nature genetics

Complete sequencing and characterization of 21,243 full-length human cDNAs

Toshio Ota[1,2], Yutaka Suzuki[3], Tetsuo Nishikawa[1,4], Tetsuji Otsuki[1], Tomoyasu Sugiyama[1], Ryotaro Irie[1], Ai Wakamatsu[1], Koji Hayashi[1], Hiroyuki Sato[1], Keiichi Nagai[1], Kouichi Kimura[4], Hiroshi Makita[4], Mitsuo Sekine[5], Masaya Obayashi[2], Tatsunari Nishi[2], Toshikazu Shibahara[3,6], Toshihiro Tanaka[3], Shizuko Ishii[1], Jun-ichi Yamamoto[1], Kaoru Saito[1], Yuri Kawai[1], Yuko Isono[1], Yoshitaka Nakamura[1], Kenji Nagahari[1], Katsuhiko Murakami[4], Tomohiro Yasuda[4], Takao Iwayanagi[4], Masako Wagatsuma[7], Akiko Shiratori[7], Hiroaki Sudo[7], Takehiko Hosoiri[7], Yoshiko Kaku[7], Hiroyo Kodaira[7], Hiroshi Kondo[7], Masanori Sugawara[7], Makiko Takahashi[7], Katsuhiro Kanda[7], Takahide Yokoi[7], Takako Furuya[7], Emiko Kikkawa[7], Yuhi Omura[7], Kumi Abe[7], Kumiko Kamihara[7], Naoko Katsuta[7], Kazuomi Sato[7], Machiko Tanikawa[7], Makoto Yamazaki[7], Ken Ninomiya[7], Tadashi Ishibashi[8], Hiromichi Yamashita[8], Katsuji Murakawa[8], Kiyoshi Fujimori[8], Hiroyuki Tanai[8], Manabu Kimata[8], Motoji Watanabe[8], Susumu Hiraoka[8], Yoshiyuki Chiba[8], Shinichi Ishida[8], Yukio Ono[8], Sumiyo Takiguchi[8], Susumu Watanabe[8], Makoto Yosida[8], Tomoko Hotuta[8], Junko Kusano[8], Keiichi Kanehori[8], Asako Takahashi-Fujii[9], Hiroto Hara[9], Tomo-o Tanase[9], Yoshiko Nomura[9], Sakae Togiya[9], Fukuyo Komai[9], Reiko Hara[9], Kazuha Takeuchi[9], Miho Arita[9], Nobuyuki Imose[9], Kaoru Musashino[9], Hisatsugu Yuuki[9], Atsushi Oshima[9], Naokazu Sasaki[10], Satoshi Aotsuka[10], Yoko Yoshikawa[10], Hiroshi Matsunawa[10], Tatsuo Ichihara[10], Namiko Shiohata[10], Sanae Sano[10], Shogo Moriya[10], Hiroko Momiyama[10], Noriko Satoh[10], Sachiko Takami[10], Yuko Terashima[10], Osamu Suzuki[10], Satoshi Nakagawa[2], Akihiro Senoh[2], Hiroshi Mizoguchi[2], Yoshihiro Goto[6], Fumio Shimizu[6], Hirokazu Wakebe[6], Haretsugu Hishigaki[6], Takeshi Watanabe[6], Akio Sugiyama[11], Makoto Takemoto[11], Bunsei Kawakami[11], Masaaki Yamazaki[12], Koji Watanabe[12], Ayako Kumagai[12], Shoko Itakura[12], Yasuhito Fukuzumi[12], Yoshifumi Fujimori[12], Megumi Komiyama[12],Hiroyuki Tashiro[12], Akira Tanigami[6], Tsutomu Fujiwara[6], Toshihide Ono[6], Katsue Yamada[6], Yuka Fujii[6], Kouichi Ozaki[6], Maasa Hirao[6], Yoshihiro Ohmori[3,6], Ayako Kawabata[2], Takeshi Hikiji[2], Naoko Kobatake[2], Hiromi Inagaki[2], Yasuko Ikema[2], Sachiko Okamoto[2], Rie Okitani[2], Takuma Kawakami[3,6], Saori Noguchi[3], Tomoko Itoh[3], Keiko Shigeta[3], Tadashi Senba[3], Kyoka Matsumura[3], Yoshie Nakajima[3], Takae Mizuno[3], Misato Morinaga[3], Masahide Sasaki[3], Takushi Togashi[3], Masaaki Oyama[3], Hiroko Hata[3], Manabu Watanabe[3], Takami Komatsu[3], Junko Mizushima-Sugano[3], Tadashi Satoh[3,4], Yuko Shirai[3], Yukiko Takahashi[3], Kiyomi Nakagawa[3], Koji Okumura[13], Takahiro Nagase[14], Nobuo Nomura[14,15], Hisashi Kikuchi[5], Yasuhiko Masuho[1], Riu Yamashita[3], Kenta Nakai[3], Tetsushi Yada[3], Yusuke Nakamura[3], Osamu Ohara[14], Takao Isogai[1] & Sumio Sugano[3,15]

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

As a base for human transcriptome and functional genomics, we created the "full-length long Japan" (FLJ) collection of sequenced human cDNAs. We determined the entire sequence of 21,243 selected clones and found that 14,490 cDNAs (10,897 clusters) were unique to the FLJ collection. About half of them (5,416) seemed to be protein-coding. Of those, 1,999 clusters had not been predicted by computational methods. The distribution of GC content of nonpredicted cDNAs had a peak at ~58% compared with a peak at ~42%for predicted cDNAs. Thus, there seems to be a slight bias against GC-rich transcripts in current gene prediction procedures. The rest of the cDNAs unique to the FLJ collection (5,481) contained no obvious open reading frames (ORFs) and thus are candidate noncoding RNAs. About one-fourth of them (1,378) showed a clear pattern of splicing. The distribution of GC content of noncoding cDNAs was narrow and had a peak at ~42%, relatively low compared with that of protein-coding cDNAs.

Now that most of the human genomic sequence has been determined[1–5] efforts are focused on its annotation. Full-length cDNAs, which are complete copies of mRNAs, are a particularly important resource for identifying genes and determining their structural features, forming a basis for transcriptome analysis. Physical cDNA clones are also indispensable reagents in the experimental analysis of gene functions, particularly in higher eukaryotes, such as humans.

In 1999, we started the FLJ project, which aimed to collect and determine the complete sequences of putatively full-length human cDNAs. At that time, there were only about 6,000 cDNAs in RefSeq, the curated

[1]Helix Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan. [2]Kyowa Hakko Kogyo, Tokyo Research Laboratory, 3-6-6 Asahi-machi, Machida, Tokyo 194-8533, Japan. [3]The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan. [4]Hitachi, Central Research Laboratory, 1-280 Higashi-koigakubo, Kokubunj, Tokyo 185-8601, Japan. [5]National Institute of Technology and Evaluation, 2-49-10 Nishihara, Shibuya-ku, Tokyo 151-0066, Japan. [6]Otsuka Pharmaceutical, 463-10 Kagasuno Kawauchi-cho, Tokushima 771-0192, Japan. [7]Hitachi, Life Science Group, 1-3-1 Minamidai, Kawagoe, Saitama 350-1165, Japan. [8]Hitachi Science Systems, 1-280 Higashi-koigakubo, Kokubunji, Tokyo 185-8601, Japan. [9]Takara Shuzo, 2257 Sunaike, Noji, Kusatu, Shiga 525-0055, Japan. [10]Nisshinbo Industries, 1-2-3 Onodai, Midori-ku, Chiba 267-0056, Japan. [11]Toyobo, 10-24 Toyo-cho, Tsuruga, Fukui 914-0047, Japan. [12]Fujiya, 228 Soya, Hadano, Kanagawa 257-0031, Japan. [13]Aisin Cosmos R&D, 1698 Yana, Kisarazu, Chiba 292-0812, Japan. [14]Kazusa DNA Research Institute,1532-3 Yana, Kisarazu, Chiba 292-0812 Japan. [15]BIRC, AIST, 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan. Correspondence should be addressed to S. Sugano (ssugano@ims.u-tokyo.ac.jp).

Published online 21 December 2003; doi:10.1038/ng1285

informational resource[6], with full-length cDNA sequence information. One million cDNAs clones were available from the IMAGE collection, the physical resource for cDNA clones. Unfortunately, most IMAGE clones were only partly sequenced (as expressed-sequence tags; ESTs[7]) and there were no good clues to indicate which ones were full-length. Therefore, we began the large-scale collection and full-length sequencing of cDNAs not only to obtain cDNA sequence information but also to provide a physical source of cDNA clones. Here, we report the first characterization of 21,243 clones.

RESULTS

Collecting and sequencing full-length cDNAs

To facilitate the large-scale collection and sequencing of full-length cDNAs, we constructed 107 human cDNA libraries enriched for full-length cDNA clones representing 61 tissues, 21 primary cell cultures and 16 cell lines. We used a cap-targeted selection method called oligo-capping[8,9] for all but one spleen library, which was constructed using a highly selective size-fractionation method to clone cDNAs of long mRNAs[10]. **Supplementary Table 1** online shows the cDNA libraries we used. The average frequency of full-length cDNA clones in the libraries was 85%.

We randomly picked cDNA clones from these libraries and subjected them to one-pass sequencing. In total, we obtained the 5′-end one-pass sequences from 1,154,510 cDNA clones. In some cases, we compared these sequences to GenBank using BLAST searching[11]. We found that 40–60% of our cDNA sequences matched RefSeq entries at the time of the search. The rest matched only human ESTs or had no matches. We selected 21,243 cDNA clones from the two latter categories and determined their complete sequences. For each of the cDNAs, we completely sequenced both strands, mainly using the primer walking method[12]. Judging from the sequence quality scores calculated using Phred[13], the accuracy of the sequence data was more than 99.99%. The average length of the cDNAs was 2,314 bp (**Supplementary Fig. 1** online). All the sequence data have been registered in public databases through the DNA database of Japan (DDBJ). Searches for the cDNAs by accession numbers, chromosomal positions, various keywords or sequence similarities are enabled at our website (http://fldb.hri.co.jp/cgi-bin/cDNA3/public/publication/index.cgi). All physical cDNA clones are freely available for research use on request. Requests for physical cDNA clones should be sent to flcdna@ims.u-tokyo.ac.jp or isogai-t@reprori.jp.

Comparison of the FLJ collection with predicted genes

We used BLAST searches to compare our full-length cDNA sequences in series with RefSeq (and other relevant data sets, as outlined in **Fig. 1**). For 14,490 cDNAs, there were no matches in RefSeq (excluding the 2,313 RefSeq entries that were derived from the FLJ collection), indicating that the full sequences of these cDNAs were unique to the FLJ collection ('FLJ-unique'). The remainder (6,753 cDNAs in 4,263 clusters) at least partially matched sequences in RefSeq. This overlap was because many cDNA sequences that did not have matches at the time of the selection were later sequenced by other researchers during the course of this project. The matches also included alternatively spliced isoforms in RefSeq. There were also a number of cDNAs that were identical to sequences in RefSeq but were selected because of inaccurate one-pass sequence data or human error when picking the clones. We clustered the 14,490 FLJ-unique cDNAs pairwise to remove the redundancy, resulting in 10,897 nonredundant cDNA clusters ('nonredundant FLJ-unique').

To determine what proportion of these 10,897 nonredundant FLJ-unique clusters had been previously predicted from the genome sequence (for chromosome assignments of the FLJ cDNAs, see **Supplementary Table 2** online), we used BLAST searches to compare these cDNAs with Ensembl genes (using version 4.28.1; 29,076 genes), which are representative predicted genes based on comprehensive analyses of a wide range of evidence and are expected to cover most human genes[14]. Ensembl genes supported by RefSeq were removed from the search. Among the 10,897 clusters, 2,774 at least partly matched Ensembl genes ('Ensembl-predicted') and 8,123 did not ('Ensembl-nonpredicted').

We then examined whether the 8,123 Ensembl-nonpredicted clusters were predicted by *ab initio* gene prediction programs, such as DIGIT, FGENESH, GENSCAN and HMMGENES, because the criteria for identifying Ensembl genes are based on a somewhat conservative method of evaluation. As shown in **Table 1**, 643 Ensembl-nonpredicted clusters were predicted by at least one of these programs ('*ab initio*–predicted'). In total, 3,417 (31%) of the 10,897 nonredundant FLJ-unique clusters corresponded to genes predicted by one or more of the computational methods. These *ab initio* methods did not predict the existence of the remaining 7,480 clusters ('nonpredicted').

One possible reason why these sequences were not predicted by the computational methods is that most of them were non-protein-coding, and many computational methods predict only protein-coding regions as genes. To check this possibility, we determined how many

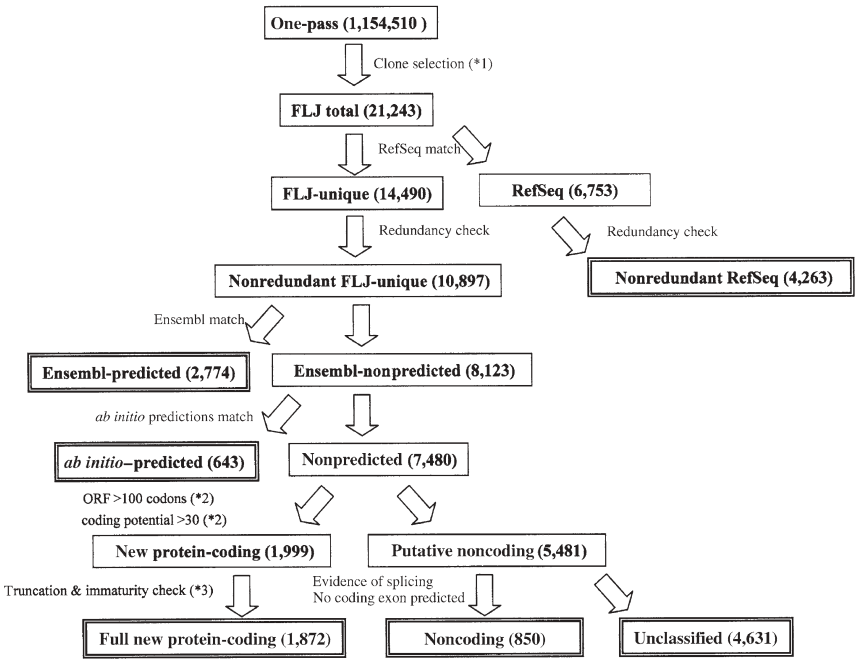

Figure 1 Flow chart of cDNA categorization. Each cDNA was categorized as shown here. For further details on the categorization (steps *1–*3), refer to **Supplementary Note** and **Supplementary Figure 6** online. Detailed descriptions of the cut-offs and the supporting evidence are also presented there.

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

Table 1 Sequence comparison of the 10,897 non-redundant FLJ unique cDNAs and computationally predicted cDNAs

ab initio prediction program used	DIGIT	FGENESH	GENSCAN	HMMGENE	All*
Total number predicted in entire human genome	2,908	15,144	16,732	10,083	45,944
Number predicted in Ensembl nonpredicted (8,123)	193	540	560	340	643

The identities of nonredundant FLJ-unique cDNAs with the Ensembl genes were searched using BLAST with a cut-off value of 1.0e-100. Ensembl genes that were supported by RefSeq were excluded from the search. For the cDNAs that did not match the Ensembl genes, we analyzed whether they could have been predicted by the *ab initio* gene prediction programs. Using DIGIT, GENSCAN, FGENESH and HMMGENES, cDNAs of genes were predicted from the human genomic sequence. The total numbers of cDNA sequences predicted by these programs are shown in the second row. The sequences of nonredundant FLJ-unique cDNAs and those of the predicted cDNAs were compared using BLAST with a cut-off value of 1.0e-100. The numbers of accurately predicted cDNAs obtained using each program are shown. *The cDNA sequences predicted by all of the four programs were merged.

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

clusters met the criteria of ORF length >100 codons and coding potential >30 (calculated according to the standard method[15]). We used these criteria because most of the ORFs registered in RefSeq satisfied them, whereas ORFs randomly occurring in the human genome seldom do (**Supplementary Figs. 2** and **3** online). Of 7,480 nonpredicted clusters, 1,999 ('new protein-coding') satisfied the criteria. Altogether, there were 5,416 protein-coding clusters (2,774 Ensembl-predicted, 643 *ab initio*–predicted and 1,999 new protein-coding clusters) among the 10,897 nonredundant FLJ-unique clusters ('protein-coding'). We found no ORFs meeting the criteria in the remaining 5,481 nonredundant FLJ-unique clusters (see also **Supplementary Note** and **Supplementary Figs. 7** and **8** online).

Non-protein-coding FLJ-unique cDNA clusters

The length distribution of the cDNAs in the 5,481 clusters that were categorized as non-protein-coding was similar to that of the other clusters (see also **Supplementary Fig. 4** online). Although some of these cDNAs might be cloning artifacts, recent evidence indicates that unexpectedly large populations of non-protein-coding transcripts exist in mammalian cells[16,17]. Genomic alignments of these clusters clearly showed that splicing had occurred for 1,378 of the 5,481 clusters. This splicing can be taken as explicit evidence that these cDNAs were derived from transcripts. We further examined whether these cDNAs contained any protein-coding-like sequences using GENSCAN, because it was possible that ORFs might be disrupted by retained introns, and because GENSCAN can identify interrupted ORFs in cDNA sequences. GENSCAN detected no coding exon–like regions in at least 850 clusters, ruling out the possibility that the retained introns disrupted the ORFs in these clusters.

To characterize the 850 non-coding-transcript clusters, we carried out BLAST searches using our one-pass sequence database (**Supplementary Table 1** online), which contains many cDNA sequences derived from various tissues. On average, the BLAST search resulted in 3.7 one-pass-sequence matches per cDNA. Many cDNAs matched a few one-pass sequences, indicating their low expression levels, and 44 clusters matched >10 one-pass sequences. By determining from which cDNA library the matching one-pass sequences were derived, we were able to identify the tissue distribution patterns of the expression of these genes. Some cDNAs had ubiquitous expression patterns, whereas others had very tissue-specific patterns. For 22 cDNAs, >20% of the matching one-pass sequences were derived from a particular cDNA library (examples are shown in **Supplementary Fig. 4** online).

GC content of new protein-coding cDNAs

To determine why 1,999 new protein-coding clusters in our FLJ-unique cDNAs were not predicted in Ensembl or by other computational methods, we compared the GC content of cDNAs between new protein-coding and 'predicted' (Ensembl-predicted and *ab initio*–predicted) clusters. The GC content of RefSeq mRNAs ranged broadly between 30% and 70% with two peaks, one at ~42% and the other at ~58% (**Fig. 2a**). The GC content of predicted and new protein-coding cDNAs showed similar broad distributions, but with only one peak each: at 42% for predicted and 58% for nonpredicted cDNAs. This suggests that there is some prediction bias against GC-rich transcripts in current gene prediction procedures. This can be seen also in the distribution of GC content of Ensembl genes. In contrast to RefSeq mRNAs, the peak at ~58% for all Ensembl genes is less pronounced.

As introns are generally more AT-rich than exons[4], the distributions of GC content of the corresponding genomic regions for both new protein-coding and predicted cDNAs shifted towards being more AT-rich. But overall patterns showed similar tendencies as cDNAs (**Fig. 2b**). Thus, current gene prediction procedures may have slight bias against predicting genes in GC-rich regions. This contradicts previous observations that the accuracy of these gene prediction methods is insensitive to the GC content (or is better in GC-rich regions)[18,19]. The discrepancy is probably caused by the fact that previous analyses were

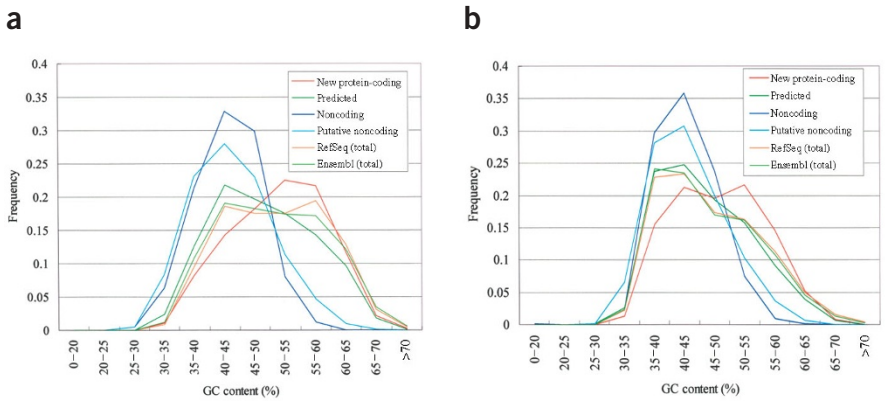

Figure 2 GC contents of the FLJ cDNAs and the corresponding genomic regions to which they were mapped. (**a**) GC contents of the new protein-coding cDNAs, non-protein-coding cDNAs, RefSeqs and Ensembl transcripts are shown. (**b**) GC contents of the genomic regions to which the corresponding transcripts were mapped (from the 5′ ends of first exons to the 3′ ends of the last exons) are shown for each category of the transcripts. For the detailed protocol for the chromosomal assignments of the FLJ cDNAs, please refer to **Supplementary Note** online (section on length distribution and chromosomal assignments of the 10,897 nonredundant FLJ-unique clusters section). Chromosomal positions of RefSeqs and Ensembl genes are as presented at University of California Santa Cruz genome browser.

based on a smaller data set and may not be extrapolated accurately to the full-genome scale.

We also calculated distribution of the GC content of noncoding and putative noncoding cDNAs (**Fig. 2**). To our surprise, both types of cDNA were relatively AT-rich (**Fig. 2a**). The new protein-coding and putative noncoding cDNAs were originally grouped together as 'non-predicted' and later separated according to the criteria of ORF length >100 codons and coding potential >30 (**Fig. 1**). Thus, we anticipated that these two categories might have similar GC content distributions. Instead, the GC content distributions of both putative noncoding and noncoding cDNAs had a peak at ~42%, similar to that of the predicted clones, but the range was much narrower. The GC content of the genomic regions where those cDNAs were mapped showed a similar AT-rich tendency (**Fig. 2b**). This raised the possibility that the noncoding cDNAs and the new protein-coding cDNAs are mainly transcribed from different regions of the human genome.

Annotation of protein-coding FLJ-unique cDNAs

For the 5,416 protein-coding clusters, we determined their amino acid sequences from the corresponding cDNA sequences. From the 1,999 new protein-coding cDNAs, we removed cDNAs that seemed to be derived from possibly truncated or immature forms of mRNA (T. Nishikawa *et al.*, unpublished data; see also **Supplementary Note** online), leaving 1,872 cDNAs ('full new protein-coding'). Thus, a total of 5,289 clusters were subjected to further analysis. The average ORF length was 335 codons (see **Supplementary Fig. 1** online for the distribution of the ORF length).

Using the amino acid sequences, we searched the protein motif databases PROSITE (version 16) and Pfam (version 5.5). In the PROSITE search, we found that 1,318 (25%) of these cDNAs had some protein signature (**Table 2**). In the Pfam search, we found that 3,112 (59%) of the cDNAs had some Pfam motif(s). In total, 1,529 kinds of Pfam motif, corresponding to 63% of the total 2,478 Pfam motifs, were represented. We also searched for putative membrane proteins and secretory proteins using SOSui and PSORT II, respectively (**Table 3**), which predicted 244 cDNAs for secretory proteins and 848 cDNAs for membrane proteins. Detailed functional annotations for each of these cDNAs are available at our websites[20,21].

Confirmation of transcripts from chromosomes 20–22

We mapped 45, 16 and 39 clusters of protein-coding transcripts to chromosomes 20, 21 and 22, respectively (**Supplementary Table 2** online). These chromosomes are 'finished' chromosomes, whose initial annotations are considered complete. As shown in **Table 4**, 491 of 10,897 nonredundant FLJ-unique clusters mapped to these chromosomes. Of these 491, 268 were protein-coding cDNAs, including 100 full new protein-coding cDNAs. There were 2,188 previously predicted genes (Ensembl genes and *ab initio*–predicted genes). Of these, 1,004 were experimentally identified cDNAs registered in RefSeq; thus, our analyses of FLJ cDNAs confirmed the presence of 168 (14%) of the 1,184 genes that had been predicted without full-length cDNA support and also identified an additional 100 protein-coding genes. This result emphasizes that full-length cDNA data should contribute to the precise annotation of the human genome.

Of these 100 full new protein-coding clusters, we experimentally confirmed the expression of 84 (84%) by RT-PCR using eight kinds of human tissue (**Supplementary Fig. 5** online). For the other cDNAs, we observed no clear band under our experimental conditions, perhaps

Table 2 Functional categorization and distribution of the protein motifs of the 5,289 protein-coding clusters

Functional categorization (PROSITE)	Number of matched clones
DNA or RNA associated proteins	553 (182)
Enzymes	198 (71)
Transport proteins	102 (38)
Structural proteins	69 (19)
Receptors	42 (14)
Post-translational modifications	10 (2)
Protein secretion and chaperones	9 (4)
Hormones and active peptides	4 (1)
Inhibitors	1 (1)
Others and domains	330 (105)
Motif (Pfam)	**Number of matched clones**
Zinc finger, C2H2 type	523 (71)
WD domain, G-beta repeat	106 (15)
Ank repeat	98 (10)
Leucine-rich repeat	88 (11)
Immunoglobulin domain	78 (9)
KRAB box	59 (15)
PHD-finger	58 (14)
Transcription factor S-II (TFIIS)	49 (5)
Zinc finger, C3HC4 type (RING finger)	38 (9)
Kelch motif	36 (18)
TPR domain	34 (6)
EF hand	26 (1)
Eukaryotic protein kinase domain	23 (5)
Fibronectin type III domain	22 (3)
BTB/POZ domain	20 (7)
TRAF-type zinc finger	19 (1)
RNA recognition motif. (RRM, RBD or RNP domain)	14 (5)
SH3 domain	11 (3)
Ras family	9 (2)
Others	1,801 (424)

PROSITE and Pfam motif databases were searched using GoodMotif[16] and the amino acid sequences deduced from the 5,289 protein-coding clusters. Matches for new protein-coding clusters are shown in parentheses.

Table 3 Predicted secretory proteins and membrane proteins

Prediction (program used)		Number of matched clones	
Secretory protein (PsortII)		244	(111)
	Number of transmembrane domains	Number of matched clones	
Transmembrane protein (SOSui)	1	373	(146)
	2	207	(94)
	3	74	(30)
	4	53	(22)
	5	33	(15)
	6	23	4)
	7	20	(7)
	8	20	(10)
	9	15	(2)
	10	16	(8)
	11	7	(3)
	12	4	(1)
	13	3	(1)
	Total	848	(343)

Secretory proteins and membrane proteins were predicted using PSORT II and SOSui, respectively. For each result, the number of matched cDNAs is shown. Matches for full new protein-coding cDNAs are shown in parentheses.

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

Table 4 FLJ cDNAs mapped to chromosomes 20, 21 and 22

			FLJ			
Chromosome	Ensembl-predicted and *ab initio* predicted	RefSeq	Mapped	Protein-coding	Full new protein-coding	Noncoding
20	940	456	222	114	45	20
21	439	183	89	44	16	11
22	809	365	180	110	39	12
Total	2,188	1,004	491	268	100	43

because their expression levels were too low or their expression was highly specific for the tissues or operation materials from which we isolated the RNAs to construct the cDNA libraries.

We also attempted RT-PCR analysis of the noncoding transcripts that mapped to chromosomes 20–22. We detected PCR bands for 32 of these clusters (74%) and observed both ubiquitous and tissue-specific expression patterns (**Supplementary Fig. 5** online). There seem to be at least hundreds of potential non-protein-coding transcripts in the human genome, some of which are expressed ubiquitously and others in a tissue-specific manner.

DISCUSSION

Here we describe the complete sequencing and characterization of 21,243 cDNA clones of our FLJ cDNA collection (the results of statistical analyses are summarized in **Supplementary Table 3** online). Of these, the full-length sequences of 14,490 cDNA clones or 10,876 cDNA clusters were unique to the FLJ project. Within the FLJ-unique clusters, 5,416 seemed to be protein-coding, and these included 2,774 Ensembl-predicted, 643 *ab initio*–predicted and 1,999 non-predicted protein-coding genes.

About two-thirds of the protein-coding genes (3,417 of 5,416) were predicted by computational methods. The total number of Ensembl-predicted genes is currently about 29,000, of which about 15,000 have been confirmed by fully sequenced cDNAs (according to RefSeq as of 1 August 2002). Here we confirmed an additional 2,774 such genes by identifying fully sequenced cDNA clusters (Ensembl-predicted). This corresponds to ~20% of the 14,000 Ensembl genes that lacked full-length cDNA support. The overlap was not as large as we expected, considering the scale of our project and the fact that Ensembl uses comprehensive evidence of protein homology from various organisms. One reason for this may be that cDNAs derived from long mRNAs were under-represented in the cDNA libraries used in our project. There is a cDNA sequencing project, KIAA, in which considerable effort is aimed at collecting cDNAs of long transcripts[10]. Such projects, as well as technical development, will be needed to collect cDNAs of long mRNAs. Another reason for the limited overlap may be the limited repertory of our cDNA libraries. Although we analyzed more than 100 different cDNA libraries, we might have missed transcripts whose expression is limited to small organs or rare cell types.

About one-third of protein-coding cDNAs (1,999 of 5,416) were not predicted by Ensembl or by *ab initio* gene predictions. We found that these new protein-coding cDNAs are relatively GC-rich. Thus, gene prediction methods that are currently in use may have some bias against GC-rich transcripts. Consistent with this finding, we found that the GC content distribution of RefSeq entries has two peaks at ~42% and ~58%, and that the peak at ~58% becomes insignificant when analyzing all Ensembl genes. This supports the suggestion that Ensembl genes that lack RefSeq support tend to be AT-rich. Bernardi proposed that the human genome could be divided into five different GC compositional categories (L1, L2 (AT-rich regions), H1, H2 and H3 (GC-rich regions)) and that L1 and L2 comprise ~70% of the human genome[22]. He predicted that GC-rich isochores, especially H3 (GC >48%), are gene-rich and that ~70% of genes are in the GC-rich region (GC >42%). Recent analysis using human genome draft sequence confirmed that known genes were rich in GC-rich regions (~70%), although the length and distribution of GC-rich regions and AT-rich regions vary widely in the human genome[4,5]. The estimate based on the gene prediction suggests that ~50% rather than ~70% of genes are present in GC-rich regions[5]. Our results suggest that there may be more genes to discover in GC-rich regions of the human genome. The new protein-coding cDNAs may be a good training set for improving the gene prediction methods.

In addition to protein-coding genes, we observed a number of putative noncoding clusters (5,481 clusters) among our nonredundant FLJ-unique clusters. About 1,300 of them were derived from spliced transcripts and 850 of them ('noncoding') contained no predicted exons. RT-PCR showed that at least some of them are transcribed *in vivo* (**Supplementary Fig. 5** online). Thus, we consider that most noncoding cDNAs come from real transcripts. The GC content of these cDNAs and the genomic regions in which they were mapped were in the range of low-GC regions. About 65% of those noncoding transcripts are transcribed from genomic regions that were low-GC regions and more than 5 kb upstream or downstream of any RefSeq- or Ensembl-predicted genes (data not shown). We categorized the remaining putative noncoding clusters as 'unclassified'. At present, we think that most of these unclassified cDNAs came from transcripts or their fragments rather than genomic DNA contaminants. For more detailed discussion of this issue, see **Supplementary Note** online.

Several large-scale projects for the systematic collection and complete sequencing of human and mouse cDNAs are underway[23–25]. The cDNAs identified in the present study, together with those from other projects, should produce a nearly complete physical collection of full-length cDNAs for human genes and those of important model organisms. For analysis of the proteome, the cDNAs are being transferred to various types of expression vectors. Recombinant proteins are being expressed with fusion tags and used for the systematic purification and identification of protein complexes[26,27]. Projects aimed at the large-scale determination of the three-dimensional structures of proteins have also been initiated based on the full-length cDNA resources[28]. Comprehensive analysis of the genome, transcriptome and proteome will lead to a better understanding of the architecture of life.

METHODS

Construction and characterization of the libraries. We constructed the cDNA libraries used in this project as previously described[9,10]. To evaluate the frequency of the full-length cDNAs in each of the libraries, we carried out BLAST searches with the cut-off value of e-100 and examined the relative positions of the 5′ end of the oligo-capped cDNAs compared with the RefSeq data, using all the one-pass sequences produced from each of the libraries. When the one-pass sequences covered the annotated coding-sequence start sites, we categorized them as 'full or near-full'. We applied this criterion because, in many cases, transcriptional start sites could not be sharply defined owing to possible slippery transcriptional start events[29], which made it difficult to determine exactly which cDNAs should be more specifically categorized as either 'full' or 'near-full'. The proportion of full-length sequences was calculated as the frequency of the full or near-full sequences among the cDNAs that had matches in RefSeq. The cDNAs that did not

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

match around the 5′ ends, possibly due to either alternative splicing or sequencing errors, were excluded from the calculation. We also excluded cDNAs in RefSeq that were derived from the FLJ project.

BLAST searches. As those that matched RefSeq sequences, the cDNAs containing at least the annotated coding sequence start sites were categorized as 'full at the 5′ end'. The RefSeq records containing 'FLJ' in the description field were excluded, because 'FLJ' indicated that the FLJ cDNA data were essential for experimental identification of the complete cDNAs of these genes, whose presence was otherwise only predicted based on the homology or partial cDNAs. For re-BLAST searching of the total 21,243 cDNA sequences against RefSeq, we used the BLAST score 1,000 for the cut-off. To further remove the redundancy from 14,490 FLJ-unique cDNAs, we carried out pairwise BLAST searches with a cut-off score of 1.0e-100. When we observed redundancy, we selected the cDNA with the longer 5′ end as representative (10,897 nonredundant FLJ-unique cDNAs). For further details on bioinformatics procedures and supporting evidence, see **Supplementary Note** online.

Ensembl prediction and *ab initio* gene predictions. We obtained Ensembl genes (using version 4.28.1; 29,076 genes) from the Ensembl website. We searched the sequences of 'Homo_sapiens.cdna.fa', which correspond to the predicted cDNA sequences, using our full-length cDNA sequences by BLAST with a cut-off score of 1.0e-100. Ensembl genes contained in RefSeq were excluded. Gene prediction programs were run against the human genomic sequence data (University of California Santa Cruz genome browser) with the default cut-off values. We used predicted exons to generate virtual cDNA sequences and did BLAST searches against them using FLJ cDNA sequences with a cut-off score of 1.0e-100.

URLs. GenBank, http://www.ncbi.nlm.nih.gov/Sitemap/index.html#GenBank; FLJ-DB, http://fldb.hri.co.jp/cgi-bin/cDNA3/public/publication/index.cgi; Ensembl, http://www.ensembl.org/Homo_sapiens/; PROSITE, http://www.expasy.ch/prosite/; Pfam, http://www.sanger.ac.uk/Pfam/; SOSui, http://sosui.proteome.bio.tuat.ac.jp/about-sosui.html/; PSORT II, http://psort.ims.u-tokyo.ac.jp/; DIGIT, http://digit.ims.u-tokyo.ac.jp; FGENESH, http://www.softberry.com/berry.phtml; GENSCAN, http://genes.mit.edu/GENSCAN.html; HMMGENES, http://www.cbs.dtu.dk/services/HMMgene/; HUNT database, http://www.hri.co.jp/HUNT/; HUGE database, http://www.kazusa.or.jp/huge/; University of California Santa Cruz genome browser, http://genome.ucsc.edu/;NEDO, http://www.nedo.go.jp/bio/index.html.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank A. Kishimoto, H. Ezoe and T. Matsuo for supporting the project and E. Nakajima for critically reading the manuscript. This project was supported by the Ministry of Economy Trade and Industry of Japan and also in part by Special Coordination Funds for Promoting Science and Technology from the Ministry of Education, Culture, Sports, Science and Technology of Japan. Requests for materials should be addressed to S. Sugano. Requests for physical cDNA clones should be addressed to S.Sugano (flcdna@ims.u-tokyo.ac.jp) or T. Isogai (isogai-t@reprori.jp). For more information on each cDNA clone, visit FLJ-DB. For general information on the FLJ project, please refer to NEDO website.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Received 7 October; accepted 1 December 2003
Published online at http://www.nature.com/naturegenetics/

1. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
2. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
3. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
4. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
6. Pruitt, K.D. & Maglott, D.R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
7. Boguski, M.S. The turning point in genome research. *Trends Biochem. Sci.* **20**, 295–296 (1995).
8. Maruyama, K. & Sugano, S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**, 171–174 (1994).
9. Suzuki, Y., Yoshitomo, K., Maruyama, K., Suyama, A. & Sugano, S. Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene* **200**, 149–156 (1997).
10. Nomura, N. *et al.* Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001-KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1. *DNA Res.* **1**, 27–35 (1994).
11. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
12. Giesecke, H., Obermaier, B., Domdey, H. & Neubert, W.J. Rapid sequencing of the Sendai virus 6.8 kb large (L) gene through primer walking with an automated DNA sequencer. *J. Virol. Methods.* **38**, 47–60 (1992).
13. 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).
14. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
15. Fickett, J.W. Predictive methods using nucleotide sequences. *Methods Biochem. Anal.* **39**, 231–245 (1998).
16. Huttenhofer, A. *et al.* RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.* **20**, 2943–2953 (2001).
17. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
18. Burset, M. & Guigo, R. Evaluation of gene structure prediction programs. *Genomics* **34**, 353–367 (1996).
19. Rogic, S., Mackworth, A.K. & Ouellette, F.B. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11**, 817–832 (2001).
20. Yudate, H.T. *et al.* HUNT: launch of a full-length cDNA database from the helix research institute. *Nucleic Acids Res.* **29**, 185–188 (2001).
21. Hattori, A. *et al.* Characterization of long cDNA clones from human adult spleen. *DNA Res.* **7**, 1–11 (2001).
22. Bernardi, G. The isochore organization of the human genome and its evolutionary history—a review. *Gene.* **135**, 57–66 (1993).
23. The FANTOM consortium and The RIKEN Genome Exploration Research Group Phase I & II team. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
24. Wiemann, S. *et al.* Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* **11**, 422–435 (2001).
25. Strausberg, R.L., Feingold, E.A., Klausner, R.D. & Collins, F.S. The mammalian gene collection. *Science* **286**, 455–457 (1999).
26. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
27. Ho, Y. *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180–183 (2002).
28. Chance, M.R. *et al.* Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci.* **11**, 723–738 (2002).
29. Suzuki, Y, *et al.* Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.* **2**, 388–393 (2001).

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics