome Technology Branch, National Human Genome Research Institute, and* [2]*NIH Intramural Sequencing Center, National Institutes of Health, Bethesda, Maryland 20892, USA*
[3]*Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA*
[4]*Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA*
[5]*Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA*
[6]*Children's Hospital Oakland Research Institute, Oakland, California 94609, USA*
[7]*The Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12201, USA*
[8]*The Institute for Systems Biology, Seattle, Washington 98103, USA*
[9]*Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064, USA*
[10]*Howard Hughes Medical Institute and Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA*

* Present addresses: Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia 30322, USA (J.W.Th.); Translational Genomics Research Institute, Phoenix, Arizona 85004 and Department of Biology, Arizona State University, Tempe, Arizona 85287, USA (J.W.To.)

The systematic comparison of genomic sequences from different organisms represents a central focus of contemporary genome analysis. Comparative analyses of vertebrate sequences can identify coding[1–6] and conserved non-coding[4,6,7] regions, including regulatory elements[8–10], and provide insight into the forces that have rendered modern-day genomes[6]. As a complement to whole-genome sequencing efforts[3,5,6], we are sequencing and comparing targeted genomic regions in multiple, evolutionarily diverse vertebrates. Here we report the generation and analysis of over 12 megabases (Mb) of sequence from 12 species, all derived from the genomic region orthologous to a segment of about 1.8 Mb on human chromosome 7 containing ten genes, including the gene mutated in cystic fibrosis. These sequences show conservation reflecting both functional constraints and the neutral mutational events that shaped this genomic region. In particular, we identify substantial numbers of conserved non-coding segments beyond those previously identified experimentally, most of which are not detectable by pair-wise sequence comparisons alone. Analysis of transposable element insertions highlights the variation in genome dynamics among these species and confirms the placement of rodents as a sister group to the primates.

The NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program aims to sequence and to analyse targeted genomic regions in multiple vertebrates. Our initial target is a genomic segment of about 1.8 Mb on human chromosome 7q31.3

containing the gene encoding the cystic fibrosis transmembrane conductance regulator[11] and nine other genes (referred to below as the 'greater *CFTR* region'). We sought to clone and to sequence the orthologous genomic segments in multiple other vertebrates (Table 1 and Methods). So far, our efforts have yielded more than 12 Mb of high-quality comparative sequence data (see Supplementary Information), over 95% of which has been finished to the standards established for human genome sequence[12]. This represents the most diverse collection of large blocks of orthologous vertebrate sequence generated to date.

To identify regions of sequence conservation, we used blastz[13] to construct pair-wise alignments of the sequences and MultiPipMaker[14] to show pair-wise percentage-identity plots of an annotated reference sequence against multiple query sequences (Fig. 1a and Supplementary Information). Alignments between the human sequence and the sequence of each of the other 12 species allowed the general patterns of conservation to be investigated. As expected, the fraction of sequence that can be aligned generally decreases with increasing evolutionary distance from humans (Fig. 1b). The only exceptions to this trend are mouse and rat, which, although considered to be closer to humans than to the other non-primate mammals included here[15], have a lower fraction of sequence that can be aligned with the human sequence. This probably reflects a particularly high rate of sequence evolution, including large deletions, in the rodent lineage[6].

For all species, reasonably consistent alignments are seen in coding exons (Fig. 1a, b). The human–fish alignments are largely limited to these, with only 14% of the aligned sequence outside coding exons; in fact, only two local human–fish alignments do not at least partially overlap a coding exon (5′ to *CAV2* and within intron 4 of *CORTBP2*; see Supplementary Information). Almost a third (31.4%) of the human coding sequence does not align to fish sequence in the orthologous region. By contrast, the human–chicken alignments exclude less than 2% of the coding sequence, which is comparable to the alignments between human and the other mammals. Within the alignments, sequence divergence relative to human (measured as the percentage of single-nucleotide mismatches) varies from a low of 1.15% for chimpanzee to a high of 35.60% for zebrafish (see Supplementary Information).

Analyses of large-scale mutational events in this genomic region support mammalian phylogenies[15,16] that place the rodents and primates as sister groups in one clade and the artiodactyls and carnivores as sister groups in another clade (Fig. 1c). Both of these groupings are confirmed by transposon insertions that are present in sister groups and absent from the other species. In particular, we found three clear examples of insertions that support the rodent–primate grouping (all MLT1A0 elements) and three that support the artiodactyl–carnivore grouping (one MLT1A0 and two L1MA9; see Supplementary Information). In each case, insertions of the same transposon subtype, present in the same orientation and with the same target-site duplication, were identified at homologous positions in all members of sister groups.

Both MLT1A0 and L1MA9 elements are thought to have been

Table 1 **Sequence generated for the greater *CFTR* region in 12 non-human vertebrates**

Species	Non-redundant sequence (total bp)	Clone gaps*	Sequence gaps†
Chimpanzee	1,317,858	2	0
Baboon	1,508,413	0	2
Cat	1,357,338	2	3
Dog	1,195,669	4	10
Cow	1,480,745	0	0
Pig	1,077,879	1	0
Rat	1,600,751	0	3
Mouse‡	1,486,509	0	0
Chicken	415,528	0	0
Fugu§	273,624	0	0
Tetraodon	257,833	0	0
Zebrafish‖	162,514	1	0

*Number of gaps between non-overlapping BACs within the clone tiling path.
†Number of gaps within the sequenced BACs.
‡Includes ~150 kb of sequence that we generated previously (GenBank accession number AF162137).
§Includes ~36 kb of sequence generated by others (GenBank accession number AJ009961.1).
‖Includes some non-orthologous sequence (see Supplementary Information for additional details).

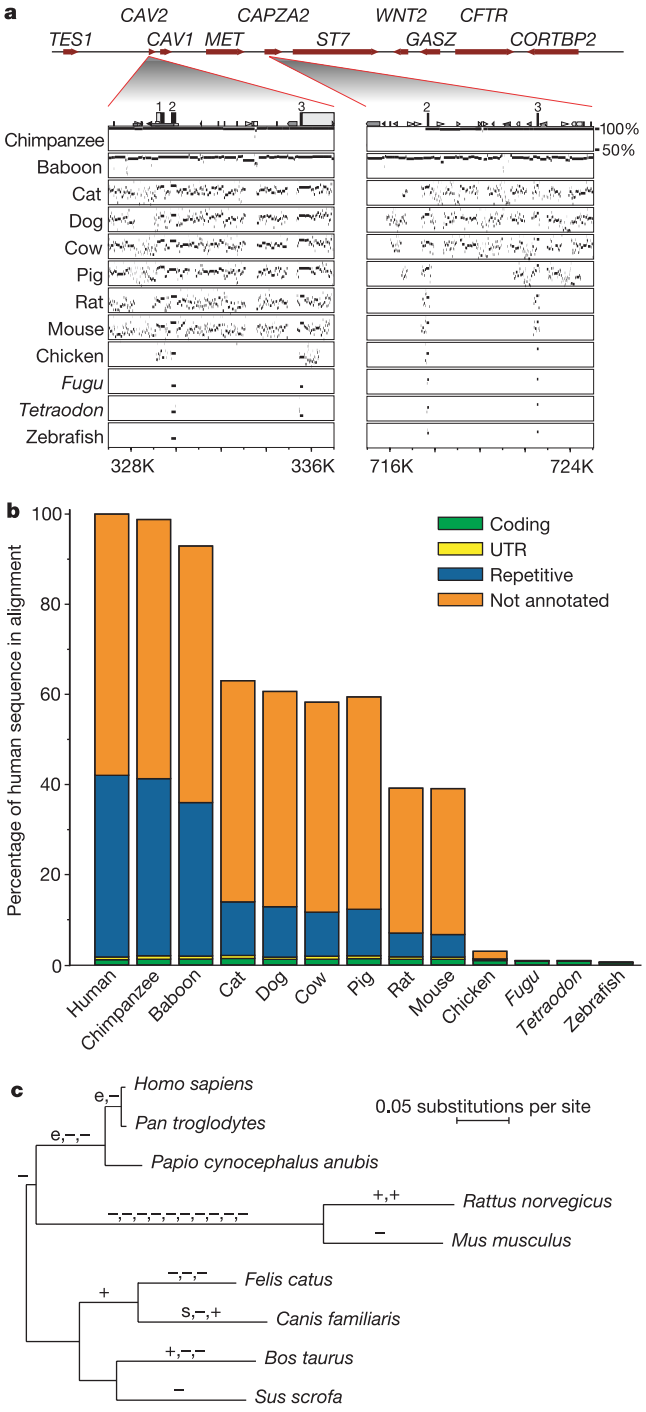

Figure 1 Patterns of sequence conservation. **a**, MultiPipMaker results for the greater *CFTR* region. Pair-wise alignments between the human reference sequence and the sequence of each indicated species were generated. The percentage identity of each gap-free alignment is indicated. Numbered boxes correspond to exons. **b**, Proportion of aligned human sequence in each of four annotated categories, shown for alignments with the sequence of each indicated species. **c**, Deduced phylogenetic tree of indicated mammalian species (see Supplementary Information). Labels on branches reflect differences in exon lengths, as determined manually by parsimony (+, insertion; −, deletion; e, extension due to alteration of splice site or stop codon; s, early stop codon).

active at around the time of the eutherian radiation, on the basis of the reconstructed phylogenies and divergence levels of the elements themselves. We found no insertions supporting alternative phylogenies (for example, rodents as an outgroup to primates, artiodactyls and carnivores). For example, out of 81 identified segments present in the primates, artiodactyls and carnivores but missing from rodents, none is an identifiable transposable element; instead, all seem to be deletions in the rodent lineage. These analyses seem to definitively refute alternative phylogenies that place the rodents as an outgroup to the other mammals studied here[16,17].

Tabulation of exon-length differences indicate a significant excess of deletions relative to insertions, which is particularly strong in the rodent lineage (consistent with previous reports for mouse genes[6]). When this excess is taken into account, the most parsimonious interpretation of the differences again favours the grouping of primates with rodents (Fig. 1c).

Neutral substitution rates estimated from sites in ancestral repeats, which are relics of transposons inserted before the eutherian radiation, also show a higher rate of substitution in the rodent lineage (Fig. 1c). The branch lengths that we estimate from ancestral repeat sites total about 1.2 substitutions per site in the mammals and are similar to previous calculations (made with these sequence data but with a different multi-sequence alignment[18]) that were based on examining untranslated regions (UTRs; total 0.93 substitutions per site), non-exonic regions (total 1.35 substitutions per site) and synonymous substitutions in codons (total 1.28 substitutions per codon).

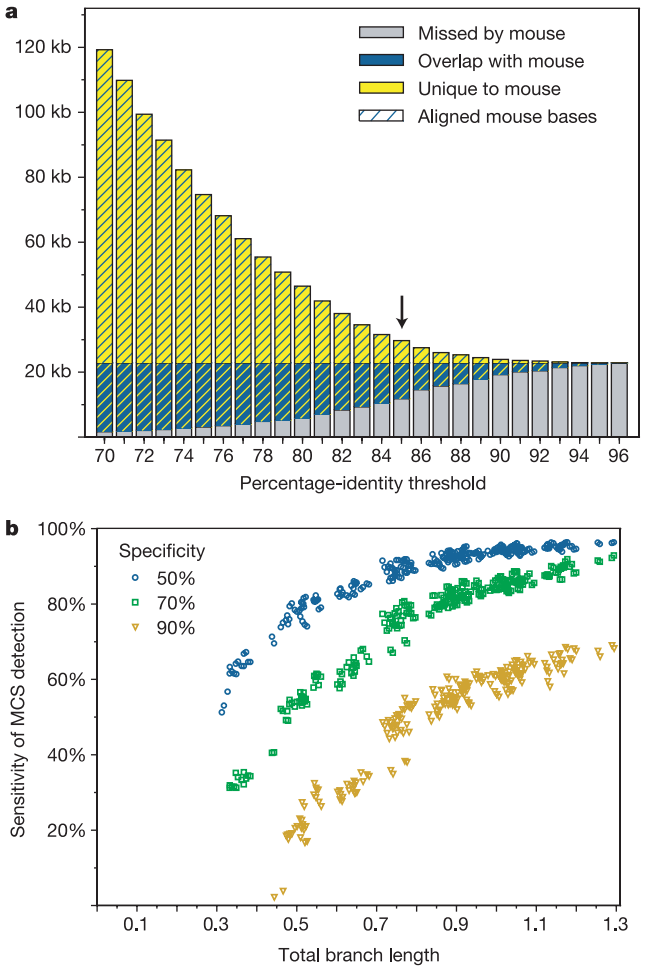

Figure 2 Detection of MCSs by using different mammalian sequences. **a**, For each indicated percentage-identity threshold (see Supplementary Information), the number of aligned mouse bases that overlap a set of 598 MCSs (see text) is shown in blue. Also shown are the number of aligned mouse bases that do not overlap MCSs (yellow) and the number of MCS bases that do not overlap aligned mouse sequence (grey). The arrow indicates the percentage-identity threshold that results in a sensitivity of 48% (with a specificity of 62%; see text). **b**, The relationship between the total branch length of the phylogenetic tree relating the species and the sensitivity of MCS detection is shown for all possible human-containing subsets of the nine mammalian sequences (see Supplementary Information). Results are plotted for three different specificities.

It is important to note, however, that there are several currently unresolved methodological issues regarding substitution analysis of non-coding genomic sequences. Alignments involving the more diverged sequences tend to have significant uncertainties, and current substitution models do not easily accommodate several known complexities in neutral mutational patterns, including context effects[19], positional variation in substitution rates[6,20] and the fact (discovered with these sequence data[21]) that there are substitution rate asymmetries associated with transcribed regions. As a result, these rate estimates must be interpreted with caution. Together, these data show that combined molecular evolution studies examining transposon events, neutral substitutions and exonic changes provide a more robust and informative phylogenetic analysis than any method alone.

Of special interest are small genomic regions that are more highly conserved across multiple species than are neutrally evolving sequences, as these may be under purifying selection (that is, selection against mutation of the base) for a functional role. By the use of two methods (E.H.M. *et al.*, manuscript in preparation, and Supplementary Information), we identified 'multi-species conserved sequences' (MCSs) across the greater *CFTR* region. Each method was calibrated such that 5% of bases in the region fell within an MCS (a value chosen to be consistent with the estimated 5% of the mammalian genome under selection based on human–mouse sequence comparisons[6]). Nearly 80% of the MCSs identified with the two methods overlap, and 75% of the MCS bases are identical.

We examined the 1,194 MCSs that overlapped between the two methods. These average 58 base pairs (bp) and represent 3.7% of the bases in the region. About 2% of MCS bases fall in ancestral repeats (corresponding to ~0.4% of the ancestral repeat sequence in the region), suggesting that the fraction of neutrally evolving sequence falsely identified as conserved is small. A total of 32% of the MCS bases overlap known coding sequence or UTRs, involving 98% (125/128) of the known coding exons and 67% (14/21; note that the 5′ UTR of the *MET* gene spans two exons) of the known UTRs. The MCSs contain 90.4% and 27.2% of coding and UTR bases, respectively. The relative paucity of UTR bases in MCSs might reflect the fact that these regions include unselected bases or are under lower selection or occasional positive selection, or a combination of these. The remaining 68% of the MCS bases are outside known exons, and virtually all of these (92%) are in MCSs that contain less than 5% repetitive sequence and are present in a single-copy fashion in the human genome (see Supplementary Information). Of the non-exonic MCSs, 16 out of 966 (1.7%; averaging 31 bp) fall within the 1-kilobase (kb) segments immediately upstream of a known transcription start site (the presumed location of most core promoters), which is about 2.3 times more frequent than would be expected if the MCS bases were randomly distributed.

Interestingly, 950 of the 1,194 MCSs (80%; averaging 44 bp) are neither exonic nor lie less than 1-kb upstream of transcribed sequence. Of these, 648 fall in introns and 302 are intergenic; this represents a 28% enrichment of MCS bases in introns (as compared with a random distribution). The detected MCSs overlap with 63% of the functionally validated regulatory elements in the region and 26% of promoters predicted by *in silico* analyses (see Supplementary Information). Several factors could account for the failure of MCSs to overlap all such elements: many of these elements are notably small (<14 bp), whereas our methods do not identify MCSs of less

than 25 bp; some of the regulatory elements may be specific to the primate lineage; and the presence or position of some of the annotated elements may be incorrect. Note that most (98%) non-exonic MCSs do not correspond to currently known regulatory elements, yielding a rich supply of candidates for future functional studies.

An important issue relevant to future genome sequencing projects is the degree to which the detection of highly conserved sequences depends on the particular set of species being studied. In the absence of a comprehensive catalogue of functional elements for any large region of the human genome, we used the detection of MCSs as a surrogate for the ability of a species' sequence to identify functionally important regions. Because a draft mouse genome sequence is now available[6], we first examined the ability of human–mouse pair-wise alignments to detect a set of 561 MCSs in a portion of the greater *CFTR* region for which sequence coverage in all species was nearly complete (see Supplementary Information). Adjustments to the stringency of the human–mouse alignments were ineffective at accurately identifying these MCSs (Fig. 2a). For example, at a percentage-identity threshold of 85%, the sensitivity (percentage of MCS bases that overlap aligned mouse sequence) is only 41%, whereas the specificity (percentage of the aligned mouse sequence that overlaps MCSs) is 77%. This is consistent with the observation that individual conserved elements often cannot be reliably detected with only the human and mouse sequences[6].

To explore more broadly how MCS detection is dependent on the specific sequences used in the analysis, we identified MCSs with each possible subset of species (always including human) and in each case calibrated the results to yield a defined specificity (percentage of bases overlapping the above 561 MCSs; see Supplementary Information). For the various subsets of mammalian species, there is strong correlation between total divergence of the subset (as measured by the combined branch length of the phylogenetic tree defined by the specific subset of species) and the ability to detect MCSs (Fig. 2b). The results at 90% specificity show that eliminating chimpanzee and baboon does not reduce the number of detected MCS bases. Eliminating the non-human primates, chicken and fish—thereby retaining only the six non-primate mammals in addition to human—reduces the number by 17%. With only one mammal from each major lineage (baboon, cat, cow and mouse), the number is reduced by 29%. Of note, chicken sequence alone detects 40% of MCS bases (representing 94% of the coding but only 29% of the non-coding MCS bases), and this value is higher than for any other single species. Fish sequences alone are effective at detecting coding exons but miss most non-coding MCSs.

The trends in Fig. 2b suggest that most MCSs are broadly distributed among the mammalian lineages, because the power to detect them seems to depend mainly on the total divergence of the subset of species rather than on the particular distribution of the species among lineages. In addition, the fact that at high specificity the sensitivity increases steadily throughout the full range of branch lengths suggests that we may be far from saturating the ability to detect such conserved sequences, even with the full set of mammals examined here. It thus seems that combined branch length will be a useful metric for guiding the selection of additional genomes to sequence.

The nature of the mutational events that have produced the observed differences in the greater *CFTR* region among the sequenced species is of fundamental evolutionary interest. Despite strict conservation in gene order and content and the absence of any observed syntenic breaks, there is significant variation in the amount of non-coding sequence, suggesting variation in genome expansion or compression. For example, the region is about tenfold smaller in the two pufferfish species than in the mammals, which themselves vary by as much as ~15% (see Supplementary Information). These findings are consistent with the relative genome sizes established for some of these species[3,5,6] and point to significant changes in this genomic region throughout vertebrate evolution. To account for these differences, we looked for evidence of molecular events that have contributed to genome expansion or compression.

The fraction of sequence corresponding to interspersed repeats, which are remnants of insertion events mediated by active or extinct transposons, varies from roughly 29% to 39% among the mammals (Fig. 3a; see Supplementary Information). The interspersed repeat content is much lower in the non-mammalian vertebrates, with the pufferfish and zebrafish sequences containing less than 1% and 13% interspersed repeats, respectively. In addition, the distribution of interspersed repeat types differs greatly among species (Fig. 3a).

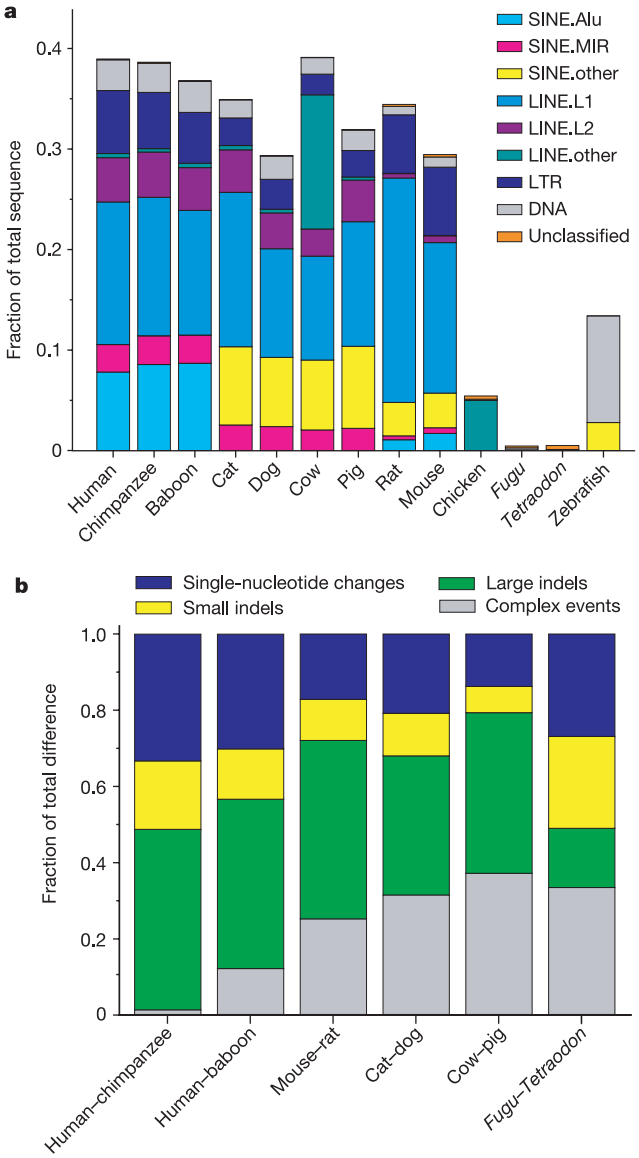

Figure 3 Comparison of genome dynamics among species. **a**, Relative content of different interspersed repeat types plotted for each species. **b**, Relative contribution of different mutational events within major lineages. Sequence alignments between the indicated species pairs were used to classify all sequence differences: single-nucleotide changes, small indels (<100 bp), large indels (>100 bp) and complex events (involving >100 bp and most probably resulting from an inversion or more than one deletion and/or insertion; see Supplementary Information). The relative fraction of nucleotide differences falling into each class is shown for the indicated species pairs. The overall percentage of single-nucleotide mismatches for the greater *CFTR* region in each species pair is 1.15%, human–chimpanzee; 5.60%, human–baboon; 13.93%, mouse–rat; 15.57%, cat–dog; 17.52%, cow–pig; and 20.50%, *Fugu–Tetraodon*. SINE, short interspersed nuclear element; LINE, long interspersed nuclear element.

Thus, the accumulation of interspersed repeats correlates with expansion of the greater *CFTR* region in the main lineages.

Differences in the pattern of interspersed repeat types are also apparent between and within mammalian lineages. Indeed, the accumulation of species-specific interspersed repeats seemingly accounts for much of the differences in relative size and interspersed repeat composition of this genomic region within these lineages (see Supplementary Information). Our findings implicate an increased rate of interspersed repeat insertions rather than variation in the extent of deletions as a primary cause of the observed size differences for this region; this contrasts with the observed higher rate of deletion versus insertion in the mouse lineage, which has resulted in a significantly smaller size of the mouse genome relative to human[6].

For the three primate species, we estimated the rate of large (>100 bp) insertions or deletions (indels) using parsimony (see Supplementary Information). This revealed variable rates of insertion and deletion, producing a slight relative expansion in the human lineage and a slight compression in both the chimpanzee and baboon lineages. We also examined the relative importance of interspersed repeat insertions and large indels compared with small indels (<100 bp) and single-nucleotide changes. Although most mutational events leading to human–chimpanzee differences are single-nucleotide changes, they account for only 33% of the bases that differ between these two species (Fig. 3b and Supplementary Information). Indeed, nearly half of the differing bases correspond to large indels (Fig. 3b; as an example, note the large deletion in the chimpanzee sequence immediately upstream of *CAPZA2* exon 2 in Fig. 1a). Similarly, at least 44% of the bases that differ between the human and baboon sequences reflect large indels. Thus, among the primates, large indels are the principal mechanism accounting for the observed sequence differences, a finding that is consistent with other studies[22,23].

Similar analyses of the other mammals identified a similar spectrum of mutational events, with large indels accounting for the largest fraction of bases that differ in each lineage (Fig. 3b). This suggests that the non-aligning regions in mammalian sequence primarily reflect the insertion of repetitive elements and the deletion of ancestral sequences. Given this assumption, the alignment of the human greater *CFTR* region to the other mammalian sequences suggests that a minimum of 63% of this segment of the human genome was present in the last common ancestor of the mammals sampled here, some 94 million years ago[24], and that the rodents represent the most derived mammalian sequence (minimum 39% ancestral; Fig. 1b). Finally, the pufferfish sequences show a distinct pattern of genomic changes (Fig. 3b). As compared with the mammals, single-nucleotide mismatches and small indels are more prominent than large indels. In part, this probably reflects the paucity of transposon insertions in pufferfish and the deleterious consequences of large indels in a more compact genome.

In summary, our approach for multi-species sequence generation and analysis is providing a previously unavailable glimpse through the window of vertebrate evolution. Our findings confirm that the genomes of rodents are changing faster than those of primates, carnivores and artiodactyls, in agreement with many reports[6,25] but in contrast to studies defending the molecular clock[26]. In addition, we have found a substantial number of sequence elements conserved across multiple species, most of which are located in non-coding regions. Although the general mutational spectrum is similar among all vertebrate lineages, differences in the relative contribution of the various types of molecular events have sculpted the genome of each species in a unique fashion. It should be noted that the findings reported here pertain to a single genomic region, whereas the vertebrate genome is known to show highly non-uniform patterns of sequence conservation[6,20]. Our efforts to sequence many other genomic targets in multiple species (see the NISC Comparative Sequencing Program website: http://www.nisc.nih.gov) should provide a broader and more informed view of such regional variation. Together, our studies point to myriad avenues for investigating the evolutionary and functional features of the vertebrate genome. □

Methods

Sequence generation and analysis

Targeted genomic regions of interest were isolated in overlapping bacterial artificial chromosome (BAC) clones[27] by using libraries derived from different vertebrate species. We used 'universal' hybridization probes, designed from small stretches of sequence conserved between human and mouse[28], to isolate BACs from multiple mammals in parallel. For isolating BACs from non-mammalian vertebrates, species-specific probes (typically designed from available gene sequences) were mostly used for clone isolation. For each species, a minimally overlapping tiling path of BACs spanning the genomic target was selected, and individual BACs were sequenced by a conventional shotgun sequencing strategy (see Supplementary Information).

To facilitate computational analyses, we compiled a single, non-redundant nucleotide sequence for each species by merging together the data generated for all sequenced BACs. This was then annotated to indicate the locations of known genes and coding regions on the basis of matches to human and/or mouse reference cDNA sequences, exons predicted by Genscan[29], and repetitive elements using RepeatMasker. Annotated sequence records are available at the NISC Comparative Sequencing Program website (http://www.nisc.nih.gov/data). A more detailed description of this assembly and annotation process is given in the Supplementary Information.

The assembled and annotated sequences were then subjected to a series of analyses, starting with the generation of multi-sequence alignments and including the studies described above and in Figs 1–3 (such as establishing orthology, examining the general patterns of sequence conservation, performing phylogenetic analyses, detecting highly conserved sequences and investigating genome dynamics). Details of these analyses are given in the Supplementary Information.

Visualization and dissemination of results

The establishment of robust and user-friendly computational-based approaches for capturing, visualizing and disseminating multi-species sequences and the data emanating from their comparative analyses represents an increasingly important challenge. By using our data as a model, we developed a viable solution to this problem by incorporating our data into the University of California Santa Cruz Genome Browser[30]. The resulting web-based resource (http://genome.ucsc.edu) serves as an additional electronic supplement to this paper; a low-resolution overview of this website is given in the Supplementary Information.

Integrating our data into the UCSC Genome Browser allows it to be visualized within the context of the growing body of information about the human and other genome sequences represented on this browser. We have configured the browser to display custom tracks showing the results of various analyses, including those involving multiple sequences (see Supplementary Information). The dynamic nature of the browser allows convenient navigation from a detected region of conservation to the underlying sequence alignment, as well as the ability to examine the results of comparative analyses using different species' sequence as the reference.

Availability of data and analysis tools

Details about the data underlying the studies reported here (including updated BAC contig maps, information about each sequenced clone, compiled and annotated sequence files for each species and links to the various electronic resources) are available on the NISC Comparative Sequencing Program website (http://www.nisc.nih.gov/data). The blastz program and MultiPipMaker network services can be obtained on the Penn State Bioinformatics Group website (http://bio.cse.psu.edu).

Received 11 April; accepted 16 June 2003; doi:10.1038/nature01858.

1. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10,** 950–958 (2000).
2. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25,** 235–238 (2000).
3. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
4. Chen, R., Bouck, J. B., Weinstock, G. M. & Gibbs, R. A. Comparing vertebrate whole-genome shotgun reads to the human genome. *Genome Res.* **11,** 1807–1816 (2001).
5. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).
6. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
7. Dubchak, I. *et al.* Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10,** 1304–1306 (2000).
8. Gottgens, B. *et al.* Analysis of vertebrate *SCL* loci identifies conserved enhancers. *Nature Biotechnol.* **18,** 181–186 (2000).
9. Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16,** 369–372 (2000).
10. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2,** 100–109 (2001).
11. Rommens, J. M. *et al.* Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245,** 1059–1065 (1989).
12. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from The Human Genome Project. *Genome Res.* **9,** 1–4 (1999).

13. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res* **13,** 103–107 (2003).
14. Schwartz, S. *et al.* MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31,** 3518–3524 (2003).
15. Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294,** 2348–2351 (2001).
16. Poux, C., Van Rheede, T., Madsen, O. & de Jong, W. W. Sequence gaps join mice and men: phylogenetic evidence from deletions in two proteins. *Mol. Biol. Evol.* **19,** 2035–2037 (2002).
17. Huelsenbeck, J. P., Larget, B. & Swofford, D. A compound Poisson process for relaxing the molecular clock. *Genetics* **154,** 1879–1892 (2000).
18. Cooper, G. M. *et al.* Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13,** 813–820 (2003).
19. Siepel, A. & Haussler, D. *Proc. 7th Annual Int. Conf. Research in Computational Molecular Biology* (ACM, New York, 2003).
20. Hardison, R. C. *et al.* Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13,** 13–26 (2003).
21. Green, P. *et al.* Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genet.* **33,** 514–517 (2003).
22. Frazer, K. A. *et al.* Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* **13,** 341–346 (2003).
23. Britten, R. J. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl Acad. Sci. USA* **99,** 13633–13635 (2002).
24. Springer, M. S., Murphy, W. J., Eizirik, E. & O'Brien, S. J. Placental mammal diversification and the Cretaceous/Tertiary boundary. *Proc. Natl Acad. Sci. USA* **100,** 1056–1061 (2003).
25. Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H. & Hewett-Emmett, D. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.* **5,** 182–187 (1996).
26. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99,** 803–808 (2002).
27. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89,** 8794–8797 (1992).
28. Thomas, J. W. *et al.* Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res.* **12,** 1277–1285 (2002).
29. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
30. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank J. Weissenbach and H. Roest Crollius for *Tetraodon* BACs; M. Diekhans for computational expertise; N. Goldman and Z. Yang for advice on phylogenetic analyses; and F. Collins and J. Mullikin for critically reading the manuscript. We acknowledge the support of the National Human Genome Research Institute (National Institutes of Health) and the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.D.G. (egreen@nhgri.nih.gov). GenBank accession numbers for BAC-derived sequences are provided in the Supplementary Information.

CYLD is a deubiquitinating enzyme that negatively regulates NF-κB activation by TNFR family members

Eirini Trompouki[1], Eudoxia Hatzivassiliou[1]*, Theodore Tsichritzis[1]*, Hannah Farmer[2], Alan Ashworth[2] & George Mosialos[1]

[1]*Institute of Immunology, Biomedical Sciences Research Center 'Alexander Fleming', 34 Alexander Fleming Street, Vari 16672, Greece*
[2]*Cancer Research UK Gene Function & Regulation Group, The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK*

* These authors contributed equally to this work

Familial cylindromatosis is an autosomal dominant predisposition to tumours of skin appendages called cylindromas. Familial cylindromatosis is caused by mutations in a gene encoding the CYLD protein of previously unknown function[1]. Here we show that CYLD is a deubiquitinating enzyme that negatively regulates activation of the transcription factor NF-κB by specific tumour-necrosis factor receptors (TNFRs). Loss of the deubiquitinating activity of CYLD correlates with tumorigenesis. CYLD inhibits activation of NF-κB by the TNFR family members CD40, XEDAR and EDAR in a manner that depends on the deubiquitinating activity of CYLD. Downregulation of CYLD by RNA-mediated interference augments both basal and CD40-mediated activation of NF-κB. The inhibition of NF-κB activation by CYLD is mediated, at least in part, by the deubiquitination and inactivation of TNFR-associated factor 2 (TRAF2) and, to a lesser extent, TRAF6. These results indicate that CYLD is a negative regulator of the cytokine-mediated activation of NF-κB that is required for appropriate cellular homeostasis of skin appendages.

Activation of NF-κB by cytokines depends on the activity of the IκB kinase (IKK) complex, which mediates phosphorylation of the NF-κB inhibitor IκB[2]. NEMO (also known as IKK-γ) is an essential component of the IKK complex. IKK-mediated phosphorylation of IκB causes its degradation by the ubiquitin–proteasome pathway and results in nuclear translocation and activation of NF-κB.

A yeast two-hybrid screen with NEMO as the bait identified complementary DNAs encoding either the alternatively spliced, 'full-length' CYLD of 953 amino acids or two truncated forms of the protein spanning amino acids 211–709 and 387–953. The interaction of CYLD with NEMO was specific, because full-length CYLD did not interact with two mutated forms of NEMO (C417R and D406V) that have an altered carboxy-terminal zinc-finger motif[3]. To determine whether this association occurs in mammalian cells, Flag-tagged CYLD and glutathione *S*-transferase (GST)-tagged NEMO were coexpressed in human embryonic kidney (HEK) 293T cells. Co-precipitation experiments indicated that CYLD and NEMO can indeed interact in mammalian cells after their coexpression (Fig. 1a, b). The amino-terminal region of CYLD mediates the interaction with NEMO, because deletion of the N-terminal 537 amino acids of CYLD abolished NEMO binding (Fig. 1c). We could not detect an interaction between endogenous CYLD and NEMO, either because of the possible transient nature of this interaction or because of the relatively low affinity of our antisera against CYLD.

The C-terminal 365 amino acids of CYLD contain motifs found in ubiquitin-specific proteases (UBPs), a subclass of deubiquitinating enzymes thought to be responsible for the removal of polyubiquitin chains from polypeptides[1,4]. Most pathogenic inactivating mutations of CYLD result in truncations or frameshift alterations of the C-terminal region of the molecule[1]. To determine whether CYLD can act as a UBP, Flag-tagged wild-type CYLD and three C-terminally truncated mutants of CYLD that are associated with familial cylindromatosis (encoding residues 1–932, 1–864 and 1–754 of CYLD; see Online Mendelian Inheritance in Man number 132700 at http://ncbi.nlm.gov/Omim/) were expressed in HEK 293T cells, immunoprecipitated and tested for their ability to cleave tetraubiquitin. Wild-type CYLD readily cleaved the tetrameric substrate to its monomer, dimer and trimer; however, the pathogenic mutations associated with familial cylindromatosis abolished the UBP activity of CYLD, indicating a correlation between tumorigenesis and loss of the deubiquitinating activity of CYLD (Fig. 1d, top). The deubiquitinating activity of CYLD did not require any other mammalian proteins, because it was detectable in bacteria (Supplementary Fig. 1). In addition, the deubiquitinating activity of CYLD was abolished by replacing the conserved catalytic residue Cys 601 with serine.

The interaction of CYLD with NEMO prompted us to investigate the potential involvement of CYLD in NF-κB activation. CD40, XEDAR and EDAR are members of the TNFR family implicated in the proper development of skin appendages from which cylindromas arise[3,5,6]. Expression of wild-type CYLD in HEK 293T cells inhibited the ability of coexpressed CD40 (Fig. 2a), XEDAR or EDAR (Supplementary Fig. 2) to activate the NF-κB pathway; by