Table 3. Statistics Obtained from the Functional Annotation Results

	Category	Number of Loci
H-Inv proteins	I. Identical to a known human protein	5,074
	II. Similar to a known protein	4,104
	III. InterPro domain containing protein	2,531
	IV. Conserved hypothetical protein	1,706
	V. Hypothetical protein	6,159
	Total number of H-Inv proteins	19,574
Non-protein-coding transcripts	Putative ncRNA	296
	Uncharacterized transcript	675
	Unclassifiable	329
	Hold	77
	Total number of non-protein-coding transcripts	1,377
Questionable transcripts		86
Total number of H-Inv loci		21,037

DOI: 10.1371/journal.pbio.0020162.t003

gene products that are related to known proteins: 5,074 (25.9%) were defined as identical to a known human protein (Category I proteins); 4,104 (21.0%) were defined as similar to a known protein (Category II proteins); and 2,531 (12.9%) as domain-containing proteins (Category III proteins). In total, we were able to assign biological function to 59.9% of H-Inv proteins by similarity or motif searches. The remaining proteins, for which no biological functional was inferred, were annotated as conserved hypothetical proteins (Category IV proteins; 1,706, 8.7%) if they had a high level of similarity to other hypothetical proteins in other species, or as hypothetical proteins (Category V proteins; 6,159, 31.5%) if they did not.

To predict the functions of hypothetical proteins (Category IV and V proteins), we used 196 sequence patterns of functional importance derived from tertiary structures of protein modules, termed 3D keynotes (Go 1983; Noguti et al. 1993). Application of the 3D keynotes to the H-Inv proteins resulted in the prediction of functions in 350 hypothetical proteins (see Protocol S1).

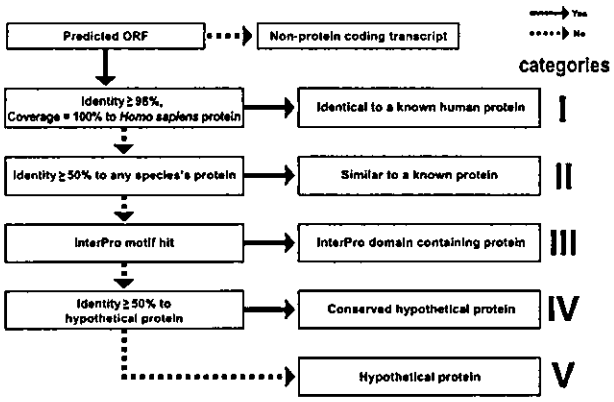

Figure 4. Schematic Diagram of Human Curation for H-Inv Proteins

The diagram illustrates the human curation pipeline to classify H-Inv proteins into five similarity categories; Category I , II, III, IV, and V proteins.

DOI: 10.1371/journal.pbio.0020162.g004

Features of ORFs deduced from human FLcDNAs. The mean and median lengths of predicted ORFs were calculated for the 19,574 H-Inv proteins. These were 1,095 bp and 806 bp, respectively (Table 4). The values obtained were smaller than those from other eukaryotes, and are inconsistent with estimates reported previously (Shoemaker et al. 2001). However, as has been seen in the earlier annotation of the fission yeast genome (Das et al. 1997), our dataset might contain stretches which mimic short ORFs. This would lead to a bias in our ORF prediction and result in an erroneous estimate of the average ORF length. We examined the size distributions of ORFs from the five categories, and found that the distribution pattern was quite similar across categories. The exception was Category V, in which short ORFs were unusually abundant (Figure S3). Judging from the length distribution of ORFs in the five categories of H-Inv proteins, the majority of ORFs shorter than 600 bps in Category V seemed questionable. In order to have a protein dataset that contains as many sequences to be further analyzed as possible, we have taken the longest ORFs over 80 amino acids if no significant candidates were detected by the sequence similarity and gene prediction (see Figure S1). The consequence of this is that Category V appears to contain short questionable ORFs, a certain fraction of which may be prediction errors. Nevertheless, these ORFs could be true. It is also possible that those ORFs were in fact translated in vivo when we curated the cDNAs manually. The existence of many functional short proteins in the human proteome is already confirmed, and there are 199 known human proteins that are 80 amino acids or shorter in the current Swiss-Prot database. We think that the H-Inv hypothetical proteins require experimentally verification in the future. Excluding the hypothetical proteins from the analysis, we obtained mean and median lengths for the ORFs of 1,368 bp and 1,130 bp, respectively, which are reasonably close to those for other eukaryotes (Table 4).

Of the 4,104 Category II proteins, 3,948 proteins (96.2%) were similar to the functionally identified proteins of

Table 4. The Features of Predicted ORFs

	Number of ORFs	Mean (bp)	Median (bp)	Percent GC of Third Codon Position
Human—H-Inv datasets (categories I–IV)	13,415	1,368	1,130	52.3
Human—all of the H-Inv datasets	19,574	1,095	806	52.4
Fly	17,878	1,580	1,212	53.9
Worm	21,118	1,327	1,038	42.9
Budding yeast	6,408	1,403	1,128	40.3
Fission yeast	4,968	1,426	1,161	39.7
Plant	27,228	1,269	1,074	44.2
Bacteria	4,289	951	834	51.9

Nonredundant proteome datasets of nonhuman species were obtained from the following URLs: fly (*Drosophila melanogaster*; http://flybase.bio.indiana.edu/), worm (*Caenorhabditis elegans*; http://www.wormbase.org/), budding yeast (*Saccharomyces cerevisiae*; http://www.pasteur.fr/externe), fission yeast (*Schizosaccharomyces pombe*; http://www.sanger.ac.uk/), plant (*Arabidopsis thaliana*; http://mips.gsf.de/proj/thal/index.html), and bacteria (*Escherichia coli* K12; http://www.ncbi.nlm.nih.gov/).
DOI: 10.1371/journal.pbio.0020162.t004

mammals (Figure S4). This implies that the predicted functions in this study were based on the comparative study with closely related species, so that the functional assignment retains a high level of accuracy if we suppose that protein function is more highly conserved in more closely related species. Moreover, the patterns of codon usage and the codon adaptation index (CAI; http://biobase.dk/embossdocs/cai.html) of H-Inv proteins were investigated (Table S2). The results indicated that the ORF prediction scheme worked equally well in the five similarity categories of H-Inv proteins.

Each H-Inv protein in the five categories was investigated in relation to the tissue library of origin (Table S3). We found that at least 30% of the clones mainly isolated from dermal connective, muscle, heart, lung, kidney, or bladder tissues could be classified as Category I proteins. Hypothetical proteins (Category V), on the other hand, were abundant in both endocrine and exocrine tissues. This bias may indicate that expression in some tissues may not have been studied in enough detail. If this is the case, then there is likely a significant gap between our current knowledge of the human proteome and its true dimensions.

Non-protein-coding genes. Over recent years, ncRNAs have been found to play key roles in a variety of biological processes in addition to their well-known function in protein synthesis (Moore and Steitz 2002; Storz 2002). Analysis of the H-Inv cDNA dataset revealed that 6.5% of the transcripts are possibly non-protein-coding, although the number is much smaller than that estimated in mice (Okazaki et al. 2002). We believe that this difference between the two species is mainly due to the larger number of mouse libraries that were used and to a rare-transcript enrichment step that was applied to these collections.

To identify ncRNAs, we manually annotated 1,377 representative non-protein-coding transcripts, which were classified into four categories (see Table 3; Figure 5): putative ncRNAs, uncharacterized transcripts (possible 3′ UTR fragments supported by ESTs), unclassifiable transcripts (possible genomic fragments), and hold transcripts (not stringently mapped onto the human genome). Of these, 296 (19.5%) were putative ncRNAs with no neighboring transcripts in the close vicinity (> 5 kb) and supported by ESTs with a poly-A signal or a poly-A tail, indicating that these may represent genuine ncRNA genes. On the other hand, a large fraction of the non-protein-coding transcripts (675; 44.5%) were classified as possible 3′ UTRs of genes that were mapped less than 5 kb upstream. The 5-kb range is an arbitrary distance that we defined as one of our selection criteria for identifying ncRNAs. However, authentic non-protein-coding genes might be located adjacent to other protein-coding genes (as described earlier). Thus, some of the transcripts initially annotated as uncharacterized ESTs may correspond to ncRNAs when these sequences satisfy the other selection criteria.

We defined a manual annotation strategy (Figure 5) that allowed us to select convincing putative ncRNAs with various

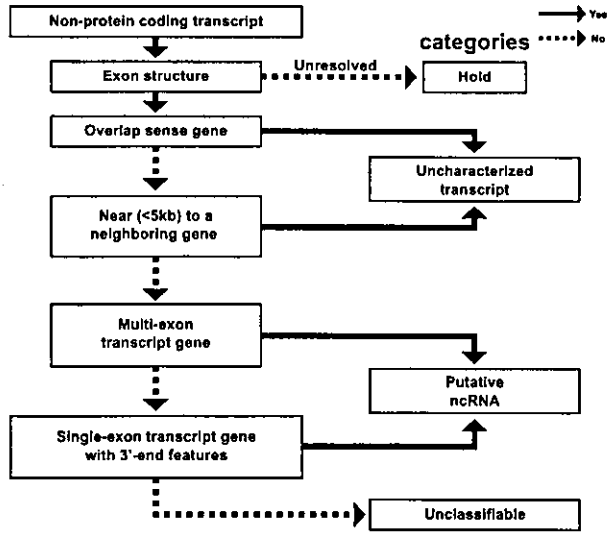

Figure 5. The Manual Annotation Flow Chart of ncRNAs

Candidate non-protein-coding genes were compared with the human genome, ESTs, cDNA 3′-end features and the locus genomic environment. The candidates were then classified into four categories: hold (cDNAs improperly mapped onto the human genome); uncharacterized transcripts (transcripts overlapping a sense gene or located within 5 kb of a neighboring gene with EST support); putative ncRNAs (multiexon or single exon transcripts supported by ESTs or 3′-end features); and unclassifiable (possible genomic fragments).
DOI: 10.1371/journal.pbio.0020162.g005

lines of supporting evidence. These are the following: absence of a neighboring gene in the close vicinity, overlap with human or mouse ESTs, occurrence in the 3′ end of cDNA sequences, as well as overlap with mouse cDNAs. Out of 296 annotated putative ncRNAs, we identified 47 ncRNAs with conserved RNA secondary structure motifs (Rivas and Eddy 2001), and nearly 60% of these were found expressed in up to eight human tissues (data not shown), indicating that the manual curation strategy employed in this study may facilitate the identification of novel non-protein-coding genes in other species.

The functions of human proteins identified through an analysis of domains. Proteins in many cases are composed of distinct domains each of which corresponds to a specific function. The identification and classification of functional domains are necessary to obtain an overview of the whole human proteome. In particular, the analysis of functional domains allows us to elucidate the evolution of the novel domain architectures of genes that life forms have acquired in conjunction with environmental changes. The human proteome deduced from the H-Inv cDNAs was subjected to InterProScan, which assigned functional motifs from the PROSITE, PRINTS, SMART, Pfam, and ProDom databases (Mulder et al. 2003). A total of 19,574 H-Inv proteins were examined, and 9,802 of them (50.1%) were assigned at least one InterPro code that was classified into either repeats (a region that is not expected to fold into a globular domain on its own), domains (an independent structural unit that can be found alone or in conjunction with other domains or repeats), and/or families (a group of evolutionarily related proteins that share one or more domains/repeats in common) when compared with those of fly, worm, budding and fission yeasts, *Arabidopsis thaliana,* and *Escherichia coli* (Table S4). Moreover, the proteins were classified according to the Gene Ontology (GO) codes that were assigned to InterPro entries (Table S5).

Identification of human enzymes and metabolic pathways. One of the most important goals of the functional annotation of human cDNAs is to predict and discover new, previously uncharacterized enzymes. In addition, revealing their positions in the metabolic pathways helps us understand the underlying biochemical and physiological roles of these enzymes in the cells. We thus searched for potential enzymes among the H-Inv proteins, and mapped them to a database of known metabolic pathways.

We could assign 656 kinds of potential Enzyme Commission (EC) numbers to 1,892 of the 19,574 H-Inv proteins based on matches to the InterPro entries and GO assignments and on the similarity to well-characterized Swiss-Prot proteins (see Dataset S2). The number of characterized human enzymes significantly increased through this analysis. The most abundant enzymes in the H-Inv proteins were protein-tyrosine kinases (EC 2.7.1.112), which is consistent with the large number of kinases found in the InterPro assignments. The other major enzymes were small monomeric GTPase (EC 3.6.1.47), adenosinetriphosphatase (EC 3.6.1.3), phosphoprotein phosphatase (EC 3.1.3.16), ubiquitin thiolesterase (EC 3.1.2.15), and ubiquitin-protein ligase (EC 6.3.2.19). These enzymes are members of large multigene families that are important for the functions of higher organisms. Furthermore, we could assign 726 EC numbers to mouse representative transcripts and proteins (Okazaki et al. 2002), and most of them appeared to be shared between human and mouse (data not shown). The high similarity of the enzyme repertoire between these two species is not surprising if we consider the close evolutionary relatedness between them. It does, however, indicate the usefulness of the mouse as a model organism for studies concerning metabolism.

We then mapped all H-Inv proteins on the metabolic pathways of the KEGG database, a large collection of information on enzyme reactions (Kanehisa et al. 2002). In total, we mapped 963 H-Inv proteins on a total of 1,613 KEGG pathways, of which 641 were based on their EC number assignments (Figure S5). Those based on EC number assignments do not necessarily function as they are assigned because they have yet to be verified experimentally. However, if all other enzymes along the same pathway exist in humans, the functional assignment has a high probability of being correct. Using this method, we discovered a total of 32 newly assigned human enzymes from the H-Inv proteins with the support of KEGG pathways (Table S6). For example, we identified (1) pyridoxamine-phosphate oxidase (EC 1.4.3.5; AK001397), an enzyme in the "salvage pathway," the function of which is the reutilization of the coenzyme pyridoxal-5′-phosphate (its role in epileptogenesis was recently reported [Bahn et al. 2002]), (2) ATP-hydrolysing 5-oxoprolinase (EC 3.5.2.9; AL096750) that cleaves 5-oxo-L-proline to form L-glutamate (whose deficiency is described in the Online Mendelian Inheritance in Man [OMIM] database [ID = 260005]), and (3) N-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25; BC018734), which catalyzes N-acetylglucosamine at the second step of its catabolism, the activity of which in human erythrocytes was detected by a biochemical study (Weidanz et al. 1996). Many of the newly identified enzymes were supported by currently available experimental and genomic data. An example is a putative urocanase (EC 4.2.1.49; AK055862) that mapped onto the "histidine metabolism" that urocanic acid catabolises. A ^{14}C Histidine tracer study unexpectedly revealed that NEUT2 mice deficient in 10-formyltetrahydrofolate dehydrogenase (FTHFD) excrete urocanic acid in the urine and lack urocanase activity in their hepatic cytosol (Cook 2001). We then found that both the FTHFD and AK055862 genes were located within the same NCBI human contig (NT005588) on Chromosome 3. Moreover, the distance between the two genes was consistent with the genetic deletion of NEUT2 (> 30 kb). We thus assumed that FTHFD and urocanase might be coincidentally defective in mice. This analysis could confirm that the AK055862 protein is a true urocanase. This example demonstrates that this kind of in silico analysis is a powerful method in defining the functions of proteins.

Polymorphism in the Transcriptome

Sites of potential polymorphism in cDNAs. Due to the rapidly increasing accumulation of genetic polymorphism data, it is necessary to classify the polymorphism data with respect to gene structure in order to elucidate potential biological effects (Gaudieri et al. 2000; Sachidanandam et al. 2001; Akey et al. 2002; Bamshad and Wooding 2003). For this purpose, we examined the relationship between publicly available polymorphism data and the structure of our H-Inv cDNA sequences. A total of 4 million single nucleotide polymorphisms (SNPs) and insertion/deletion length variations (indels) with mapping information from the Single

Table 5. The Numbers of SNPs and indels Occurring in the Representative cDNAs

		5′ UTR	Coding Region	3′ UTR
SNPs[a]	Synonymous		11,014(1/325 bp)	
	Nonsynonymous		13,215(1/1,206 bp)	
	Truncation[b]		315	
	Extension[b]		43	
	Synonymous SNP at stop codon		28	
	Total	10,715(1/569 bp)	24,679[c](1/833 bp)	31,852(1/536 bp)
Indels		381(1/15,999 bp)	452(1/45,490 bp)	1,364(1/12,553 bp)

[a]The numbers of SNPs and indels are summarized for representative cDNA sequences which were mapped on the genome. The numbers in parentheses represent the densities of SNPs and indels.
[b]SNPs that cause nonsense mutation or extension of polypeptides were classified assuming that the cDNAs represent original alleles.
[c]This figure includes 64 unclassifiable SNPs.
DOI: 10.1371/journal.pbio.0020162.t005

Nucleotide Polymorphism Database (dbSNP; http://www.ncbi.nlm.nih.gov/SNP/, build 117) (Sherry et al. 1999) were used for the search. We could identify 72,027 uniquely mapped SNPs and indels in the representative H-Inv cDNAs and observed an average SNP density of 1/689 bp. To classify SNPs and indels with respect to gene structure, the genomic coordinates of SNPs were converted into the corresponding nucleotide positions within the mapped cDNAs. The SNPs and indels were classified into three categories according to their positions: 5′ UTR, ORF, and 3′ UTR (Table 5). The density of indels was higher in 5′ UTRs (1/15,999 bp) and 3′ UTRs (1/12,553 bp) than in ORFs (1/45,490 bp). This is possibly due to different levels of functional constraints. We also examined the length of indels and found a higher frequency of indels in those ORFs that had a length divisible by three and that did not change their reading frames. We observed that the density of SNPs was higher in both the 5′ and 3′ UTRs (1/569 bp and 1/536 bp, respectively) than in ORFs (1/833 bp).

SNPs located in ORFs were classified as either synonymous, nonsynonymous, or nonsense substitutions (Table 5). We identified 13,215 nonsynonymous SNPs that affect the amino acid sequence of a gene product. At least 4,998 of these nonsynonymous SNPs are "validated" SNPs (as defined by dbSNP). This data can be used to predict SNPs that affect gene function. SNPs that create stop codons can cause polymorphisms that may critically alter gene function. We identified 358 SNPs that caused either a nonsense mutation or an extension of the polypeptide. We classified these 358 SNPs into these two types based on the alleles of the cDNA. Most of these SNPs (315/358) were predicted to cause truncation of the gene products and produce a shorter polypeptide compared with the alleles of H-Inv cDNAs. For example, Reissner's fiber glycoprotein I (AK093431) contains a nonsense SNP that results in the loss of the last 277 amino acids of the protein, and consequently the loss of a thrombospondin type I domain located in its C-terminal end. This SNP is highly polymorphic in the Japanese population, the frequencies of *G* (normal) and *T* (termination) being 0.43 and 0.57, respectively. As seen in this example, the identification of SNPs within cDNAs provides important insights into the potential diversity of the human transcriptome. Thus, polymorphism data crossreferenced to a comprehensively annotated human transcriptome might prove to be a valuable tool in the hands of researchers investigating genetic diseases.

Sites of microsatellite repeats. Among the 19,442 representative protein-coding cDNAs, we identified a total of 2,934 di-, tri-, tetra-, and penta-nucleotide microsatellite repeat motifs (Table 6). Interestingly, 1,090 (37.2%) of these were found in coding regions, the majority of which (86.9%) were tri-nucleotide repeats. Di-, tetra-, and penta-nucleotide repeats made up the greatest proportion of repeats in 5′ UTRs and 3′ UTRs. Coding regions contained mostly tri-

Table 6. The Numbers of Microsatellite Repeat Motifs That Occurred in the Representative cDNAs

	Microsatellite Repeats				
	Di-	Tri-	Tetra-	Penta-	Total
5′ UTR	162 (50)	394 (3)	117 (4)	21 (1)	694 (58)
Coding region	70 (13)	947 (10)	63 (2)	10 (0)	1,090 (25)
3′ UTR	482 (121)	340 (3)	281 (8)	47 (1)	1,150 (133)
Total	714 (184)	1,681 (16)	461 (14)	78 (2)	2,934 (216)

Microsatellites were defined as those sequences having at least ten repeats for di-nucleotide repeats and at least five repeats for tri-, tetra-, and penta-nucleotide repeats. Numbers of polymorphic microsatellites inferred by comparisons of cDNA and genomic sequences are shown in parenthesis. See Table S2 for a list of accession numbers for these cDNAs.
DOI: 10.1371/journal.pbio.0020162.t006

nucleotide repeats. This result is consistent with the idea that microsatellites are prone to mutations that cause changes in numbers of repeats. Only tri-nucleotide repeats can conserve original reading frames when extended or shortened by mutations. A previous study showed that many of the microsatellite motifs identified in human genomic sequences, including those in coding regions, are highly polymorphic in human populations (Matsuzaka et al. 2001). We found this to be the case in our study: 36 of the microsatellite repeats we detected were found to be polymorphic in human populations according to dbSNP records (data not shown). We identified 216 microsatellite repeats in 213 genes that showed contradictory numbers of repeats between cDNA and genome sequences (see Dataset S3). This figure includes 25 microsatellites in ORFs that have the potential to alter the protein sequences. Individual cases need to be verified by further experimental studies, but many of these microsatellites may really be polymorphic in human populations and have marked phenotypic effects.

There were 382 cDNAs that possessed two or more microsatellites in their nucleotide sequences. This is illustrated in RBMS1 (BC018951), a cDNA which encodes an RNA-binding motif. This cDNA has four microsatellites, $(GGA)_7$, $(GAG)_9$, $(GAG)_6$, and $(GCC)_6$, in its 5′ UTR. These microsatellites are all located at least 98 bp upstream of the start codon, but they could still have pronounced regulatory effects on gene expression. Another example is the cDNA that encodes CAGH3 (AB058719). This cDNA has four microsatellites, $(CAG)_8$, $(CAG)_6$, $(CAG)_8$, and $(CAG)_8$, all of which are located within the ORF. These microsatellites all encode stretches of poly-glutamine, which are known to have transcription factor activity (Gerber et al. 1994) and often cause neurodegenerative diseases when the number of repeats exceeds a certain limit. A typical example of a disorder caused by these repeats is Huntington's disease (Andrew et al. 1993; Duyao et al. 1993; Snell et al. 1993).

We also searched for repeat motifs containing the same amino acid residue in the encoded protein sequences. We located a total of 3,869 separate positions where the same amino acid was repeated at least five times. The most frequent repetitive amino acids are glutamic acid, proline, serine, alanine, leucine, and glycine. The glutamine repeats of this nature were found in 160 different locations.

Evolution of the Human Transcriptome

Beyond the study of individual genes, the comparison of numerous complete genome sequences facilitates the elucidation of evolutionary processes of whole gene sets. Moreover, the FLcDNA datasets of humans and mice give us an opportunity to investigate the genome-wide evolution of these two mammals by using the sequences supported by physical clones. Here we compared our human cDNA sequences with all proteins available in the public databases. Focusing on our results, we discuss when and how the human proteome may have been established during evolution. Furthermore, the evolution of UTRs is examined through comparisons with cDNAs from both primates and rodents.

Conserved and derived protein-coding genes in humans. An advantage of large-scale cDNA sequencing is that it can generate a nearly complete gene set with good evidence for transcription. The human proteome deduced from the FLcDNA sequences gives us an opportunity to decipher the evolution of the entire proteome. Here we compare the representative H-Inv cDNAs with the Swiss-Prot and TrEMBL protein databases using FASTY (Pearson 2000), and we describe the distributions of the homologs among taxonomic groups at two different similarity levels. The number of representative H-Inv cDNAs that have homolog(s) in a given taxon was counted (Figure S6), and the cDNAs were classified into functional categories (Figure 6). These results indicated that homologs of the human proteins were probably conserved much more in the animal kingdom than in the others at both moderate ($E < 10^{-10}$) and weak ($E < 10^{-5}$) similarity levels (see Figure S6). Moreover, human sequences had as many nonmammalian animal homologs as mammalian homologs, with seemingly little bias to any one function (see Figure 6). This suggests that the genetic background of humans may have already been established in an early stage of animal evolution and that many parts of the whole genetic system have probably been stable throughout animal evolution despite the seemingly drastic morphological differences between various animal species. This result is consistent with our previous observation that the distribution of the functional domains is highly conserved among animal species (see Table S4). The number of homologs may have been inflated by recent gene duplication events within the human lineage. Hence we counted the number of paralog clusters instead of cDNAs that had homologs in the databases, and obtained essentially the same results (Figure S7).

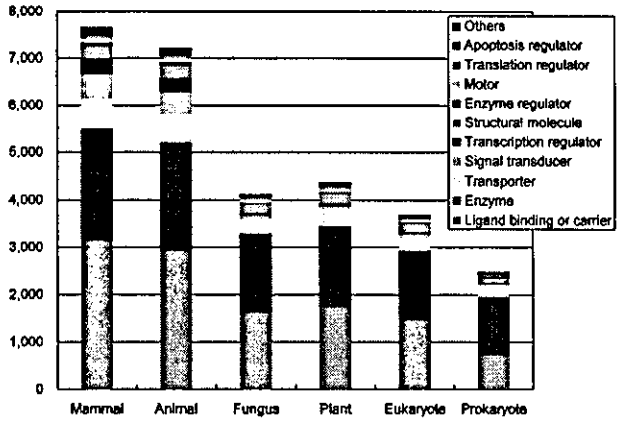

Figure 6. The Functional Classification of H-Inv Proteins That Are Homologous to Proteins in Each Taxonomic Group

The numbers of representative H-Inv cDNAs with sequence homology to other species' proteins ($E < 10^{-5}$) were calculated. The cDNAs for which we could not assign any functions were discarded. Mammalian species were excluded from the "animal" group. "Eukaryote" represents eukaryotic species other than those included in the mammal, animal, fungi, and plant groups. See also Table S7.
DOI: 10.1371/journal.pbio.0020162.g006

This analysis also revealed a number of potential human-specific proteins, which did not have any homologs in the current sequence databases. In this case the creation of lineage-specific genes through speciation is not completely excluded. However, most ORFs with no similarity to known proteins would not be genuine for the reasons discussed above. Therefore, the number of "true" human-specific proteins is expected to be relatively small.

We conducted further BLASTP searches matching entries from the Swiss-Prot database against the H-Inv dataset itself.

As a result, 12,813 (45.3%) of 28,263 vertebrate proteins had homologs in nonvertebrates at $E < 10^{-30}$. Taking into account that the dataset is relatively small (approximately 12,000 sequences) and as a result may be biased, animal species may conceivably share a similar protein-coding gene set.

Ohno (1996) proposed that the emergence of a large number of animal phyla in a short period of time would endow them with almost identical genomes. These were collectively referred to as the pananimalia genome. Our data support Ohno's hypothesis from the perspective that the basic gene repertoires of animals are essentially highly similar among diverse species that have evolved separately since the Cambrian explosion. Subsequently, morphological evolution seems to have been brought about mainly by changes in gene regulation. The number of transcription regulator homologs is different between animals and other phyla (Table S7). In this analysis it was not possible to examine the genes recently deleted from the human lineage. However, the similarity of the proteome sets between distantly related mammals such as human and mouse (Waterston et al. 2002) suggests that not many genes have been deleted specifically from humans since humans and mice diverged.

A unique feature of the Animalia proteome is, for example, the presence of apoptosis regulator homologs, which are found widely in the animal kingdom, whilst they are rare in the other phyla (Table S7). Since apoptosis plays an important role during the development of multicellular animals, this observation indicates that apoptosis was established independently of both plants and fungi during the early evolution of multicellularization in the kingdom Animalia. Likewise, signal transducers and cell-adhesion proteins are distinctive. In contrast, enzymes, translation regulators, molecular chaperones, etc. were highly conserved among all taxonomic groups. These proteins may have played such essential roles that any alterations were eliminated by strong purifying selection. It is assumed some functions were presumably derived from ancient endocellular symbionts (mitochondria and chloroplasts) (Martin 2002).

Evolution of untranslated regions. The UTRs of mRNA are known to be involved in the regulation of gene expression at the posttranscriptional level through control of translation efficiency (Kozak 1989; Geballe and Morris 1994; Sonenberg 1994), mRNA stability (Zaidi and Malter 1994; McCarthy and Kollmus 1995), and mRNA localization (Curtis et al. 1995; Lithgow et al. 1997). Only a few studies on very limited datasets have been carried out so far to describe quantitatively either the evolutionary dynamics of mRNA UTRs (Larizza et al. 2002), or their general structural and compositional features (Pesole et al. 1997). The human transcriptome presented here along with the murid data obtained mainly from the FANTOM2 project enables us to stabilize a mammalian genome perspective on the subject (Table S8). A sliding window analysis of UTR sequence identities between humans and mice revealed a positive correlation between the number of indels in an untranslated region and the distance from the coding sequence (Figure 7). Unlike indels, mismatches are distributed equally along whole untranslated regions. In other words, indels seem to be less tolerated in close proximity to a coding sequence, while substitutions are evenly distributed along the untranslated regions of the mRNAs. This seems to be a general pattern observed similarly in other species (data not shown). Indels in

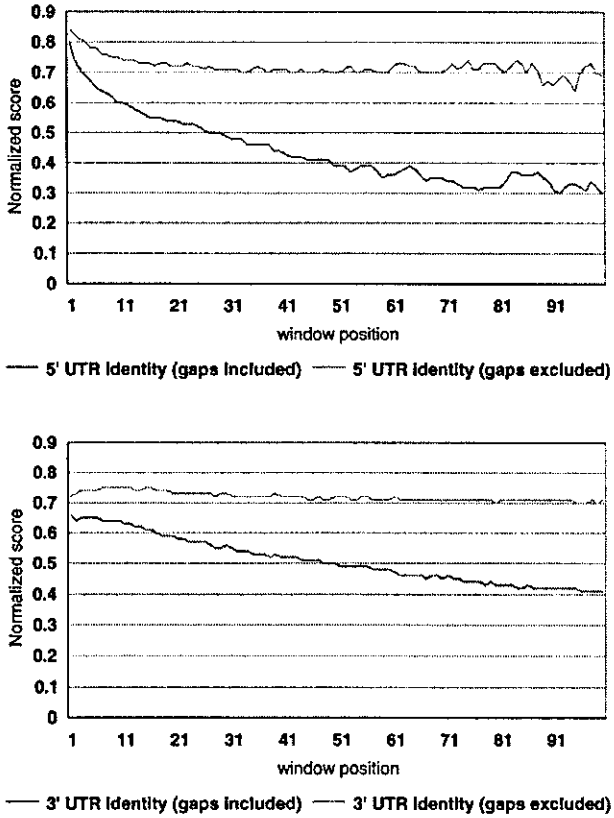

Figure 7. Window Analysis of Similarity between Human and Mouse UTRs

Results for 5′ UTRs presented above and for 3′ UTRs below. The whole mRNA sequences were aligned using a semiglobal algorithm as implemented in the map program (Huang 1994) with the following parameters: match 10, mismatch −3, gap opening penalty −50, gap extension penalty −5, and longest penalized gap 10; the terminal gaps are not penalized at all. A window size of 20 bp was used with a step of 10 bp. The analysis window was moved upstream and downstream of start and stop codons, respectively. The normalized score for a given window is calculated as a fraction of an average score for all UTRs in a given window over the maximum score observed in all 5′ or 3′ UTRs, respectively.

DOI: 10.1371/journal.pbio.0020162.g007

UTRs may have been avoided so that the distance between the coding region and a signal sequence for regulation in the UTR could be conserved throughout evolution, while purifying selection against substitutions appeared to be relatively weak.

Untranslated region replacement. A replacement of the entire UTR may lead to drastic changes in gene expression, especially if a UTR having a posttranscriptional signal is replaced by another. We compared the evolutionary distances of UTRs between primate and rodent orthologous sequences. We based our analysis on the UTR sequence distances that contradicted the expected phylogenetic tree of relatedness. We could detect 149 UTR replacements distributed among different species. Some of the observed replacements may result from selection of different AS isoforms of a single locus in different species. This is particularly likely if an AS event involves an alternative first or last exon. It seems that UTR replacements are more frequent in rodents than in primates, but the difference is not statistically significant at the 5%

significance level (Table S9). We detected a UTR replacement in less than 2% of the analyzed sequences. The evolutionary consequences could be significant because the UTR replacement might result in changes in expression level or the loss of an mRNA localization signal.

The H-Invitational Database

All the results of the mapping of the FLcDNA sequences onto the human genome, the clustering of FLcDNA sequences, sequence alignments, detection of AS transcripts, sequence similarity searches, functional annotation, protein structure prediction, subcellular localization prediction, SNP mapping, and evolutionary analysis, as well as the basic features of FLcDNA sequences, are stored in the H-InvDB (Figure S8). The H-InvDB is a unique database that integrates annotation of sequences, structure, function, expression, and diversity of human genes into a single entity. It is useful as a platform for conducting in silico data mining. The database has functions such as a keyword search, a sequence similarity search, a cDNA search, and a searchable genome browser. It is hoped that the H-InvDB will become a vital resource in the support of both basic and applied studies in the fields of biology and medicine.

We constructed two kinds of specialized subdatabases within the H-InvDB. The first is the Human Anatomic Gene Expression Library (H-Angel), a database of expression patterns that we constructed to obtain a broad outline of the expression patterns of human genes. We collected gene expression data from normal and diseased adult human tissues. The results were generated using three methods on seven different platforms. These included iAFLP (Kawamoto et al. 1999; Sese et al. 2001), DNA arrays (long oligomers, short oligomers [Haverty et al. 2002], cDNA nylon microarrays [Pietu et al. 1999], and cDNA glass slide microarrays [Arrays/IMAGE-Genexpress]), and cDNA sequence tags (SAGE [Velculescu et al. 1995; Boon et al. 2002], EST data [Boguski et al. 1993; Kawamoto et al. 2000], and MPSS [Brenner et al. 2000]). By normalizing levels of gene expression in experiments conducted with different methods, we determined the gene expression patterns of 19,276 H-Inv loci in ten major categories of tissues. This analysis allowed us to clearly distinguish broadly and evenly expressed housekeeping genes from those expressed in a more restricted set of tissues (details will be published elsewhere). The H-Angel database comprises the largest and most comprehensive collection of gene expression patterns currently available. Also provided is a classification of human genes by expression pattern.

The second subdatabase of the H-InvDB is DiseaseInfo Viewer. This is a database of known and orphan genetic diseases. We tried to relate H-Inv loci with disease information in two ways. Firstly, 613 H-Inv loci that correspond with known, characterized disease-related genes were identified by creating links to entries in both LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/) and OMIM (Hamosh et al. 2002). To explore the possibility that cDNAs encoding unknown proteins may be related to "orphan pathologies" (diseases that have been mapped to chromosomal regions, but for which associated genes have not yet been described), we generated a list of H-Inv loci that co-localized with these cytogenetic regions. The nonredundant orphan disease dataset we created consists of 586 diseases identified through OMIM (http://www.ncbi.nlm.nih.gov/Omim/, ver. Jan. 2003), with an additional 108 identified from GenAtlas (http://www.dsi.univ-paris5.fr/genatlas/, ver. Jan. 2003). Using the OMIM and GenAtlas databases in conjunction with the annotation results from the H-InvDB may accelerate the process of identifying candidate genes for human genetic diseases.

Concluding Remarks

There are a number of established collections of nonhuman cDNAs, such as those of *Drosophila melanogaster* (Stapleton et al. 2002), *Danio rerio* (Clark et al. 2001), *Arabidopsis thaliana* (Seki et al. 2002), *Plasmodium falciparum* (Watanabe et al. 2002), and *Trypansoma cruzi* (Urmenyi et al. 1999). The most extensive collection of mammalian cDNAs so far has been that of the RIKEN/FANTOM mouse cDNA project (Kawai et al. 2001; Okazaki et al. 2002). This wealth of information has spurred a wide variety of research in the areas of both gene expression profiling (Miki et al. 2001) and protein–protein interactions (Suzuki et al. 2001). The H-InvDB provides an integrative means of performing many more such analyses based on human cDNAs.

The most important findings that have resulted from the cDNA annotation are summarized here.

(1) The 41,118 H-Inv cDNAs were found to cluster into 21,037 human gene candidates. Comparison with known and previously predicted human gene sets revealed that these 21,037 hypothesized gene clusters contain 5,155 new gene candidates.

(2) The primary structure of 21,037 human gene candidates was precisely described. For the majority of them we observed that both first introns and last exons tended to be longer than the other introns and exons, respectively, implying the possible existence of intriguing mechanisms of transcriptional control in first introns.

(3) We discovered the existence of 847 human gene candidates that could not be convincingly mapped to the human genome. This result suggested that up to 3.7%–4.0% of the human genome sequences (NCBI build 34 assembly) may be incomplete, containing either unsequenced regions or regions where sequence assembly has been performed in error.

(4) Based on H-Inv cDNAs, we were able to define an experimentally validated AS dataset. The dataset was composed of 3,181 loci that encoded a total of 8,553 AS isoforms. In the 55% of ORFs containing AS isoforms, the pattern of alternative exon usage was found to encode different functional domains at the same loci.

(5) A standardized method of human curation for the H-Inv cDNAs was created under the tacit consensus of international collaborations. Using this method, we classified 19,574 H-Inv proteins into five categories based on sequence similarity and structural information. We were able to assign functional definitions to 9,139 proteins, to locate function- or family-defining InterPro domains in 2,503 further proteins, and to identify 7,800 transcripts as good candidates for hypothetical proteins.

(6) A total of 1,892 H-Inv proteins were assigned identities as one of 656 different EC-numbered enzymes. This enzyme library includes 32 newly identified human enzymes on known metabolic pathway maps and comprises the largest collection of computationally validated human enzymes.

(7) Based on a variety of supporting evidence, 6.5% of H-

Inv loci (1,377 loci) do not have a good protein-coding ORF, of which 296 loci are strong candidates for ncRNA genes.

(8) We identified and mapped 72,027 SNPs and indels to unique positions on 16,861 loci. Of these, 13,215 non-synonymous SNPs, 358 nonsense SNPs, and 452 indels were found in coding regions and may alter protein sequences, cause phenotypic effects, or be associated with disease. In addition, we identified 216 polymorphic microsatellite repeats on 213 loci, 25 of which were located in coding regions.

(9) During human proteome analysis, it was suggested that the basic gene set of humans might have been established in the early stage of animal evolution. Our analysis of UTRs revealed that insertions or deletions near coding regions were rare when compared with substitutions, though in some cases drastic changes such as UTR replacements occurred.

(10) A consequence of the annotation process and our related research was the development of the H-InvDB to contain our annotation work. H-InvDB is a comprehensive database of human FLcDNA annotations that stores all information produced in this project. As a subdivision of H-InvDB, we developed two other specialized subdatabases: H-Angel and DiseaseInfo Viewer. H-Angel is a database of gene expression patterns for 19,276 loci. DiseaseInfo Viewer is a database of known disease-related genes and loci co-localized with 694 orphan pathologies. These pathologies were mapped onto the genome but were not identified experimentally.

In the H-Inv project, we collected as many FLcDNAs as possible and conducted extensive analyses concerning the quality of cDNAs, such as detection of frameshift errors, retained introns, and internal poly-A priming, under a unified criterion. Although these analyses are still in an elementary state, we store these results in H-InvDB to share this information with the biological community. We believe that this is an important contribution of our project, because it will provide a reliable way to control the quality of the cDNA clones. In the future, this information will be useful for improving the methods of clone library construction.

It has been suggested that the human genome encodes 30,000 to 40,000 genes. In this study we comprehensively evaluated more than 21,000 human gene candidates (up to 70% of the total). Thus, efforts should be continued by the H-Inv consortium and others to "fully" characterize the human transcriptome. For this purpose new technologies should be implemented that are more sensitive in detecting rarely expressed genes and AS transcripts. Nevertheless, there are unavoidable limitations for human cDNA collections, such the identification of embryo-specific genes, for which other approaches should be employed. One alternative is the use of ab initio predictions from genomic sequences, in conjunction with expression profiling studies, to identify rarely expressed genes that share structural similarity to known genes. Additionally, a better characterization of *cis*-regulatory element units may help to define the boundary of other genes that are undetected by current gene prediction programs. Another area that remains to be explored is the identification of potential hidden RNA gene families that may play vital roles, such as the recently uncovered family of microRNA genes, which is involved in the regulation of expression of other genes (for review see Ambros 2001; Moss 2002).

The proteome determination aspects of this project, including the identification of new enzymes and hypothetical proteins, should stimulate more focused biochemical studies. The functional classifications may allow definition of sub-proteomes that are related to different physiological processes. The H-Inv transcriptome based on the definition of a consensus proteome (the H-Inv proteins) links both the analysis of genomic DNA and direct proteome analysis with the study of expressed mRNA analysis from different tissues, cells, and disease states. It creates a standard for the comparison of disease-related alterations of the human proteome. Moreover, comparison with pathogen proteomes may yield many possible drug target proteins. Also, the annotation of ncRNAs raises the possibility of novel "smart" therapeutics that could either inhibit or mimic the mechanisms of these RNAs.

The H-Inv project is the first ever comprehensive compilation of curated and annotated human FLcDNAs. The project may lead to a more complete understanding of the human transcriptome and, as a result, of the human proteome. The preceding examples of the importance of the H-Inv data in understanding human physiology and evolution represent just a small fraction of the research potential of the H-InvDB.

In conclusion, the H-InvDB platform constructed to hold the results of the comprehensive annotations performed by our international team of collaborators represents a substantial contribution to resources that are needed for further exploration of both human biology and pathology.

Materials and Methods

cDNA resources. 41,118 H-Inv cDNAs were sequenced by the Human Full-Length cDNA Sequencing Project (Ota et al. 1997; Yudate et al. 2001; Ota et al. 2004) at the Helix Research Institute, the Institute of Medical Science at the University of Tokyo, and the Kazusa DNA Research Institute (20,999 sequences in total); the Kazusa cDNA Sequencing Project (Kikuno et al. 2002) at the Kazusa DNA Research Institute (2,000 sequences); the Mammalian Gene Collection (Strausberg et al. 1999) at the National Institutes of Health in the United States (11,806 sequences); the German Human cDNA Project (Wiemann et al. 2001) coordinated by the Deutsches Krebsforschungszentrum in Heidelberg (5,555 sequences); and the Chinese National Human Genome Center at Shanghai (Hu et al. 2000) (758 sequences).

Mapping human cDNAs to the human genome and the comparison of the mapped H-Inv cDNAs with other annotated datasets. We have mapped human cDNA sequences to the human genome sequence corresponding to the NCBI build 34 assembly. The datasets we used were a set of 41,118 H-Inv cDNAs and a set of 37,488 human RefSeq sequences available on 15 July 2002 and on the 1 September 2003, respectively. All the revisions for H-Inv cDNA sequences until August 2003 were applied in the datasets. Before performing the mapping procedure, all the repetitive and low-complexity sequences in all the cDNA sequences were masked using RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and Repbase 7.5. Then we used the cross_match program to mask the remaining vector sequences in each cDNA sequence. Any poly-A tails were also masked by using a custom-made Perl script. In the first step of the mapping procedure, we conducted BLASTN (ver.2.2.6) searches of all the sequences against the human genome sequence and extracted the corresponding genomic regions for each query sequence. Then we used est2genome (EMBOSS package ver.2.7.1) to align each sequence to the genomic region with a threshold of 95% identity and 90% coverage. Coverage of each cDNA sequence was calculated excluding those from the vector and poly-A tails that were masked in the previous step. If the sequences were mapped to multiple positions on the human genome, then we selected their best locus based on the identity, length coverage, and number of exons of those sequences. As a result, 77,315 sequences (including 40,140 cDNAs from the H-Inv project) were successfully mapped onto the human genome and were clustered into 38,587 clusters based on sharing at least 1 bp of an

exon on the same chromosome strand. We used all the mapped sequences, including human RefSeq sequences, to compare the clusters that included H-Inv cDNAs with those that consisted of only human RefSeq sequences. 20,190 clusters out of 38,587 consisted of only H-Inv cDNAs or both H-Inv cDNAs and human RefSeq sequences. The rest of the clusters consisted of RefSeq sequences only. All of the mapped cDNAs and the overlap with the RefSeq sequences can be viewed using G-integra in the H-InvDB (http://www.jbirc.aist.go.jp/hinv/g-integra/html/). The mapping procedure for all the unmapped cDNAs against the mouse genome was also performed, using a threshold of 60% identity and 90% coverage.

Clustering of unmapped sequences. The sequences that were not mapped onto the human genome were clustered by a single linkage clustering method. The similarity search was performed among all the unmapped sequences. The program used was MegaBLAST version 2.2.6 (Zhang et al. 2000). As with to the mapping strategy, some distinctive sequences (repetitive regions, contaminations from cloning vectors and poly-A tails) were excluded from the queries of the similarity search. The similarity was evaluated using the expected value (*E*-value) between two sequences. Only when the *E*-value of the two sequences was calculated to be 0, did we assume that a significant level of similarity was detected between the two sequences.

Identification of gene structure. In order to identify gene structure, we used only the representative H-Inv cDNAs. When detecting repetitive elements in cDNAs, RepeatMasker was conducted in a similar manner to the previous phase. We used curated cDNAs in which frameshift errors and remaining introns were removed.

Prediction of ORFs. We predicted ORFs in all 41,118 H-Inv cDNAs, as illustrated in Figure S1, based on the alignment of similarity searches by FASTY (Pearson 2000; Mackey et al. 2002) (ver. 3.4t11) and BLASTX (Altschul et al. 1990) (ver. 2.0.11), and gene prediction by GeneMark (McIninch et al. 1996) (http://opal.biology.gatech.edu/GeneMark/) (Table S10). Prior to the prediction of ORFs, we judged if the sequence had any frameshift errors or remaining introns (see Figure S1). During ORF prediction, we corrected the aforementioned sequence irregularities computationally.

Procedure of computational and human annotation. Prior to the human curation, we performed two computational automated annotation processes to select the representative clone for each locus and to predict function of H-Inv proteins (see Figure S2). We then assigned the most suitable data source ID to each H-Inv protein following a scheme illustrated in Figure S2 and referring to the information using newly developed annotation viewers, named SOUP location viewer, SOUP annotation viewer, and Similarity Motif ORF (SMO) Viewer (Figure S9). Questionable transcripts were determined by human curation based upon evidence such as the following: sequences with no similarity to a known protein or domain, sequences with a very short ORF, cDNAs with only a single exon, and sequences with no EST support. Only 959 (4.9%) of the computationally selected 19,574 representative H-Inv proteins had to be manually corrected. Another 3,142 (16.1%) of the H-Inv proteins had their functional assignment altered by manual curation.

Assignment of functional motifs. Nonredundant proteome datasets were obtained for fly (http://flybase.bio.indiana.edu/), worm (http://www.wormbase.org/), budding yeast (http://www.pasteur.fr/externe), fission yeast (http://www.sanger.ac.uk/), plant (http://mips.gsf.de/proj/thal/index.html), and a bacteria (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/). The H-Inv proteins and other nonredundant proteome datasets were assigned InterPro codes by InterProScan ver. 3.1 (Mulder et al. 2003). The codes corresponded to families, domains, and repeats. GO terms were also assigned (see Table S5).

Evolutionary relationship of proteomes. The top 40 InterPro entries for the human proteome were compared with their equivalents from the fly, worm, yeasts, plant, and bacteria proteomes (see Table S4).

Protein domains and low-complexity inserted sequences. Folds were assigned by reverse PSI-BLAST (Altschul et al. 1997) searches of the amino acid sequences derived from the H-Inv cDNA against the SCOP database (Lo Conte et al. 2000). Information on protein and gene structures, with the exception of mouse and puffer fish, was obtained from the individual genome projects (Blattner et al. 1997; Kunst et al. 1997; CESC 1998; Adams et al. 2000; AGI 2000; Wood et al. 2002). The data for mouse and puffer fish were obtained from the Ensembl database (Hubbard et al. 2002).

Subcellular localization. Subcellular localization targeting signals and transmembrane helices of 40,352 H-Inv proteins were predicted using the PSORT II (Nakai and Horton 1999), TargetP (Emanuelsson et al. 2000), TMHMM, and SOSUI (Hirokawa et al. 1998) computer programs.

UTR sequences. We obtained the UTR sequences from three primates (*Pan troglodytes*, chimpanzee; *Macaca fascicularis*, crab-eating macaque; and *Macaca mulatta*, rhesus monkey) and two rodents (*Mus musculus*, house mouse; and *Rattus norvegicus*, Norwegian rat) that corresponded to UTRs from *Homo sapiens*. In order to do this, we mapped the cDNAs to the human or mouse genome. The corresponding rodent cDNAs were determined by using a human-mouse genome alignment provided by Ensembl. cDNAs of the primates and rodents were retrieved from the DDBJ/EMBL/GenBank databases using the cut off date of 15 July 2002. Additionally, we used the FANTOM2 mouse sequences released on 5 December 2002, and 4,063 5′ ESTs of chimpanzees (Sakate et al. 2003). Corresponding UTRs between human and other species were identified by aligning 5′ and 3′ ends of the human ORFs. To compare evolutionary distances, we analyzed 3,061 and 5,277 orthologous groups that consisted of at least three species' information for the 5′ and 3′ UTR sequences, respectively.

Supporting Information

Dataset S1. List of Library Origins of H-Inv cDNAs (182 Libraries)

The dataset consists of 41,118 H-Inv cDNAs that were cloned from cDNA libraries derived from 182 varieties of cell and tissue.

Found at DOI: 10.1371/journal.pbio.0020162.sd001 (33 KB XLS).

Dataset S2. List of H-Inv Proteins with Potential EC Numbers (1,892 H-Inv Proteins)

The allotted EC numbers are based on the corresponding DNA databank records, UniProt/Swiss-Prot and TrEMBL records that show sequence similarity to the proteins, and InterPro records that the proteins hit.

Found at DOI: 10.1371/journal.pbio.0020162.sd002 (247 KB XLS).

Dataset S3. List of Polymorphic Microsatellites Inferred by Comparisons between the H-Inv cDNAs and Genomic Sequences

Found at DOI: 10.1371/journal.pbio.0020162.sd003 (56 KB XLS).

Figure S1. Prediction of ORFs

(A) Schematic diagram for the prediction of ORFs. This diagram illustrates the ORF prediction method used on all H-Inv cDNAs. The method was based upon the alignment of similarity searches using FASTY and BLASTX. Gene prediction was carried out using GeneMark. Prior to the prediction of ORFs, we judged if a sequence had any frameshift errors or remaining introns. During ORF prediction, we corrected those sequence irregularities computationally. Details of how sequence irregularities were predicted are described in (B) and (C).
(B) Schematic diagram for prediction of unspliced introns. This schematic diagram illustrates the prediction method used for unspliced introns.
(C) Schematic diagram for prediction of frameshift errors. Frameshift errors were inferred from cDNA–genome pairwise alignment gaps due to insertion or deletion, exception of multiple of 3 bp, or over 10 bp in either the query cDNA or genome.
(D) The statistics for the predicted frameshifts and unspliced introns.

Found at DOI: 10.1371/journal.pbio.0020162.sg001 (49 KB PDF).

Figure S2. Scheme of Prediction for Functional Annotation

(A) Schematic diagram for determining a representative transcript for each locus. The procedure of computational autoannotation is illustrated. Prior to the human curation of the representative transcript of each H-Inv cluster, we performed computational autoannotation.
(B) Schematic diagram for functional prediction of H-Inv proteins. This schematic diagram illustrates the H-Inv autofunctional annotation pipeline that can determine the most appropriate data source ID, avoiding the following keywords that suggest proteins without experimental verification in the description; (1) hypothetical, (2) similar to, (3) names of cDNA clones (Rik, KIAA, FLJ, DKFZ, HSPC, MGC, CHGC, and IMAGE) and (4) names of InterPro domain frequent hitters.

Found at DOI: 10.1371/journal.pbio.0020162.sg002 (34 KB PDF).

Figure S3. Size Distribution of Predicted ORFs

The size distribution of all H-Inv proteins among the five similarity categories.

Found at DOI: 10.1371/journal.pbio.0020162.sg003 (24 KB PDF).

Figure S4. Features of Category II Proteins

A total of 4,104 H-Inv proteins were classified as Category II based on sequence similarity to functionally validated proteins. The table and figure show source species of proteins in public databases to which the Category II proteins were similar.

Found at DOI: 10.1371/journal.pbio.0020162.sg004 (9 KB PDF).

Figure S5. H-Inv KEGG Analysis Results (Images of KEGG Pathways)

The images illustrate the metabolic pathways of KEGG database based on the EC number assignments to H-Inv proteins.

Found at DOI: 10.1371/journal.pbio.0020162.sg005 (47 KB PDF).

Figure S6. Numbers of Representative H-Inv cDNAs That Are Homologous to Proteins in Each Taxonomic Group

Two thresholds ($E < 10^{-5}$, white bars, and $E < 10^{-10}$, black bars) were employed. The "animal" group does not include mammalian species. The "eukaryote" group represents eukaryotic species other than animals, fungi, and plants.

Found at DOI: 10.1371/journal.pbio.0020162.sg006 (9 KB PDF).

Figure S7. A Functional Classification of H-Inv Protein Families That Have Homologs in Each Taxonomic Group

H-Inv protein families were identified by clustering H-Inv proteins using the single-linkage clustering method. Then, the number of homologs for each H-Inv protein family was calculated. Mammalian species are excluded from the "animal" group. "eukaryote" represents eukaryotic species other than animals, fungi, and plants.
Single-linkage clustering. All of the H-Inv proteins were compared with themselves by BLASTP and clustered with the thresholds of E-values of 10^{-30} and 10^{-50}. The numbers of singleton families detected were 11,890 and 13,938 at the E-value of 10^{-30} and 10^{-50}, respectively.

Found at DOI: 10.1371/journal.pbio.0020162.sg007 (49 KB PDF).

Figure S8. A Sample View of the H-Invitational Database (H-InvDB; http://www.h-invitational.jp/)

A FLcDNA (BC003551) is shown with its detailed annotations, e.g., gene structure, functional annotation, ORF predictions, protein structure prediction by GTOP, etc. The H-InvDB has links to other internal databases (red boxes) such as a genome map viewer (G-integra) and gene expression library (H-Angel). Green boxes show internal viewers for the results of clustering (Clustering Viewer showing results by H-Inv, STACK, TIGR, UniGene, etc.), the prediction of subcellular localization (TOPOViewer showing results of TMHMM, SOSUI, TargetP, and PsortII), and the disease-related information (DiseaseInfo Viewer linking to OMIM and GenAtlas). The H-InvDB also has links to many external public databases (black boxes), including DDBJ/EMBL/GenBank, RefSeq, UniProt/Swiss-Prot and TrEMBL, Genew, InterPro, 3D Keynote, Ensembl, GeneLynx, LocusLink, PubMed, LIFEdb, dbSNP, GO, and GTOP, and to homepages by original data producers of FLcDNA clones and sequences (blue boxes), including the Chinese National Human Genome Center (CHGC), the Deutsches Krebsforschungszentrum (DKFZ/MIPS), Helix Research Institute (HRI), the Institute of Medical Science at the University of Tokyo (IMSUT), the Kazusa DNA Research Institute (KDRI), the Mammalian Gene Collection (MGC/NIH), and the FLJ project.

Found at DOI: 10.1371/journal.pbio.0020162.sg008 (2,650 KB PDF).

Figure S9. H-Inv Annotation Viewers

(A) G-integra: A genome mapping viewer.
(B) SOUP Locus annotation viewer.
(C) SOUP cDNA annotation viewer.
(D) SMO Viewer: The similarity, motif, and ORF information viewer.

Found at DOI: 10.1371/journal.pbio.0020162.sg009 (2,022 KB PDF).

Table S1. Gene Structure

(A) Gene structure of the cDNAs.
(B) The frequencies and varieties of repetitive sequences found in the cDNAs. A list of the 20,899 loci representing cDNAs that RepeatMasker showed contained repetitive elements.
(C) The positions (5′ UTR, ORF, and 3′ UTR) of repetitive sequences in the protein-coding cDNAs. A total of 1,863 cDNAs contained repetitive sequences in their ORF, of which 549 had repetitive sequences within their most probable ORF. Repetitive sequences appeared in 2,240 and 5,401 cDNAs in their 5′ UTRs and 3′ UTRs, respectively.

Found at DOI: 10.1371/journal.pbio.0020162.st001 (20 KB PDF).

Table S2. CAI and Codon Usage

(A) CAI was measured for all H-Inv proteins. CAI is a measure of biased patterns for synonymous codon usage (http://biobase.dk/embossdocs/cai.html).
(B) Codon usage in predicted ORFs of H-Inv proteins. Total trinucleotide frequencies (forward strand) for the sequences of each species are shown. Nonredundant proteome datasets for nonhuman species were obtained from the following sites: fly (*Drosophila melanogaster*; http://flybase.bio.indiana.edu/), worm (*Caenorhabditis elegans*; http://www.wormbase.org/), budding yeast (*Saccharomyces cerevisiae*; http://www.pasteur.fr/externe), fission yeast (*Schizosaccharomyces pombe*; http://www.sanger.ac.uk/), plant (*Arabidopsis thaliana*; http://mips.gsf.de/proj/thal/index.html), and bacteria (*Escherichia coli* K12; ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/).

Found at DOI: 10.1371/journal.pbio.0020162.st002 (20 KB PDF).

Table S3. Tissue Library Origins of H-Inv Proteins

The results of classification into five similarity categories for each of ten tissue classes.
(A) Numbers of H-Inv proteins.
(B) Histogram.

Found at DOI: 10.1371/journal.pbio.0020162.st003 (10 KB PDF).

Table S4. The InterPro IDs Identified in H-Inv Proteins

The top 40 InterPro IDs identified in H-Inv proteins and proteins from other species are listed for all types (A) and for each type of family, domain, and repeat (B–D). Analyses were conducted by InterPro ver. 3.1. Nonredundant proteome datasets of other species were obtained from the following sites: fly (*Drosophila melanogaster*; http://flybase.bio.indiana.edu/), worm (*Caenorhabditis elegans*; http://www.wormbase.org/), budding yeast (*Saccharomyces cerevisiae*; http://www.pasteur.fr/externe), fission yeast (*Schizosaccharomyces pombe*; http://www.sanger.ac.uk/), plant (*Arabidopsis thaliana*; http://mips.gsf.de/proj/thal/index.html), and bacteria (*Escherichia coli* K12; ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/).

Found at DOI: 10.1371/journal.pbio.0020162.st004 (36 KB PDF).

Table S5. GO Term Assignment to H-Inv Proteins

(A) Molecular function.
(B) Cellular component.
(C) Biological process.

Found at DOI: 10.1371/journal.pbio.0020162.st005 (74 KB PDF).

Table S6. List of Newly Assigned Human Enzymes (32 H-Inv Proteins)

All these 32 H-Inv proteins were newly assigned enzyme numbers with the support of the KEGG pathway. These enzyme assignments were previously unrepresented in *Homo sapiens*.

Found at DOI: 10.1371/journal.pbio.0020162.st006 (33 KB PDF).

Table S7. A Functional Classification of Representative H-Inv cDNAs That Have Homologs in Other Species

(See also Figure 6.)

Found at DOI: 10.1371/journal.pbio.0020162.st007 (9 KB PDF).

Table S8. Basic Statistics for UTR Sequences Analyzed

Found at DOI: 10.1371/journal.pbio.0020162.st008 (8 KB PDF).

Table S9. UTR Replacements in Primates and Rodents

One hundred and forty-seven UTR replacements distributed among different species were detected.

Found at DOI: 10.1371/journal.pbio.0020162.st009 (9 KB PDF).

Table S10. List of the Databases and Software Used in the H-Inv Project

Found at DOI: 10.1371/journal.pbio.0020162.st010 (31 KB PDF).

Protocol S1. A Detailed Functional Annotation Based on Protein Modules

Found at DOI: 10.1371/journal.pbio.0020162.sd004 (25 KB PDF).

Acknowledgments

This paper is dedicated to the late Dr. Yoshimasa Kyogoku, the Director of the Biological Information Research Center, National

Institute of Advanced Industrial Science and Technology, who passed away on February 27, 2003.

The authors express their most sincere gratitude to Drs. David Lipman, Graham Cameron, Joakim Lundeberg, and Francis Collins for their support, the Research Association for Biotechnology of Japan, the International Human Genome Sequencing Consortium, and the Chromosome 22 Group at the Sanger Institute for providing sequence and annotation data. We are grateful to T. Hasui, T. Habara, K. Yamaguchi, H. Kawashima, F. Todokoro, N. Yamamoto, Y. Makita, R. Aono, Y. Tanada, H. Kubooka, H. Maekawa, Y. Sasayama, T. Yamamoto, S. Okiyama, K. Nakamura, A. Matsuya, Y. Mimiura, R. Matsumoto, K. Takabayashi, Y. Hayakawa, H. Zhang, S. Nurimoto, T. Sugisaki, T. Kawamura, O. Nakano, S. Hosoda, N. Yoshimura, and T. Endo for their technical support. This research is financially supported by the Ministry of Economy, Trade, and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science, and Technology of Japan (MEXT), the Japan Biological Informatics Consortium (JBIC), the New Energy and Industrial Technology Development Organization (NEDO), the United States Department of Energy, the National Institutes of Health of the United States, the Bundesministerium für Bildung und Forschung (BMBF) of Germany, the European Union through the EURO-IMAGE Consortium (grant BMH4-CT97-2284 coordinated by Charles Auffray), the 863 and 973 Program of the Ministry of Science and Technology of China, and CNRS of France. The work on Module 3D-keynote is supported by Grants-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" to Mitiko Go and Kei Yura, and for Scientific Research (B) to MG, from MEXT. KY is also supported by a Grant-in-Aid for Encouragement of Young Scientists from MEXT. The work on subcellular localization is supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from MEXT and the National Project on Protein Structural and Functional Analyses from the same Ministry.

Conflicts of interest. The authors have declared that no conflicts of interest exist.

Author contributions. The project was conceived and designed by T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, Y. Yamaguchi-Kabata, S. Miyazaki, K. Ikeo, A. Kasprzyk, T. Nishikawa, M. Stodolsky, W. Makalowski, M. Go, K. Nakai, T. Takagi, M. Kanehisa, Y. Sakaki, J. Quackenbush, Y. Okazaki, Y. Hayashizaki, W. Hide, R. Chakraborty, K. Nishikawa, H. Sugawara, Y. Tateno, Z. Chen, M. Oishi, P. Tonellato, R. Apweiler, K. Okubo, L. Wagner, S. Wiemann, R. L. Strausberg, T. Isogai, C. Auffray, N. Nomura, T. Gojobori, and S. Sugano.

The data were analyzed by T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, K. O. Koyanagi, R. A. Barrero, T. Tamura, Y. Yamaguchi-Kabata, M. Tanino, K. Yura, S. Miyazaki, K. Ikeo, K. Homma, A. Kasprzyk, T. Nishikawa, M. Hirakawa, J. Thierry-Mieg, D. Thierry-Mieg, J. Ashurst, L. Jia, M. Nakao, M. A. Thomas, N. Mulder, Y. Karavidopoulou, L. Jin, S. Kim, T. Yasuda, B. Lenhard, E. Eveno, Y. Suzuki, C. Yamasaki, J.-I. Takeda, C. Gough, P. Hilton, Y. Fujii, H. Sakai, S. Tanaka, C. Amid, M. Bellgard, M. de Fatima Bonaldo, H. Bono, S. K. Bromberg, A. Brookes, E. Bruford, P. Carninci, C. Chelala, C. Couillault, S. J. De Souza, M.-A. Debily, M.-D. Devignes, I. Dubchak, T. Endo, A. Estreicher, E. Eyras, K. Fukami-Kobayashi, G. Gopinathrao, E. Graudens, Y. Hahn, M. Han, Z.-G. Han, K. Hanada, H. Hanaoka, E. Harada, K. Hashimoto, U. Hinz, M. Hirai, T. Hishiki, I. Hopkinson, S. Imbeaud, H. Inoko, A. Kanapin, Y. Kaneko, T. Kasukawa, J. F. Kelso, P. Kersey, R. Kikuno, K. Kimura, B. Korn, V. Kuryshev, I. Makalowska, T. Makino, S. Mano, R. Mariage-Samson, J. Mashima, H. Matsuda, H.-W. Mewes, S. Minoshima, K. Nagai, H. Nagasaki, N. Nagata, R. Nigam, O. Ogasawara, O. Ohara, M. Ohtsubo, N. Okada, T. Okido, S. Oota, M. Ota, T. Ota, T. Otsuki, D. Piatier-Tonneau, A. Poustka, S.-X. Ren, N. Saitou, K. Sakai, S. Sakamoto, R. Sakate, I. Schupp, F. Servant, S. Sherry, R. Shiba, N. Shimizu, M. Shimoyama, A. J. Simpson, B. Soares, C. Steward, M. Suwa, M. Suzuki, A. Takahashi, G. Tamiya, H. Tanaka, T. Taylor, J. D. Terwilliger, P. Unneberg, V. Veeramachanen, S. Watanabe, L. Wilming, N. Yasuda, H.-S. Yoo, W. Makalowski, M. Go, K. Nakai, Y. Okazaki, W. Hide, R. Chakraborty, Z. Chen, P. Tonellato, K. Okubo, L. Wagner, S. Wiemann, T. Isogai, C. Auffray, N. Nomura, T. Gojobori, and S. Sugano.

The paper was written by T. Imanishi, T. Itoh, Y. Suzuki, S. Fukuchi, K. O. Koyanagi, R. A. Barrero, T. Tamura, Y. Yamaguchi-Kabata, M. Tanino, K. Yura, K. Homma, M. Hirakawa, L. Jia, M. Nakao, B. Lenhard, C. Yamasaki, C. Gough, P. Hilton, Y. Fujii, S. Tanaka, C. Chelala, M.-D. Devignes, T. Hishiki, I. Hopkinson, W. Makalowski, K. Nakai, W. Hide, P. Tonellato, C. Auffray, N. Nomura, T. Gojobori, and S. Sugano. ∎

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, et al. (1991) Complementary DNA sequencing: Expressed sequence tags and human genome project. Science 252: 1651–1656.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. (2000) The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

[AGI] Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12: 1805–1814.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 30: 3389–3402.

Ambros V (2001) microRNAs: Tiny regulators with great potential. Cell 107: 823–826.

Andrew S, Goldberg Y, Kremer B, Telenius H, Theilmann J, et al. (1993) The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat Genet 4: 398–403.

Ashburner M (2000) A biologist's view of the *Drosophila* genome annotation. Genome Res 10: 391–393.

Auffray C, Behar G, Bois F, Bouchier C, DaSilva C, et al. (1995) IMAGE: Integrated molecular analysis of the human genome and its expression. C R Acad Sci III, Sci Vie 318: 263–272.

Bahn JH, Kwon OS, Joo HM, Ho Jang S, Park J, et al. (2002) Immunohistochemical studies of brain pyridoxine-5′-phosphate oxidase. Brain Res 925: 159–168.

Bamshad M, Wooding S (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4: 99–111.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235–242.

Blattner FR, Plunkett Gr, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1474.

Boguski MS, Lowe TMJ, Tolstoshev CM (1993) dbEST: Database for expressed sequence tags. Nat Genet 4: 332–333.

Bono H, Kasukawa T, Furuno M, Hayashizaki Y, Okazaki Y (2002) FANTOM DB: Database of functional annotation of RIKEN mouse cDNA clones. Nucleic Acids Res 30: 116–118.

Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, et al. (2002) An anatomy of normal and malignant gene expression. Proc Natl Acad Sci U S A 99: 11287–11292.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd D, et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18: 630–634.

Camargo A, Samaia H, Dias-Neto E, Simao D, Migotto I, et al. (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc Natl Acad Sci U S A 98: 12103–12108.

[CESC] *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: A platform for investigation biology. Science 282: 2012–2018.

Clark M, Hennig S, Herwig R, Clifton S, Marra M, et al. (2001) An oligonucleotide fingerprint normalized and expressed sequence tag characterized zebrafish cDNA library. Genome Res 11: 1594–1602.

Cook RJ (2001) Disruption of histidine catabolism in NEUT2 mice. Arch Biochem Biophys 392: 226–232.

Curtis D, Lehmann R, Zamore PD (1995) Translational regulation in development. Cell 81: 171–178.

Cyranoski D (2002) Geneticists lay foundations for human transcriptome database. Nature 419: 3–4.

Das S, Yu L, Gaitatzes C, Rogers R, Freeman J, et al. (1997) Biology's new Rosetta Stone. Nature 385: 29–30.

Deininger PL, Batzer MA (2002) Mammalian retroelements. Genome Res 12: 1455–1465.

Duyao M, Ambrose C, Myers R, Novelletto A, Persichetti F, et al. (1993) Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat Genet 4: 387–392.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005–1016.

Gaudieri S, Dawkins RL, Habara K, Kulski JK, Gojobori T (2000) SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. Genome Res 10: 1579–1586.

Geballe AP, Morris DR (1994) Initiation codons within 5′-leaders of mRNAs as regulators of translation. Trends Biochem Sci 19: 159–164.

Gerber H, Seipel K, Georgiev O, Hofferer M, Hug M, et al. (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263: 808–811.

Gieser L, Swaroop A (1992) Expressed sequence tags and chromosomal localization of cDNA clones from a subtracted retinal pigment epithelium library. Genomics 13: 873–876.
Go M (1983) Modular structural units, exons, and function in chicken lysozyme. Proc Natl Acad Sci U S A 80: 1964–1968.
[GOC] Gene Ontology Consortium (2001) Creating the gene ontology resource: Design and implementation. Genome Res 11: 1425–1433.
Hamosh A, Scott A, Amberger J, Bocchini C, Valle D, et al. (2002) Online Mendelian inheritance in man (OMIM): A knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30: 52–55.
Haverty PM, Weng Z, Best NL, Auerbach KR, Hsaio LL, et al. (2002) HugeIndex: A database with visualization tools for high-density oligonucleotide array data from normal human tissues. Nucleic Acids Res 30: 214–217.
Hawkins JD (1988) A survey on intron and exon lengths. Nucleic Acids Res 16: 9393–9908.
Hirokawa T, Boon-Chieng S, Mitaku S (1998) SOSUI: Classification and secondary structure prediction system for membrane proteins. Bioinformatics 14: 378–379.
Hirosawa M, Nagase T, Ishikawa K, Kikuno R, Nomura N, et al. (1999) Characterization of cDNA clones selected by the GeneMark analysis from size-fractionated cDNA libraries from human brain. DNA Res 6: 329–336.
Houlgatte R, Mariage-Samson R, Duprat S, Tessier E, Bentolila S, et al. (1995) The genexpress index: A resource for gene discovery and the genic map of the human genome. Genome Res 5: 272–304.
Hu RM, Han ZG, Song HD, Peng YD, Huang QH, et al. (2000) Gene expression profiling in the human hypothalamus-pituitary-adrenal axis and full-length cDNA cloning. Proc Natl Acad Sci U S A 97: 9543–9548.
Hubbard T, Barker D, Birney E, Cameron G, Chen Y, et al. (2002) The Ensembl genome database project. Nucleic Acids Res 30: 38–41.
Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30: 42–46.
Kawabata T, Fukuchi S, Homma K, Ota M, Araki J, et al. (2002) GTOP: A database of protein structures predicted from genome sequences. Nucleic Acids Res 30: 294–298.
Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, et al. (2001) Functional annotation of a full-length mouse cDNA collection. Nature 409: 685–690.
Kawamoto S, Ohnishi T, Kita H, Chisaka O, Okubo K (1999) Expression profiling by iAFLP: A PCR-based method for genome-wide gene expression profiling. Genome Res 9: 1305–1312.
Kawamoto S, Yoshii J, Mizuno K, Ito K, Miyamoto Y, et al. (2000) BodyMap: A collection of 3′ ESTs for analysis of human gene expression information. Genome Res 10: 1817–1827.
Kent WJ, Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. Genome Res 11: 1541–1548.
Khan AS, Wilcox AS, Polymeropoulos MH, Hopkins JA, Stevens TJ, et al. (1992) Single pass sequencing and physical and genetic mapping of human brain cDNAs. Nat Genet 2: 180–185.
Kikuno R, Nagase T, Waki M, Ohara O (2002) HUGE: A database for human large proteins identified in the Kazusa cDNA sequencing project. Nucleic Acids Res 30: 166–168.
Kozak M (1989) The scanning model for translation: An update. J Cell Biol 108: 229–241.
Kriventseva EV, Gelfand MS (1999) Statistical analysis of the exon–intron structure of higher and lower eukaryote genes. J Biomol Struct Dyn 17: 281–288.
Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390: 249–256.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
Larizza A, Makalowski W, Pesole G (2002) Structural and evolutionary analysis of eukaryotic mRNA untranslated regions. Comput Chem 26: 479–490.
Lithgow T, Cuezva JM, Silver PA (1997) Highways for protein delivery to the mitochondria. Trends Biochem Sci 22: 110–113.
Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, et al. (2000) SCOP: A structural classification of proteins database. Nucleic Acids Res 28: 257–259.
Lorenc A, Makalowski W (2003) Transposable elements and vertebrate protein diversity. Genetica. In press
Mackey AJ, Haystead TA, Pearson WR (2002) Getting more from less: Algorithms for rapid protein identification with multiple short peptide sequences. Mol Cell Proteomics 1: 139–147.
Makalowski W (2000) Genomic scrap yard: How genomes utilize all that junk. Gene 259: 61–67.
Maroni G (1996) The organization of eukaryotic genes. Evol Biol 29: 1–19.
Marshall E (2001) Rat genome spurs an unusual partnership. Science 291: 1872.
Martin W (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A 99: 12246–12251.
Matsuzaka Y MS, Makino S, Nakajima K, Tomizawa M, Oka A, et al. (2001) New polymorphic microsatellite markers in the human MHC class III region. Tissue Antigens 57: 397–404.
McCarthy JEG, Kollmus H (1995) Cytoplasmic mRNA-protein interaction gene expression. Trends Biochem Sci 20: 191–197.
McIninch JD, Hayes WS, Borodovsky M (1996) Applications of GeneMark in multispecies environments. Proc Int Conf Intell Syst Mol Biol 4: 165–175.
Miki R, Kadota K, Bono H, Mizuno Y, Tomaru Y, et al. (2001) Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. Proc Natl Acad Sci U S A 98: 2199–2204.
Moonen HJ, Briede JJ, van Maanen JM, Kleinjans JC, de Kok TM (2002) Generation of free radicals and induction of DNA adducts by activation of heterocyclic aromatic amines via different metabolic pathways in vitro. Mol Carcinogen 35: 196–203.
Moore PB, Steitz TA (2002) The involvement of RNA in ribosome function. Nature 418: 229–235.
Moss EG (2002) MicroRNAs: Hidden in the genome. Curr Biol 12: R138–R140.
Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31: 315–318.
Nagase T, Ishikawa K, Kikuno R, Hirosawa M, Nomura N, et al. (1999) Prediction of the coding sequences of unidentified human genes. XV. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. DNA Res 6: 337–345.
Nagase T, Kikuno R, Ohara O (2001) Prediction of the coding sequences of unidentified human genes. XXI. The complete sequences of 60 new cDNA clones from brain which code for large proteins. DNA Res 8: 179–187.
Nakai K, Horton P (1999) PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24: 34–35.
Noguti T, Sakakibara H, Go M (1993) Localization of hydrogen bonds within modules in barnase. Proteins 16: 357–363.
Nomura N, Miyajima N, Sazuka T, Tanaka A, Kawarabayasi Y, et al. (1994) 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.
Ohno S (1996) The notion of the Cambrian pananimalia genome. Proc Natl Acad Sci U S A 93: 8475–8478.
Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420: 563–573.
Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, et al. (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet 2: 173–179.
Ota T, Nishikawa T, Suzuki Y, Maruyama K, Sugano S, et al. (1997) Full-length cDNA project toward a high throughput functional analysis. Microb Comp Genomics 2: 204–205.
Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36: 40–45.
Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132: 185–219.
Pesole G, Liuni S, Grillo G, Saccone C (1997) Structural and compositional features of untranslated regions of eukaryotic mRNAs. Gene 205: 95–102.
Pietu G, Mariage-Samon R, Fayein NA, Matingou C, Eveno E, et al. (1999) The Genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. Genome Res 9: 195–209.
Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29: 137–140.
Reese M, Hartzell G, Harris N, Ohler U, Abril J, et al. (2000) Genome annotation assessment in *Drosophila melanogaster*. Genome Res 10: 483–501.
Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2: 8–27.
Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928–933.
Saha S, Spark A, Rago C, Akmaev V, Wang C, et al. (2002) Using the transcriptome to annotate the genome. Nat Biotechnol 20: 508–512.
Sakate R, Osada N, Hida M, Sugano S, Hayasaka I, et al. (2003) Analysis of 5′-end sequences of chimpanzee cDNAs. Genome Res 13: 1022–1026.
Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, et al. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. Science 296: 141–145.
Sese J, Nikaidou H, Kawamoto S, Minesaki Y, Morishita S, et al. (2001) BodyMap incorporated PCR-based expression profiling data and a gene ranking system. Nucleic Acids Res 29: 156–158.
Sherry ST, Ward M, Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9: 677–679.
Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, et al. (2001) Experimental annotation of the human genome using microarray technology. Nature 409: 922–927.
Snell R, MacMillan J, Cheadle J, Fenton I, Lazarou L, et al. (1993) Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. Nat Genet 4: 393–397.
Sonenberg N (1994) mRNA translation: Influence of the 5′ and 3′ untranslated regions. Curr Opin Genet Dev 4: 310–315.
Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12: 1060–1067.

Stapleton M, Liao G, Brokstein P, Hong L, Carninci P, et al. (2002) The *Drosophila* gene collection: Identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. Genome Res 12: 1294–1300.

Storz G (2002) An expanding universe of noncoding RNAs. Science 296: 1260–1263.

Strausberg RL, Feingold EA, Klausner RD, Collins FS (1999) The mammalian gene collection. Science 286: 455–457.

Strausberg RL, Feingold E, Grouse L, Derge J, Klausner R, et al. (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci U S A 99: 16899–16903.

Suyama M, Nagase T, Ohara O (1999) HUGE: A database for human large proteins identified by Kazusa cDNA sequencing project. Nucleic Acids Res 27: 338–339.

Suzuki H, Fukunishi Y, Kagawa I, Saito R, Oda H, et al. (2001) Protein-protein interaction panel using mouse full-length cDNAs. Genome Res 11: 1758–1765.

Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. Gene 200: 149–156.

Urmenyi T, Bonaldo M, Soares M, Rondinelli E (1999) Construction of a normalized cDNA library for the *Trypanosoma cruzi* genome project. J Eukaryot Microbiol 46: 542–544.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484–487.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.

Watanabe J, Sasaki M, Suzuki J, Sugano S (2002) Analysis of transcriptomes of human malaria parasite *Plasmodium falciparum* using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. Gene 291: 105–113.

Waterston R, Lindblad-Toh K, Birney E, Rogers J, Abril J, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.

Weidanz JA, Campbell P, Moore D, DeLucas LJ, Roden L, et al. (1996) N-acetylglucosamine kinase and N-acetylglucosamine 6-phosphate deacetylase in normal human erythrocytes and *Plasmodium falciparum*. Br J Haematol 95: 645–653.

Wiemann S, Weil B, Wellenreuther R, Gassenhuber J, Glassl S, et al. (2001) Toward a catalog of human genes and proteins: Sequencing and analysis of 500 novel complete protein coding human cDNAs. Genome Res 11: 422–435.

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. Nature 415: 871–880.

Yudate HT, Suwa M, Irie R, Matsui H, Nishikawa T, et al. (2001) HUNT: Launch of a full-length cDNA database from the Helix Research Institute. Nucleic Acids Res 29: 185–188.

Yulug IG, Yulug A, Fisher EM (1995) The frequency and position of Alu repeats in cDNAs, as determined by database searching. Genomics 27: 544–548.

Zaidi SHE, Malter JS (1994) Amyloid precursor protein mRNA stability is controlled by a 29-base element in the 3′-untranslated region. J Biol Chem 269: 24007–24013.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7: 203–214.

D78–D81 Nucleic Acids Research, 2004, Vol. 32, Database issue
DOI: 10.1093/nar/gkh076

DBTSS, DataBase of Transcriptional Start Sites: progress report 2004

Yutaka Suzuki[1,*], Riu Yamashita[1,2], Sumio Sugano[1] and Kenta Nakai[1,2]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan, and [2]Undergraduate Program for Bioinformatics and Systems Biology, Faculty of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

Received September 15, 2003; Revised and Accepted October 1, 2003

DDBJ/EMBL/GenBank accession nos*

ABSTRACT

DBTSS (http://dbtss.hgc.jp) was originally constructed based on a collection of experimentally determined TSSs of human genes. Since its first release in 2002, it has been updated several times. First, the amount of stored data has increased significantly: e.g. the number of clones that match both the RefSeq mRNA set and the genome sequence has increased from 111 382 to 190 964, now covering 11 234 genes. Second, the positions of SNPs in dbSNP were displayed on the upstream regions of contained human genes. Third, DBTSS now covers other species such as mouse and the human malaria parasite. It will become a central database containing data for many more species with oligo-capping and related methods. Lastly, the database now serves for comparative promoter analyses: in the current version, comparative views of potentially orthologous promoters from human and mouse are presented with an additional function of searching potential transcription-factor binding sites, which are either conserved or diverged between species.

INTRODUCTION

The knowledge of exact transcriptional start sites (TSSs) of genes is valuable in many ways: it makes the prediction of translational start sites more accurate; it can be used for exploring sequence determinants of TSSs; and it makes the analysis of upstream regulatory regions (promoters) more precise. In principle, information of a TSS is obtained by mapping the corresponding transcript onto the genome sequence. Nevertheless, it is widely known that many mRNA sequence data stored in public databases, lack information about their 5′ ends because of the difficulty in obtaining full-length cDNAs. Thus, even after the completion of human genome sequencing, it is not easy to locate TSSs systematically. To overcome this problem, we have developed a method to construct full-length enriched cDNA libraries using a cap selection technique, the oligo-capping method, and have been systematically collecting full-length cDNA data with this method [(1); T.Ota *et al.* submitted]. Initial computational characterization of human TSSs has been carried out (2,3) and a database [DataBase of Transcriptional Start Sites (DBTSS)] containing the TSS information of 7889 human genes has been constructed (4). In this report, we summarize the updates of DBTSS since its first release, including its new departure as a basis of comparative promoter analyses.

NEW FEATURES

Compared with its initial version, the current DBTSS (version 3) has been upgraded in at least five ways. First, the number of processed one-pass human cDNA clones has increased significantly (from 217 402 to 400 225). Since one of the important findings from our TSS analysis was that the TSS position of a gene is not always fixed but rather often fluctuates for ~50 bp on average (3), the distribution of TSS positions should become clearer as the number of mapped cDNA clones increases. As always, we constructed a so-called RefFull sequence set (11 234 sequences) by extending the 5′-end sequences of RefSeq mRNA sequences (5), if necessary. On average, 6042 sequences were extended by 71.6 bp. At the genomic level, the average difference between 5′-ends of two data sets becomes 4396 bp because of internal introns. Thus, it is clear that our data make promoter analysis of human genes much easier. For more details of the statistics of the DBTSS, see the Statistics section of the DBTSS web page.

Second, to facilitate promoter analysis of human genes, we mapped the positions of single nucleotide polymorphisms (SNPs) stored in a public database, dbSNP (5), on the −1000:+200 region of each representative TSS for each human gene (a sample output is shown in Fig. 1). These SNPs are candidates of functional regulatory SNPs (rSNPs) that affect the promoter activity. We also plan to add SNP data from other sources. In DBTSS, it is also possible to enlist the name of genes located within a specified distance from each SNP.

The third, and probably the most important, upgrade of DBTSS is that it now supports data from multiple species. To date, we have constructed many full-length cDNA libraries of various species upon requests from many researchers. In addition, large-scale collections of cDNAs determined using a related method by Yoshihide Hayashizaki's group are also publicly available (6,7). In the current version, we added the

*To whom correspondence should be addressed. Tel: +81 3 5449 5343; Fax: +81 3 5449 5416; Email: ysuzuki@ims.u-tokyo.ac.jp

*BP192706–BP383670

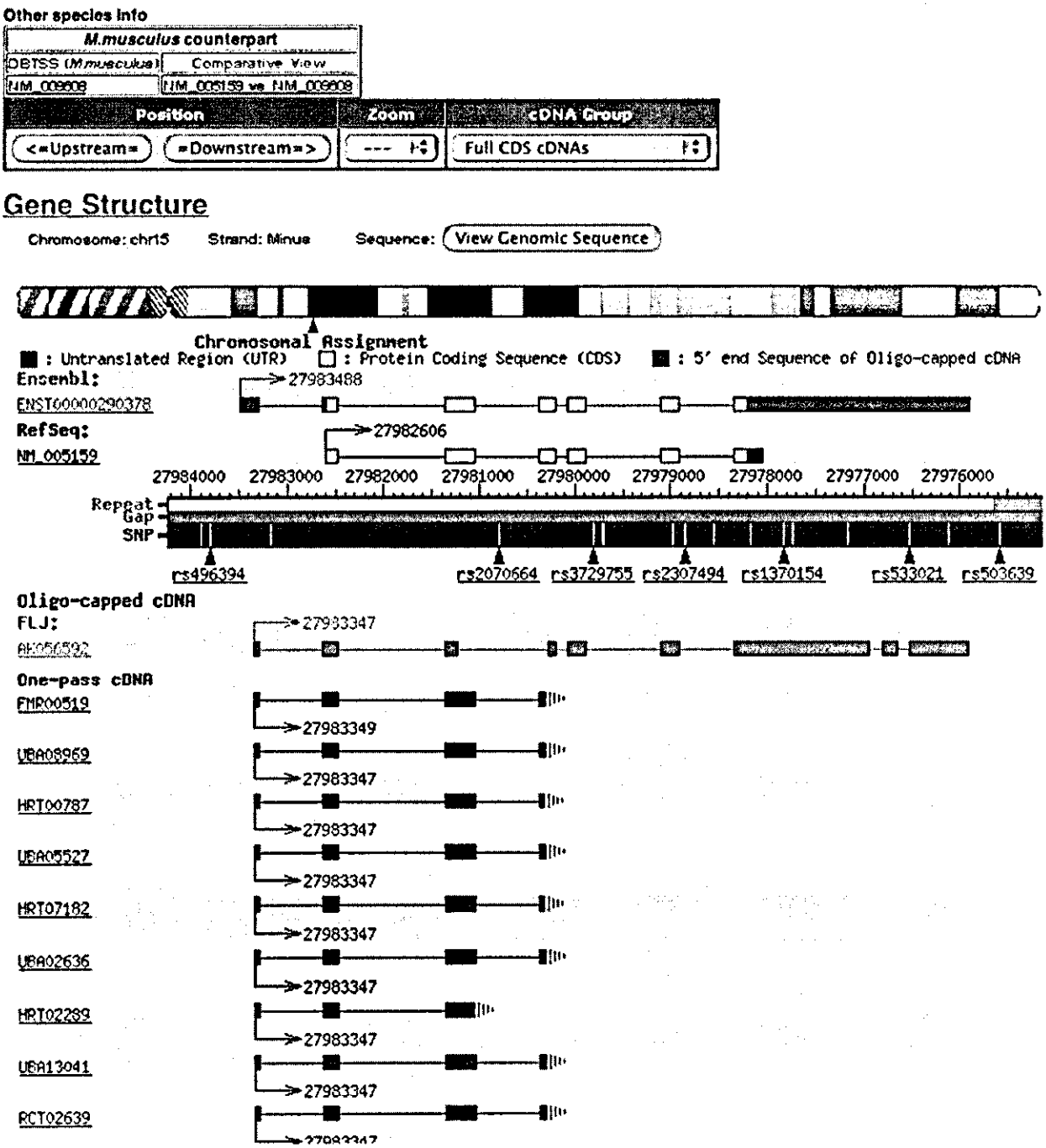

Figure 1. Example of the output of a human gene including the correspondence with a mouse gene, gene position in the chromosome, comparison with Ensembl and RefSeq data, SNP positions and graphical representations of one-path cDNA clones.

data of 2490 clones of *Plasmodium falciparum*, the human malaria parasite (8) and 580 209 full-length cDNA sequences of *Mus musculus* (7). The number of Ref-full members for mouse is 6875 (for more details, see Y.Suzuki *et al.*, submitted). We will add data for other species whenever we get the agreement. They include data for *Caenorhabditis elegans*, chimpanzee, macaque, *Cyanidioschyzon melorae* (unicellular red alga), zebrafish and sorghum.

The remaining two novel features will be explained in the next section.

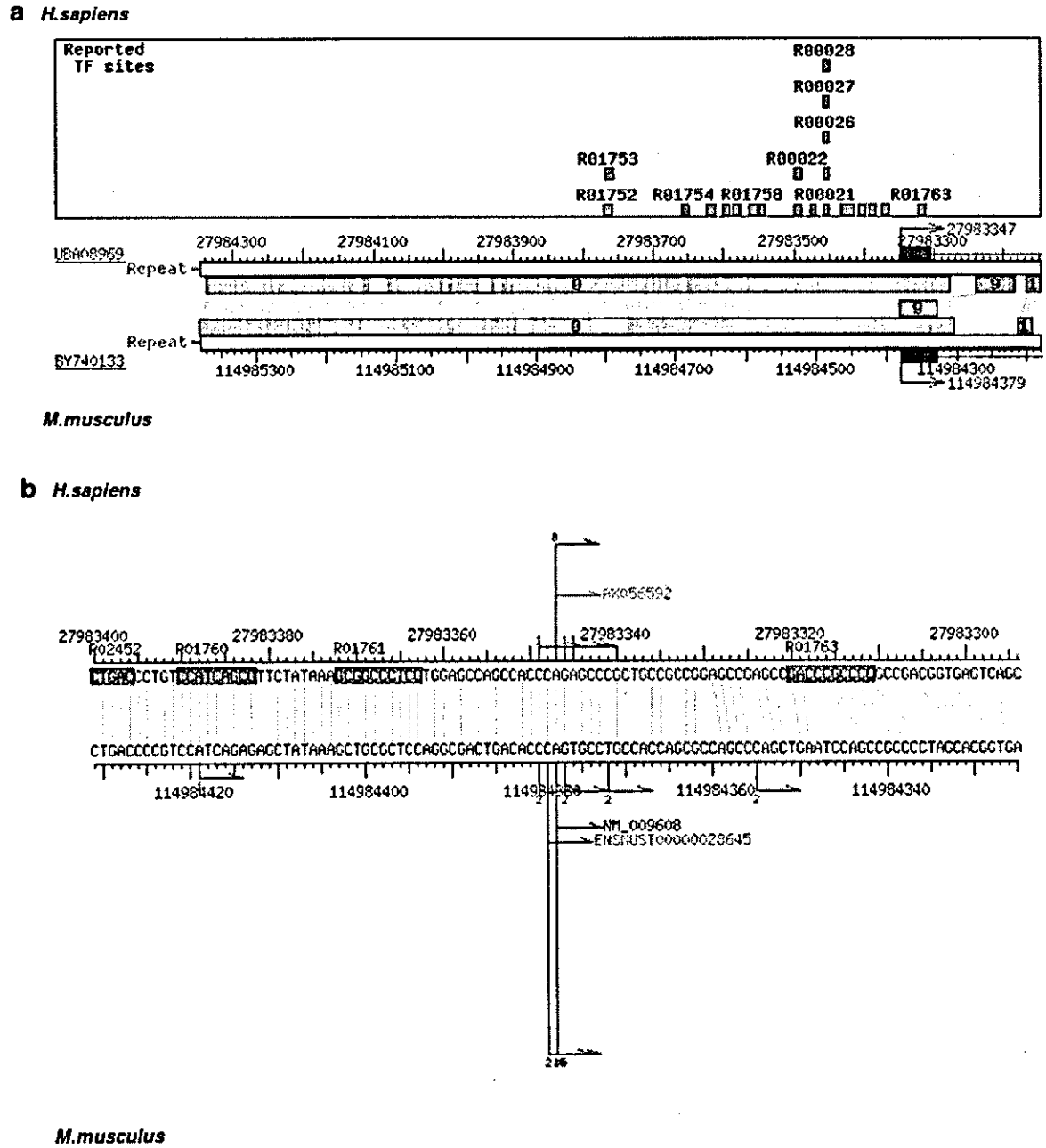

Figure 2. A comparative view of human and mouse promoters. (a) Global view with potential transcription factor binding sites. Locally similar sequence segments are shown in boxes and the corresponding boxes are represented by the same number (e.g. '0'). (b) More detailed view around the corresponding TSSs.

PROMOTER COMPARISON AND SEARCH OF *CIS*-ELEMENTS

The fourth novel feature of the DBTSS (version 3) is that it provides users with comparative views of human and mouse promoters that are probably orthologous. The potentially orthologous gene set was obtained from the LocusLink database (5) and our own sequence comparison. As a result, promoters of 3324 gene pairs can now be displayed. In each pair, locally similar sequence segments were detected by a local alignment program, LALIGN (9) and their correspondences are shown graphically (Fig. 2).

The fifth novel feature is a function for locating positions similar to known transcription-factor binding sites, which are stored in the TRANSFAC database (10). More specifically, we support TRANSFAC Public-based search (for searches using TRANSFAC Professional, which is a commercial version, users should follow its condition of use, which are shown in

our web page). To reduce the number of potentially spurious hits, users can choose various levels of cut-off values and target regions/strands. Moreover, it is also possible to restrict hits within conserved regions between the two species. It is also possible for users to enlist gene names that specify combinations of the above conditions: e.g. genes that harbor both potential binding sites of factors A and B on their upstream regions could be selected with arbitrary cut-off values. With this function, the DBTSS can now be regarded as a platform of systematic promoter analyses.

DBTSS is available at http://dbtss.hgc.jp/ and will continue to expand, incorporating our in-house data and others.

ACKNOWLEDGEMENTS

We thank T. Hasui, K. Abe, M. Morinaga, M. Ishizawa, M. Kawamura, T. Mizuno, A. Kanai and H. Hata for technical support; J. Mizushima-Sugano and E. Nakajima for helpful discussion; Y. Hayashizaki for permission to incorporate their mouse data into DBTSS; and E. Wingender and A. Kel for enabling TRANSFAC-based search. This study was supported by a Grant-in-Aid for Scientific Research on Priority Areas and by special coordination funds for promoting science and technology (SCF), both from the Ministry of Education, Culture, Sports, Science and Technology in Japan.

REFERENCES

1. Suzuki,Y. and Sugano,S. (2003) Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.*, **221**, 73–91.
2. Suzuki,Y., Tsunoda,T., Sese,J., Taira,H., Mizushima-Sugano,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Nakamura,Y. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.*, **11**, 677–684.
3. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Morishita,S. *et al.* (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.*, **2**, 388–393.
4. Suzuki,Y. Yamashita,R., Nakai,K. and Sugano S. (2002) DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
5. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
6. Carninci,P. and Hayashizaki,Y. (1999) High-efficiency full-length cDNA cloning. *Methods Enzymol.*, **303**, 19–44.
7. The FANTOM consortium and the RIKEN Genome Exploration Research Group Phase I & II Team (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
8. Watanabe,J., Sasaki,M., Suzuki,Y. and Sugano,S. (2002) Analysis of transcriptomes of human malaria parasite *Plasmodium falciparum* using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. *Gene*, **291**, 105–113.
9. Huang,X.Q., Hardison,R.C. and Miller,W. (1990) A space-efficient algorithm for local similarities. *Comput. Appl. Biosci.*, **16**, 373–381.
10. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A. E,, Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Sequence Comparison of Human and Mouse Genes Reveals a Homologous Block Structure in the Promoter Regions

Yutaka Suzuki,[1,3] Riu Yamashita,[1] Matsuyuki Shirota,[1] Yuta Sakakibara,[1,2] Joe Chiba,[2] Junko Mizushima-Sugano,[1] Kenta Nakai,[1] and Sumio Sugano[1]

[1]*Human Genome Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, 108-8639, Japan;*
[2]*Department of Biological Science and Technology, Science University of Tokyo, Noda-shi, Chiba, 278-8510, Japan*

Comparative sequence analysis was carried out for the regions adjacent to experimentally validated transcriptional start sites (TSSs), using 3324 pairs of human and mouse genes. We aligned the upstream putative promoter sequences over the 1-kb proximal regions and found that the sequence conservation could not be further extended at, on average, 510 bp upstream positions of the TSSs. This discontinuous manner of the sequence conservation revealed a "block" structure in about one-third of the putative promoter regions. Consistently, we also observed that G+C content and CpG frequency were significantly different inside and outside the blocks. Within the blocks, the sequence identity was uniformly 65% regardless of their length. About 90% of the previously characterized transcription factor binding sites were located within those blocks. In 46% of the blocks, the 5′ ends were bounded by interspersed repetitive elements, some of which may have nucleated the genomic rearrangements. The length of the blocks was shortest in the promoters of genes encoding transcription factors and of genes whose expression patterns are brain specific, which suggests that the evolutional diversifications in the transcriptional modulations should be the most marked in these populations of genes.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to DDBJ under accession nos. BP192706–BP383670.]

As fellow mammals, humans share many physiological, anatomical, and metabolic parallels with mice (Nadeau and Taylor 1984). However, there are striking differences between the two species as well, that is, alterations in size, shape, and longevity. Above all, humans but not mice have developed highly complex neural systems in the brain. It has long been supposed that the genetic basis for these similarities/differences lies, at least in part, in alterations in the expression of genes rather than changes in the functions of their encoded protein products (King and Wilson 1975; Tautz 2000). Differential regulation of gene expression seems a likely explanation for many differences between humans and mice. Between humans and mice, many of the protein functions themselves have been shown to be comparable (Boguski 2002). To understand the molecular machinery that makes humans distinct from mice, the features in the transcriptional networks that are unique to humans should be identified. On the other hand, if the mechanisms that constitute the basic framework of the genetic network are to be delineated, the investigation should be focused on the features that are shared between humans and mice.

However, only limited knowledge has been accumulated about how and to what extent the transcriptional modulatory mechanisms are conserved or divergent between human and mouse genes. Although there are pioneering studies phylogenetically comparing the genomic sequences involved in transcriptional regulations (for reviews, see Ureta-Vidal et al. 2003; Wray et al. 2003), our understanding of the comprehensive systems of transcriptional regulation is still at a very primitive stage. To address this issue, it is essential to enrich our basic knowledge of the molecular mechanisms underlying the regulation of the transcription of each gene.

One of the most important regulatory steps for transcription is the initiation step. For many genes, it has been shown that the transcription level is regulated by controlling the efficiency of the formation of the RNA polymerase II pre-initiation complex (Mitchell and Tjian 1989; Roeder 1996). The DNA sequence just adjacent to the transcriptional start sites (TSSs) plays an important role in the regulation. This region is called the promoter, and several *cis*-regulatory sequence elements are embedded in it. These *cis*-acting elements are recognized by general transcription factors (GTFs), various kinds of transcription regulatory factors (TFs), or other protein factors. When these proteins are recruited to the promoter, they accelerate/inhibit the formation of the preinitiation complex through direct interaction or by changing the conformation of the docking platform (Novina and Roy 1996). To understand the molecular mechanisms of such transcriptional regulation, it is essential to identify and characterize what kinds of *cis*-elements are embedded within the promoters and what kinds of TFs are recruited onto the promoters (http://www.epd.isb-sib.ch; Eukaryotic Promoter Database; and http://www.gene-regulation.com/; TRANSFAC; Praz et al. 2002; Kel et al. 2003).

With the near completion of the human and mouse genome sequencing projects (http://genome.ucsc.edu/downloads.html; UCSC Genome Browser; Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002), the basic materials to start genome-wide analyses of promoters have become available. Because the promoters are located proximal to or overlapping with the TSSs and because the 5′ ends of full-length cDNA sequences correspond to the TSS, it is possible to retrieve the putative promoter sequences

[3]Corresponding author.
E-MAIL ysuzuki@ims.u-tokyo.ac.jp; FAX +81 3 5449 5416.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2435604.

(called "putative promoter regions" [PPRs] hereafter) from large volumes of genomic sequences by combining the information about genomic DNA and full-length cDNAs.

We previously developed a method to construct full-length cDNA libraries and have been collecting full-length cDNAs (Carninci and Hayashizaki 1999; Suzuki and Sugano 2003). So far, we have accumulated 400,225 human and 580,209 mouse cDNAs (http://fantom.gsc.riken.go.jp/; FANTOM), from a wide variety of tissues and cultured cells (Kawai et al. 2001; Okazaki et al. 2002; Waterston et al. 2002). Based on the data for these full-length cDNAs, in the present study we were able to determine the exact positions of their TSSs on the genomic sequences and retrieve the PPR sequences for 8793 human and 6875 mouse RefSeq genes (http://dbtss.hgc.jp/; DBTSS; and http://www.ncbi.nlm.nih.gov/RefSeq/; RefSeq). Of these, 3324 promoters could be paired with each other between mutually 1:1 homologous genes (Statistics of the data set used in the present study are summarized in Supplemental data Table 1; for further details refer to Suzuki et al. 2004). This collection of PPR sequences enabled us, for the first time, to precisely distinguish which parts of the genomic sequences correspond to the exonic regions, TSSs, and upstream regions. Here we report our first large-scale comparative sequence analyses of PPRs between human and mouse genes.

RESULTS

Sequence Comparison of Promoters Between Human and Mouse Genes

We aligned the PPR sequences of 3324 pairs of human and mouse genes over the regions proximal to the TSSs, from −1 kb to +200 bp (the TSS was designated as 0). The sequence identities calculated for these regions were 46% on average. Consistent with a previous report (Waterston et al. 2002), the average sequence identity was the highest in the −100-bp to +100-bp region, and it decreased as the distance from the TSSs increased (Fig. 1A).

For the alignment, we used the sequence alignment program LALIGN (Huang et al. 1992), because it is a relatively simple local alignment program that is robust against gaps (a typical example of the results is shown in Supplemental data Fig. 1). We also used CLUSTALW (Thompson et al. 1994), which is one of the most popular global alignment programs. However, CLUSTALW was inappropriate for our purpose. When CLUSTALW was used for the alignment, a relatively short gap disturbed the overall alignment in many cases (data not shown).

We further examined the sequence alignments and found that the aligned sequences did not always cover the entire 1-kb upstream region. Very frequently, the sequence alignments disappeared at particular positions within the 1-kb regions, which made the aligned parts look like "blocks" (a typical example is illustrated in Supplemental data Fig. 1). The boundary of the block was defined as the most distal aligned region according to the result of LALIGN. The observed patterns of gradually decreasing average identities mainly accounted for the difference in the frequency of the blocks covering the corresponding regions (Fig. 1B). The average length of the blocks was 510 bp (Fig. 2A). The sequence identity inside the blocks was uniformly around 65% irrespective of the block's length (Fig. 2B,C). The overall sequence similarities of the upstream sequences were mainly dependent on the length of the blocks. We performed similar analyses using different parameters for gap-opening penalties and gap-extension penalties. We observed essentially the similar results unless the effects parameter changes resulted in disruption of the alignments themselves (for further details, see Supplemental data Fig. 2).

We also examined whether this discontinuous manner of the sequence conservation was specific to the PPRs using the sequences of the nongenic regions. Positional information of the putative syntenic regions of the human and mouse genomes were obtained from UCSC Genome Browser and those regions at least 100 kb apart from the so-called Ensembl regions were selected (http://genome.ucsc.edu/goldenPath/14nov2002/database/; Ensembl; and http://www.ensembl.org/; Ensembl). Using the distal sequences (−1 kb to 200 bp) of those putative "homologous" regions, a similar analysis was performed. As shown in Figure 2D, the discontinuity of the sequence conservation was also observed in the nongenic regions throughout the genome (for further details on these homologous regions in the nongenic regions, see Supplemental data Fig. 3).

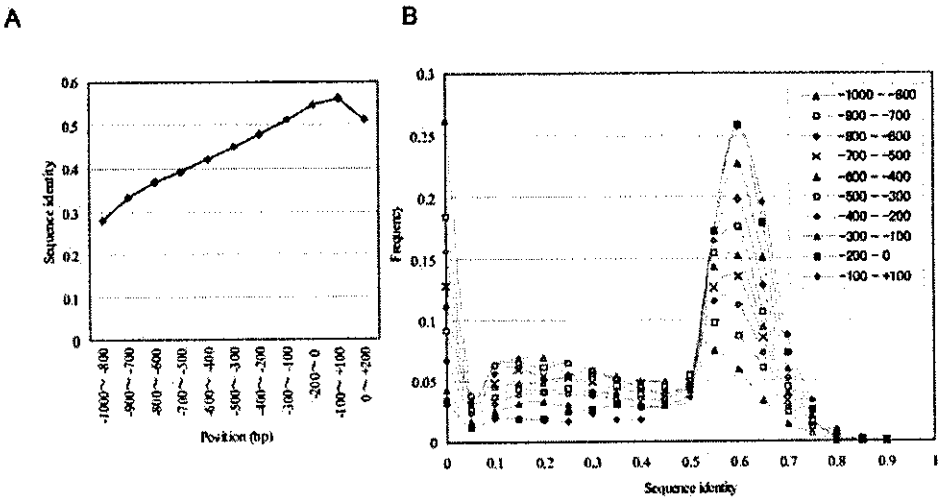

Figure 1 Sequence identity between human and mouse PPRs. Sequence alignments were calculated using LALIGN with the default parameters. The sequence identity was evaluated as the number of aligned nucleotides in the regions of −1000 to +200 (TSS: 0). The average sequence identities were calculated for each region (*A*). (*B*) The PPRs were separated into the 200-bp windows at the positions indicated in the *inset*. Sequence identity was calculated for each of the windows. Frequency as to which of the windows belong to which of the sequence identity groups represented on the horizontal axis is plotted.

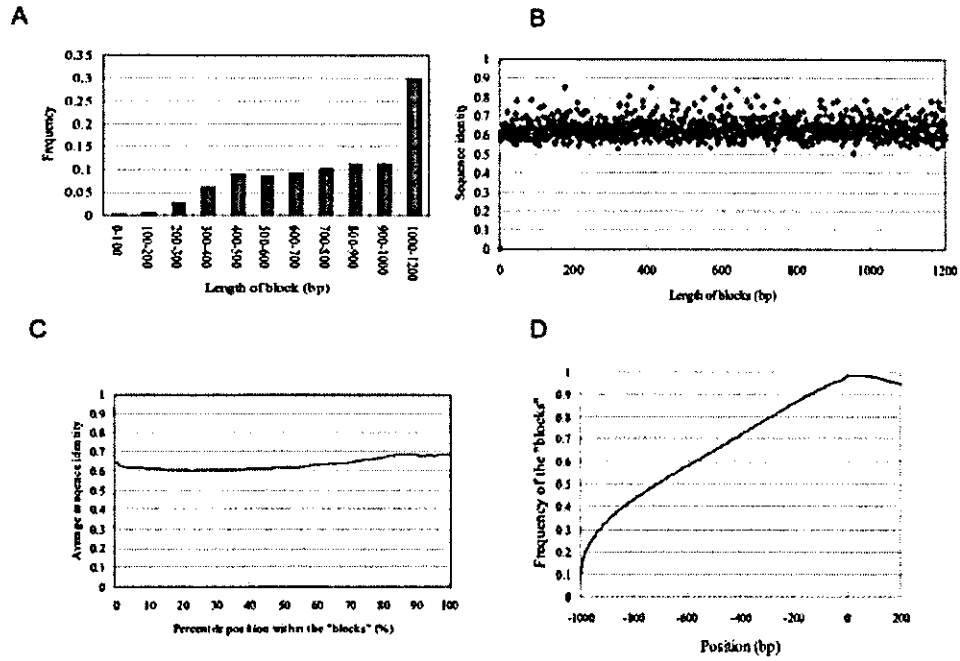

Figure 2 Sequence alignments of the block structure in PPRs and nongenic regions. (*A*) Frequency of the blocks belonging to each population is shown. (*B*) Relation between length of the block and the average sequence identity within it. (*C*) Relation between percentile position within the block and the average sequence identity. (*D*) Alignment of the nongenic sequences using LALIGN. The sequences ranging from −1 kb to + 200 bp of the putative syntenic regions located in nongenic regions as in UCSC genome browser were aligned and the frequencies of the aligned nucleotides were calculated at each of the positions. Vertical line represents the frequency of the nucleotide at the indicated position being located within the block. (Note that the vertical axis in Figure 1 represents the frequency of the sequence "identity").

Sequences Around the Distal Regions of the Block Structure

We examined why the sequence alignments could not be further extended at the edge of the blocks. It was rare that the alignments were terminated at the positions of sequence gaps (incompleteness in the genome sequencing) in either the human or mouse genomes. In humans, 31% of the boundaries were flanked by interspersed repetitive elements (Table 1). Of these, 16% corresponded to *Alu* elements, which are primate-specific repetitive elements (Mitchell and Tjian 1989; Deininger and Batzer 2002). Similarly, in mice, insertion of repetitive elements was observed for 20% of the boundaries, of which 8% were B1 elements, which are *Alu* superfamily elements in rodents. Taken together, in 46% of the blocks, repetitive elements were found at the boundaries in either the human or mouse genome. For this population, it is possible that the sequence alignments were disrupted because the repetitive elements were inserted into otherwise continuous regions. It was possible that LALIGN could not allocate "gaps" to them in the alignments. To address this issue, we excised the repetitive elements and generated the sequence alignments again. Still, we could not identify sequence similarity significantly greater than 30% in essentially any case. This is similar to the results obtained from the analyses of the remaining 54% of the edges of the blocks. In either case, the sequences outside of the blocks seemed completely lost from the corresponding parts of the counter genomes.

Sequences Are Conserved in a Block Manner in the Promoters

To determine whether the observed block structures were derived from algorithmic artifacts of LALIGN, we aligned the PPR sequences using another type of sequence alignment program, SSEARCH (Smith and Waterman 1981; Pearson 1996). This program is based on the simple Smith–Waterman algorithm and gives the most precise alignments, though it is computationally expensive. Using the SSEARCH alignments, we again observed the similar block structures, and the sequence identities sharply dropped just outside the blocks. In these cases, the results were robust against changes of the parameters, as is the case for LALIGN (also see Supplemental data Fig. 4).

When a similar analysis was performed using the sequences around the 5′-end boundaries of the second exons (note that PPRs were defined as the regions upstream of the first exons), the SSEARCH scores dropped sharply at the 5′ ends of the exons (Fig. 3). Thus, the boundaries of the block structures were overlapped with the exon–intron boundaries in these cases. The boundaries between exonic and intronic sequences can be considered as transition points from the regions where most of the sequences play biologically significant roles to the regions where most of the sequences are biologically less relevant. Similarly, it can be suggested that, in the promoters, most of the biologically significant elements should be embedded inside rather than outside the blocks. It was also significant that such discontinuity in the sequence conservation has frequently been observed in the proximal regions of both the boundaries of the blocks in the PPRs and the exon–intron boundaries.

Differences in G+C Content and CpG Frequency Between the Sequences Inside and Outside the Blocks

We compared the G+C contents and the frequencies of the dinucleotides, CpG, between the sequences inside and outside the blocks (Table 2). Promoters are frequently associated with the G+C-rich regions with increased frequency of the CpG (Cross and Bird 1995). For humans, when the sequences of 200 bp