Functional annotation of a full-length mouse cDNA collection

The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium*

* A full list of authors appears at the end of the paper

....

The RIKEN Mouse Gene Encyclopaedia Project, a systematic approach to determining the full coding potential of the mouse genome, involves collection and sequencing of full-length complementary DNAs and physical mapping of the corresponding genes to the mouse genome. We organized an international functional annotation meeting (FANTOM) to annotate the first 21,076 cDNAs to be analysed in this project. Here we describe the first RIKEN clone collection, which is one of the largest described for any organism. Analysis of these cDNAs extends known gene families and identifies new ones.

In mammals and higher plants, interpreting the genome sequence is not straightforward: coding regions are interspersed with noncoding DNA, and an individual gene may give rise to many gene products. Thus, genomic sequence cannot be reliably decoded to identify the spectrum of messenger RNAs (the transcriptome) and their corresponding protein products (the proteome). This problem is illustrated by the different estimates of the number of human genes (30,000, 35,000 and 120,000)[1–3]. Although gene prediction programs have become more accurate and sensitive, the sequence of a full-length cDNA clone provides more reliable evidence for the existence and structure of a gene. The Mouse Gene Encyclopaedia Project aims to identify and sequence every transcript encoded by the mouse genome. Here, we report the characterization of our first cDNA set of 21,076 mouse clones (some of which are derived from the same transcripts).

Strategies

In the first phase of the project, we prepared around 160 full-length enriched[4,5], normalized and subtracted[6] cDNA libraries from various tissues and developmental stages. From these, we collected and clustered 930,000 3′ end sequences to produce about 128,600 groups that were targeted for sequencing.

In the second phase of the project, we selected a single clone from each cluster for sequencing. Preference was given to clones from libraries estimated to contain the highest representation of full-length transcripts. To expedite sequencing, we focused on relatively short cDNAs (Fig. 1), which are probably biased in favour of 5′ truncated clones. To increase the likelihood of discovering new genes, we also biased our selection towards clones with novel 3′ end sequences. We sequenced 21,076 cDNA clones, with average length 1,257 base pairs (bp); the longest clone sequenced was 6,327 bp (see Supplementary Information Fig. 1A). All sequences have been registered in the public sequence database DDBJ, except for 1,908 cDNAs assembled using sequences from public expressed sequence tag (EST) databases (available at http://genome.gsc.riken.go.jp/genome/fantom/viewer/est/). We estimated using the PHRED base-calling program[7,8] that the average accuracy of our sequences was 99.1%; 72% (15,236 clones) of clones showed > 99% accuracy and 6,739 sequences (32%) were determined at > 99.9% accuracy (see Supplementary Information Fig. 1B).

We extracted the open reading frame (ORF) of each full-length sequence using the RIKEN DECODER program (see Methods). DECODER corrected frame-shifts in 3,376 (15%) of 21,076 clones. The likelihood that the sequence selected was correct is given by a score (Va) calculated in light of the Kozak consensus, preferred codon usage and position of the initiation codon. The probability that a frame shift occurred was determined using quality values (PHRED scores).

Annotation of cDNAs

An international meeting was held to facilitate functional annotation of the cDNA sequences. Participants contributed to the development of a web-based annotation interface that should expedite future annotation of additional clones in the Mouse Gene Encyclopaedia project. We agreed on annotation vocabularies and the application of Gene Ontology (GO) terms (http://genome.gsc.riken.go.jp/FANTOM/). Before the FANTOM meeting, a set of RIKEN clones with significant similarity to mouse genes represented in the databases of Mouse Genome Informatics (MGI) was annotated; 4,248 RIKEN clones were found to be identical by human curation to mouse genes in MGI (referred to as the MGI-confirmed set).

There was significant redundancy in the cDNA set. Duplication may have resulted from a number of factors including mistakes made when samples were regridded, internal initiation of reverse transcription, incomplete or variable splicing and differences in polyadenylation site usage, which may account for about 19% of true 3′ end variability[9]. To cluster redundant clones, we compared all the sequences pairwise using FLAST, a sequence comparison program based on DDS[10], and grouped them on the basis of sequence similarity. To assess cluster fidelity, sequences were assembled using CAP3[11] and aligned using CLUSTALW[12], and visually inspected. This placed 8,207 clones into 2,957 clusters, reducing the size of the cDNA clone set to 15,826 unique genes and the MGI-confirmed set to 2,921 unique genes. Further analysis of RIKEN clones in the MGI-confirmed set revealed some instances where non-overlapping clones could be added to existing clusters or grouped together on the basis of curatorial association with the same MGI gene. Therefore, the actual number of genes in the MGI set was reduced from 2,921 to 2,390, and the number of genes represented by the whole RIKEN set was reduced to 15,295. This is an overestimate of the total gene number in the RIKEN set, as we expect a similar compression to occur for clones outside the MGI set with further cluster-orientated analyses that consider external data sets. On the basis of the observed redundancy in the MGI set (roughly 20%), we have estimated the number of genes in the non-MGI set to be at least 10,500. Therefore, there should around 12,890 unique genes in the complete collection.

Redundant clones were not eliminated from the RIKEN database, because many clusters contain genuine alternate transcripts from single genes (see below). All nonredundant clones were annotated; for clusters, a single sequence was annotated and the annotation extended to other clones within a given cluster. The number of genes in each category ('MGI-confirmed', 'identical to', 'similar to' and so on) is shown in Table 1.

We functionally classified cDNAs by assigning GO terms[13] (see Methods). We assigned one or more GO terms to 3,025 of the

 685

9,902 RIKEN clones with definitive coding potential (Table 2). The putative functions of the clones are well distributed among the major categories. (http://www.gsc.riken.go.jp/genome/ fantom/ viewer/).

Analysis of length of cDNAs

Three approaches were used to assess to what extent the clones in the RIKEN cDNA collection were 'full-length'. (1) Clones in the MGI-confirmed set were compared with other published cDNAs for the same genes containing complete coding sequences (CDS), to determine whether the clones span the entire coding sequence; (2) the fraction of computationally predicted CDSs among all clones was calculated; (3) and the fraction of clones annotated as full-length by curators was determined. These analyses gave reasonably similar estimates of the percentage of full-length clones in the

collection: 63%, 53% and 59%, respectively. As the sizes of clone inserts in the categories designated 'motif-containing proteins' and 'hypothetical proteins' were similar to the sizes of clone inserts in the 'MGI-confirmed', 'identical to', 'homologue to', 'similar to' and 'related to' categories (see Supplementary Information Table 2A), we believe that 60–70% of the cDNAs encoding motif-containing and hypothetical proteins (potentially the most novel ones in the RIKEN set) will also be full-length. This validates the cap-trapping method and shows that our clones will be a valuable resource for studying transcription regulatory elements in the 5′ untranslated regions (UTRs; see Supplementary Information Table 5B).

A significant number of RIKEN cDNAs are shorter than published cDNAs encoding known genes (see Supplementary Information Table 2A). Some are likely to be truncated and unspliced forms, although alternative transcripts generated from genuine functional promoters, and transcripts generated from cryptic internal promoters, may be other sources of such clones.

Alternative splicing leading to exon skipping, extension, deletion or truncation increases the complexity of gene expression products and the proteome. One study[14] indicated that 22% of human genes may be alternatively spliced. An EST-based analysis of 475 disease-associated genes suggested that one in three genes exhibits alternative splicing[15]. This is also evident in the RIKEN cDNA set, of which 220 display potential alternative splicing (see Supplementary Information Table 2B). Furthermore, about 6% of the MGI clones are probably splice variants of known genes.

Ninety-five clones (1.2%) were in the reverse orientation. We confirmed that the inserts were probably inserted into the vector correctly; whether these clones represent antisense transcripts and are important physiologically remains to be determined.

cDNAs representing metabolic enzymes

Over 100 clones representing newly identified genes in mouse were assigned to various metabolic pathways. This was achieved by converting the GO numbers assigned to clones to EC numbers that designate enzymes (see https://genomec.gsc.RIKEN.go.jp/genome/fantom/bono/pathway/index.cgi?org = mmu).
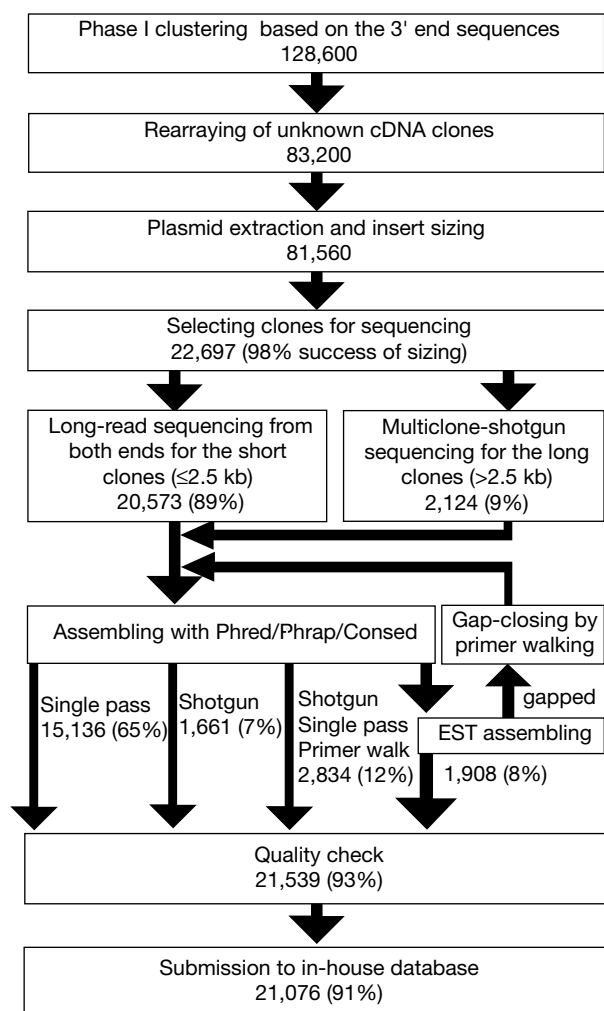
Figure 1 Phase II full-insert sequencing flow chart. We listed 83,200 clones for full-insert sequencing from 128,600 Phase I clusters. To expedite sequencing and increase the likelihood of discovering new genes, we focused on short cDNAs with novel 3′ end sequences (22,697 clones). Clones shorter than 2.5 kilobases (kb) were sequenced from both ends using long-read Licor DNA4200 sequencers; clones longer than 2.5 kb were subjected to multi-clone shotgun sequencing (M. Yoshino *et al.*, unpublished) using RISA (RIKEN Integrated Sequence Analysis)[23,24]. Electropherograms from the four sequencers used (RISA (Shimadzu), Licor DNA 4200 (Licor), ABI377 and ABI3700 (Applied Biosystems Inc.)) were base-called by PHRED, assembled by PHRAP and edited with CONSED[7,8,25] in three steps. The computer-assisted system assembled the raw sequence data (15,136 and 1,661 clones, respectively, from Licor and Shotgun strategies); we closed the gaps (1,908 assemblies) using public EST databases; and remaining gaps were resequenced by primer walking (2,834 assemblies).

Table 1 Gene categories

Gene categories	Genes	Clones
MGI-confirmed	2,390	4,248
Identical to	488	841
Similar to	703	930
Homologue to	3,550	5,525
Related to	564	841
Motif-containing protein	573	740
Hypothetical protein	1,634	2,086
Unclassifiable transcript	3,251	3,626
Unclassifiable	2,142	2,239
Total	15,295	21,076

Number of genes and clones assigned to each gene category. The definition of each gene category is described in the Methods. Here, a gene refers to a representative clone or a singleton.

Table 2 Number of clones assigned to GO functional categories

GO term	GO ID	Number
Enzyme	GO:0003824	943
Nucleic acid binding	GO:0003676	725
Ligand binding or carrier	GO:0005488	687
Structural protein	GO:0005198	459
Signal transduction	GO:0004871	361
Transporter	GO:0005215	332
Motor	GO:0003774	208
Chaperone	GO:0003754	94
Enzyme inhibitor	GO:0004857	75
Cell cycle regulator	GO:0003750	60
Other		68
Total		4,012

Any given sequence may have been assigned to more than one category. We assigned 4,012 GO functional terms to 3,025 genes as described in the text.

686

NATURE | VOL 409 | 8 FEBRUARY 2001 | www.nature.com

Orthologues of human disease genes

Identifying orthologues of human disease genes in model organisms and creating animal models of human disorders should help us to understand the relationships between genetic variants and human diseases, and may be useful for testing diagnostic and therapeutic strategies. To find orthologues of human disease genes in the RIKEN cDNA set, we compared the clones to 288 human disease gene orthologues compiled for the *Drosophila* genome sequence paper[16]. Of this list, 118 genes (44%) share significant protein sequence identity with one or more of the RIKEN clones. We found novel mouse orthologues for ten of these human disease genes: two cancer-related genes (DEK oncogene and BCR), three genes related to neurological disorders (dysferlin, MJD and USH2A), three genes related to malformation syndromes (CKN1, PEX1 and Tafazzin), and two genes related to haematological disorders (α-haemoglobin and XK).

Identification of new mouse genes

Some of the cDNAs in the 'similar to', 'homologue to', 'related to', 'motif-containing protein', 'hypothetical protein' and 'unclassifiable transcript' categories are likely to represent new mouse genes. As in other completed genomes[17], many new genes are members of large multigene families associated with cellular differentiation and signal transduction. For example, we identified 251 genes in the non-redundant clone set encoding zinc-finger containing proteins. For many, there was no evident relationship to any other known gene. Ten cDNAs encoded proteins that contain the SAP domain, a DNA-binding motif[18] (see Supplementary Information Fig. 2). Five of the ten mouse proteins containing this domain have not been characterized, including those that match newly identified predicted human genes. An additional 31 RIKEN clones were annotated as homeobox genes, including seven for which this was the only annotated feature.

Seventy-four of the cDNAs showed recognizable homology to known protein phosphatases, or contained a phosphatase motif. Of these, perhaps the most interesting are 14 clones that are predicted to be members of the dual-specificity protein phosphatase family (see Supplementary Information Table 4A). Each is likely to be a new member of this family and, on the basis of the literature, a possible candidate disease gene or tumour suppressor.

One-hundred and ninety-three cDNA clones in the RIKEN set were identified as protein kinases. Most are closely related to known serine-threonine and tyrosine kinases; only 14 were identified solely on the basis of the consensus kinase signature motif (see Supplementary Information Table 4B). Most of these correspond only to the kinase domain of known genes and may therefore represent novel transcripts.

Although the RIKEN set seems to contain many novel cDNAs, it contains relatively few known or new host defence genes. Most such genes are induced by immunological challenges, and lymphoid organs such as spleen contain complex mixtures of cell types in different states of activation. Therefore, even following induction, immune-related transcripts have low abundance in total tissue mRNA. Future efforts in compiling the mouse gene encyclopaedia will include production and sequencing of libraries from various stimulated immune cells.

Determining protein domains and families

For sequences without recognizable homology to a known gene, we searched for functional motifs in an attempt to predict their protein products. We performed FASTA searches against InterPro and HMMER searches against the TIGR-FAM database using DECODER-predicted protein sequences. In addition, we searched the Pfam database with the program ESTwise using the RIKEN clone nucleotide sequences. The search of RIKEN clones against InterPro data is shown in Supplementary Information Table 4C. InterPro motifs were identified in 3,204 new mouse genes.

Whenever possible, InterPro identifiers were used to allocate the GO terms.

New protein motifs

We used maximum density subgraph analysis[19] to identify six new motifs in our cDNAs, which were not present in the Pfam, ProDom and InterPro databases. Hidden Markov models (HMM) were constructed for these motifs and used to search the Swiss-Prot-TrEMBL nonredundant database using HMMER version 2.1.1. Two such searches resulted in the discovery of motifs in the organic-anion transporting polypeptide (Oatp) family sequences (see alignment at OATP, Supplementary Information Fig. 3B). A phylogenetic analysis (see Supplementary Information Fig. 3A) indicates that RIKEN clones 7516, 15434 and 18937 may belong to a new Oatp subfamily. The results of other HMM searches are shown in Supplementary Information Table 5A.

Untranslated regions

Sequences affecting the translation and stability of mRNAs are found in the 5′ and 3′ UTRs. Using PatSearch[20], we screened our clone set for UTR-specific functional motifs in UTRsite (http://bigarea.area.ba.cnr.it:8000/EmbIT/UTRHome). This occasionally added to the functional annotation of clones. For example, a histone 3′ UTR stem–loop structure in two unclassifiable RIKEN cDNAs (RIKEN clones 10172 and 22851) indicates that these clones correspond to the 3′ UTRs of mRNAs encoding histone proteins (see Supplementary Information Table 5B for these and other common motifs in the RIKEN clones).

Chromosomal mapping of cDNA clones

RIKEN clones corresponding to those in the Whitehead Mouse and
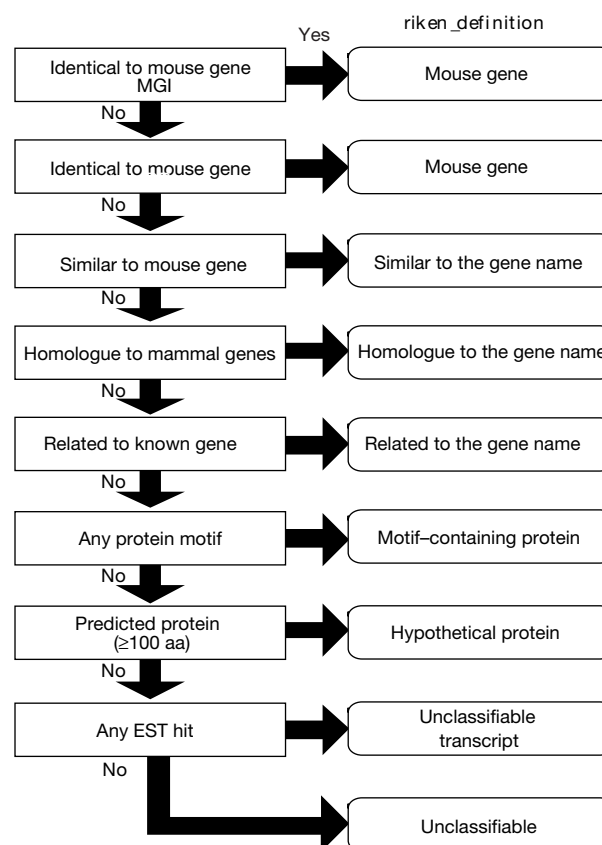
Figure 2 The criteria used in assigning RIKEN definitions (riken_defs). See Methods for details.

Jackson Laboratory radiation hybrid databases were directly mapped onto the mouse genome (see Supplementary Information Fig. 4 and Table 6). We also mapped 12,191 clones to human genomic sequences and then onto the mouse genome on the basis of synteny between the mouse and human genomes (http://www.gsc.riken.go.jp/e/FANTOM/map/).

We identified 817 hypothetical transcripts for which there were no corresponding human genes in the RefSeq, human UniGene or Ensembl transcript databases. Of these, 485 mapped to one or more GenScan-predicted exons in the human draft sequence. Among the 485 putative transcripts, 174 perfectly matched the GenScan predictions; 311 showed only partial matches because GenScan seemed not to have predicted one or more exons. The remaining 332 cDNAs did not hit any exon predicted by GenScan. These data strongly support the importance of cDNA sequencing in the identification of genes that would not otherwise be discovered in genomic sequences and indicate the need for caution when using *ab initio* predictions as the primary source for genome annotation.

Discussion

Functional annotation of the first RIKEN mouse cDNA set clearly validates the overall strategy, but also suggests the need for further refinements and for similar projects in other species. At least some unclassifiable transcripts probably represent unprocessed nuclear RNA that could potentially be avoided by isolating cytoplasmic RNA. Future prioritization will include 5′ as well as 3′ end sequencing. In this clone set, we focused on shorter transcripts and this may have had the unintended consequence of enriching for truncations and abundant gene products already present in the MGI database.

Information about these clones is available at RIKEN (http://www.gsc.riken.go.jp/e/FANTOM/viewer/) and Mouse Genome Informatics (http://www.informatics.jax.org and mirror sites). We would welcome suggestions for annotation. Ultimately, however, the annotation of this and other clone sets, as well as the human, mouse, rat and other genomes, will come from careful experimental analysis of the identified coding sequences. A variety of techniques, including microarray analysis and two-hybrid screens, will lend significant experimental support for functional assignments and lead to the discovery of pathways and gene families. As the new mouse genes in this set become better characterized, revised nomenclature and other biological data will be incorporated into their MGI and FANTOM records. Further computational analyses, including cross-species comparisons, will further elucidate the functions of the newly identified genes and may assist in the identification of genomic regulatory regions. □

Methods

Phase I

All cDNA libraries were prepared from C57BL/6J mouse mRNA using a strategy designed to enhance representation of full-length transcripts. About 160 cDNA libraries were enriched in full-length inserts by applying several technologies including cap trapping[4,5], thermoactivation of reverse transcriptase by trehalose[21], normalization and subtraction[6] and vectors designed for the preferential cloning of long inserts (P. Carninci, manuscript in preparation).

Phase II

The representative clones were regridded. The sequencing strategy (Fig. 1) and sequence editing approaches are described in Supplementary Information methods.

Gene assignments and functional annotation of genes

We used a variety of software programs, including BLASTN, BLASTX, FASTA/FASTY (ftp://ftp.virginia.edu/pub/fasta/), DECODER, EST-WISE (http://www.sanger.ac.uk/Software/Wise2/) and HMMER (http://hmmer.wustl.edu/), to search databases including NCBI-nr, Locus Link, SwissProt, SwissProt TrEMBL, TIGR nraa, PFAM, TIGR-FAM, UniGene, the TIGR Gene Indices, UTRdb and UTRsite, and a number of species-specific databases (see Supplementary Information Table 7A and B). (DECODER[26] is an amino-acid translation program designed to suggest the position of experimental frame-shift errors, and predict amino-acid sequences for full-length cDNA sequences with PHRED scores. The program generates artificial insertions into and artificial deletions from the low-accuracy base positions of the original sequence, thereby generating many candidate sequences. The validity of the most probable sequence (the likelihood that it represents the actual protein) is evaluated by using a score (V_a) that is calculated in light of the Kozak consensus, preferred codon usage and position of the initiation codon.) Additional analyses were performed using the bioSCOUT program (LION Bioscience). Protein domain analyses were conducted at the European Bioinformatics Institute using the InterPro software program.

Curators annotated clones with the help of the FANTOM+ interface, which allowed users to view pre-computed similarity and motif search results, to launch additional searches, and to transfer the annotation from any of these to the FANTOM database.

The aim of the FANTOM meeting was to assign each RIKEN clone a RIKEN definition (riken_def) to indicate its most likely function and/or status on the basis of similarity to known genes. A supplementary RIKEN definition line (riken_def_suppl) was available in the interface for additional annotation. Annotation of RIKEN clones with significant similarity to known sequences was guided by the gene or gene product descriptors of the reference sequences to which the RIKEN clones were similar. In general, the riken_def was derived from the gene descriptor of the reference sequence that had the highest similarity to the RIKEN clone sequence. When the RIKEN clone was highly similar to several genes, an annotation hierarchy was used to choose the riken_def, based on the species of origin and descriptor content for the candidate reference sequences (Fig. 2).

Priority was given to reference sequence descriptors from which functional information could be inferred, even if sequences with less informative descriptors were more similar to the clones. Annotations from highly curated databases (MGI and SwissProt) were preferred and provided convenient entry points into the GO vocabularies. Informative descriptors from mouse genes identical to RIKEN clones were the first choice for annotation. Official gene nomenclature was used preferentially for the 'MGI-confirmed' set. For RIKEN clones identical to mouse genes not represented in MGI ('identical-to') or with non-identical similarity to known genes, riken_defs were derived from informative gene descriptors according to the following species priority: identical mouse > non-identical mouse > non-mouse mammal > non-mammal. The controlled vocabulary prefix terms 'similar to', 'homologue to' and 'related to' were used in the riken_def line to indicate that a gene descriptor was derived from non-identical mouse, non-mouse mammal or non-mammal sources, respectively. RIKEN clones with no significant sequence similarity to known genes were named on the basis of coding potential, protein motif signature and representation in mouse, human or rat EST databases. RIKEN clones with no significant similarity to known genes, but with predicted protein motifs found in Pfam and/or InterPro, were named '⟨motif name⟩-containing protein'. Clones with no known sequence similarity or domain hits, but with coding potential ≥100 amino acids and EST representation, were named 'hypothetical protein'. Clones belonging to none of the above groups, but with matches to ESTs, were referred to as 'unclassifiable transcript'. Clones with no EST matches were called 'unclassifiable'. New mouse genes discovered in the RIKEN clone set will be assigned official nomenclature in MGI according to a defined syntax: gene symbol, ⟨Riken Clone Identifier⟩Rik; gene name, Riken cDNA ⟨Riken Clone Identifier⟩ gene (for example, 2610307C23Rik; Riken cDNA 2610307C23 gene). For novel genes represented by RIKEN clusters, nomenclature will be taken from the clone identifiers of the representative clones for each cluster.

Computational identification of full-length clones

See Supplementary Information Table 2A.

Assignment of gene ontology (GO) terms

See Supplementary Information Table 7.

Mapping RIKEN clones using mouse radiation hybrid (RH) data

Repeat-masked RIKEN clone sequences were BLAST-searched against the Whitehead Mouse RH database and the Jackson Laboratory RH database. Identity ≥98% over more than 100 bp was considered an exact correspondence. In total, 8,960 sequences were unique after eliminating redundancy between these two databases. The RIKEN sequences were searched by BLASTN against this nonredundant set. Among the RIKEN clones, 3,398 were matched; 2,469 and 3,085 RIKEN clones were mapped onto the Whitehead Mouse RH database and The Jackson Laboratory RH database, respectively.

Mapping of RIKEN full-length cDNAs onto the human genomic sequence

To detect even short exons (20 bp), we conducted a BLASTN search (E-value = 1.0) between repeat-masked RIKEN clones and the human genome sequences (15 June 2000 version) provided by the Center for Biomolecular Science and Engineering, UCSC, which comprises three billion bases and represents each chromosome in one continuous contig (at the time of analysis 19.7% of the sequence was incomplete, that is, N base). These selected exons were used for mapping our cDNAs. The criterion for mapping was much more stringent, based upon the sum of the lengths of homologues being > 200 bp. All nonredundant cDNAs were compared pairwise to the 10,239 human reference genes of RefSeq, 81,963 UniGene clusters and 37,720 Ensembl transcripts. To identify the candidates for the hypothetical genes, we eliminated the RIKEN cDNAs that showed homology (≥ 100 bp at > 70% identity by BLASTN) to the above database. For mapping, an average of 85–88% identity was reported between mouse and human mRNAs of orthologous sequences[22].

The RIKEN mouse cDNA clones will be publicly available in May 2001 when we have replicated the clones and sent them to the distributor. Information on how to obtain these clones can be obtained from http//genome.gsc.riken.go.jp.

1. 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).
2. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
3. Liang, F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**, 239–240 (2000).
4. Carninci, P. & Hayashizaki, Y. High-efficiency full-length cDNA cloning. *Methods Enzymol.* **303**, 19–44 (1999).
5. Carninci, P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**, 327–336 (1996).
6. Carninci, P. *et al.* Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**, 1617–1630 (2000).
7. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
8. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
9. Gautheret, D., Poirot, O., Lopez, F., Audic, S. & Claverie, J. M. Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.* **8**, 524–530 (1998).
10. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37–45 (1997).
11. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
12. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
13. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
14. Croft, L. *et al.* ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genet.* **24**, 340–341 (2000).
15. Hanke, J. *et al.* Alternative splicing of human genes: more the rule than the exception? *Trends Genet.* **15**, 389–390 (1999).
16. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
17. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
18. Aravind, L. & Koonin, E. V. SAP- a putative DNA-binding motif involved in chromosomal organization. *Trends Biochem. Sci.* **25**, 112–114 (2000).
19. Matsuda, H. Detection of conserved domains in protein sequences using a maximum-density subgraph algorithm. *IEICE Trans. Fundamentals Electron. Commun. Comput. Sci.* **E83-A**, 713–721 (2000).
20. Pesole, G., Liuni, S. & D'Souza, M. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics* **16**, 439–450 (2000).
21. Carninci, P. *et al.* Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc. Natl Acad. Sci. USA* **95**, 520–524 (1998).
22. 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).
23. Itoh, M. *et al.* Automated filtration-based high-throughput plasmid preparation system. *Genome Res.* **9**, 463–470 (1999).
24. Shibata, K. *et al.* RIKEN integrated sequence analysis (RISA) system- 384-format sequencing pipline with 384 multicapillary sequencer. *Genome Res.* **10**, 1757–1771 (2000).
25. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
26. Fukunishi, Y. & Hayashizaki, Y. Amino-acid translation program for full-length cDNA sequences with frame-shift error. *Physiol. Genomics.* (in the press).

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*. It is also available on the RIKEN Genomic Sciences Center web site (https://genomec.gsc.riken.go.jp/genome/fantom1/paper/Fig./nature/supplement/; user name: fantom1; password: fntm0828) and at http://www.gsc.riken.go.jp/e/FANTOM/supplement/.

RIKEN Genome Exploration Research Group Phase II team:
J. Kawai[1,2], A. Shinagawa[1], K. Shibata[1,2], M. Yoshino[1], M. Itoh[1,2], Y. Ishii[1], T. Arakawa[1], A. Hara[1], Y. Fukunishi[1,2], H. Konno[1,2], J. Adachi[1], S. Fukuda[1,2], K. Aizawa[1,2], M. Izawa[1], K. Nishi[1], H. Kiyosawa[1], S. Kondo[1], I. Yamanaka[1] & T. Saito[1]

FANTOM Consortium: Y. Okazaki[1], T. Gojobori[3], H. Bono[1], T. Kasukawa[4], R. Saito[1], K. Kadota[1], H. Matsuda[5], M. Ashburner[6], S. Batalov[7], T. Casavant[8], W. Fleischmann[6], T. Gaasterland[9], C. Gissi[10], B. King[11], H. Kochiwa[12], P. Kuehl[13], S. Lewis[14], Y. Matsuo[15], I. Nikaido[16], G. Pesole[10], J. Quackenbush[17], L. M. Schriml[18], F. Staubli[19], R. Suzuki[12], M. Tomita[12], L. Wagner[18], T. Washio[12], K. Sakai[1], T. Okido[1], M. Furuno[1], H. Aono[1], R. Baldarelli[11], G. Barsh[20], J. Blake[11], D. Boffelli[21], N. Bojunga[19], P. Carninci[1], M. F. de Bonaldo[22], M. J. Brownstein[23], C. Bult[11], C. Fletcher[7], M. Fujita[24], M. Gariboldi[25], S. Gustincich[26], D. Hill[11], M. Hofmann[19], D. A. Hume[27], M. Kamiya[1,2], N. H. Lee[17], P. Lyons[28], L. Marchionni[29], J. Mashima[3], J. Mazzarelli[30], P. Mombaerts[31], P. Nordone[32], B. Ring[33], M. Ringwald[11], I. Rodriguez[31], N. Sakamoto[34], H. Sasaki[35], K. Sato[1,44], C. Schönbach[36], T. Seya[37], Y. Shibata[1], K.-F. Storch[38], H. Suzuki[1], K. Toyo-oka[39], K. H. Wang[40], C. Weitz[38], C. Whittaker[41], L. Wilming[42], A. Wynshaw-Boris[43], K. Yoshida[1], Y. Hasegawa[4], H. Kawaji[4,5] & S. Kohtsuki[4]

General organizer: Y. Hayashizaki[1,2,44]

1, Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), Yokohama Institute 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; 2, CREST, JST, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074 Japan; 3, Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan; 4, NTT Software Corporation, 223-1 Yamashita-cho, Naka-ku, Yokohama, Kanagawa, 231-8554, Japan; 5, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan; 6, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; 7, Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, California 92121, USA; 8, The Coordinated Laboratory for Computational Genomics, University of Iowa Iowa City, Iowa 52242, USA;

9, The Rockefeller University, 1230 York Avenue, New York, New York 10021-6399, USA; 10, Dipartimento di Fisiologia e Biochimica Generali, Universita di Milano Via Celoria, 26, 20133 Milano, Italy; 11, Mouse Genome Informatics, The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609, USA; 12, Laboratory for Bioinformatics, Faculty of Environmental Information, Keio University, 5322 Endoh, Fujisawa, Kanagawa, 252-0816, Japan; 13, Department of Molecular & Cell Biology, University of Maryland at Baltimore, Baltimore, Maryland 20201, USA; 14, University of California, Berkeley, Department of Molecular & Cell Biology, 142 Life Sciences Addition #3200, Berkeley, California 94720-3200, USA; 15, Computational Proteomics Team, Bioinformatics Group, RIKEN Genomic Sciences Center (GSC), Yokohama Institute 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; 16, Tokai University, Graduate School of Marine Science and Technology, 3-20-1 Orido, Shimizu, Shizuoka,424-8610 Japan; 17, The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, Maryland 20850, USA; 18, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room 8N805, Bethesda, Maryland 20894, USA; 19, LION Bioscience AG, Im Neuenheimer Feld 515-519, D-69120 Heidelberg, Germany; 20, Stanford University School of Medicine, Beckman Centre B271A, Stanford, California 94305-5428, USA; 21, Lawrence Berkeley Laboratory, 1 Cyclotron Rd, MS84-255, Berkeley, California 94710, USA; 22, Department of Pediatrics, The University of Iowa, 200 Hawkins Drive 440B EMRB, Iowa City, Iowa 52242-1009, USA; 23, Laboratory of Genetics, NIMH/NHGRI, National Institutes of Health Building 36, Room 3D06, Bethesda, Maryland 20892, USA; 24, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan; 25, Istituto Tumori Milano, Via Venezian,1, 120133 Milano, Italy; 26, Department of Neurobiology, Harvard Medical School, 220 Longwood Ave., Boston, Massachusetts 02115, USA; 27, Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia; 28, Department of Medical Genetics, Wellcome Trust Centre for Molecular Mechanisms in Disease, University of Cambridge, Wellcome Trust/MRC building, Addenbrookes Hospital, Cambridge, CB2 2XY, UK; 29, LNCIB c/o AREA Science Park, Padriciano 99, 34012 Trieste, Italy; 30, Computational and Bioinformatics Laboratory, Center for Bioinformatics, University of Pennsylvania, 1313 Blockley Hall, 418 Guardian Drive, Philadelphia, Pennsylvania 19104-6021, USA; 31, Vertebrate Developmental Neurogenetics, The Rockefeller University, 1230 York Avenue, Box 242, New York, New York 10021-6399, USA; 32, University at Buffalo/Roswell Park Cancer Institute, 120 Meyers Rd.#615, Amherst, New York 14226, USA; 33, Department of Genetics, Stanford University, Beckman Centre B281, Stanford, California 94305, USA; 34, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan; 35, National Cancer Research Institute, 1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan; 36, Computational Genomics Team, Bioinformatics Group, RIKEN Genomic Sciences Center (GSC), Yokohama Institute 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; 37, Osaka Medical Center for Cancer, Nakamichi 1-3-3, Higashinari-ku, Osaka 537-8511 Japan; 38, Department of Neurobiology, Harvard Medical School, 220 Longwood Ave., Boston, Massachusetts 02115, USA; 39, University of California, San Diego, School of Medicine, Department of Pediatrics, 9500 Gilman Dr., Medical Teaching Facility 253, La Jolla, California 92093-0627, USA; 40, E17-353, Center for Learning and Memory, Massacusetts Institute of Technlogy, 77 Massachusetts Ave., Cambridge, Massachusetts 02139, USA; 41, Massachusetts Institute of Technology, MIT CCR, 77 Massachusetts Avenue 17-230, Cambridge, Massachusetts 02139, USA; 42, Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; 43, University of California San Diego School of Medicine, 9500 Gilman Dr., Medical Teaching Facility, Room 252, La Jolla, California 92093-0627, USA; 44, Tsukuba University, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan.