Toward a Catalog of Human Genes and Proteins: Sequencing and Analysis of 500 Novel Complete Protein Coding Human cDNAs

Stefan Wiemann,[1,12] Bernd Weil,[1,2] Ruth Wellenreuther,[1]
Johannes Gassenhuber,[1,2] Sabine Glassl,[3] Wilhelm Ansorge,[3] Michael Böcher,[4]
Helmut Blöcker,[4] Stefan Bauersachs,[5] Helmut Blum,[5] Jürgen Lauber,[6]
Andreas Düsterhöft,[6] Andreas Beyer,[7] Karl Köhrer,[7] Normann Strack,[2]
Hans-Werner Mewes[2], Birgit Ottenwälder,[8] Brigitte Obermaier,[8] Jens Tampe,[9]
Dagmar Heubner,[10] Rolf Wambutt,[10] Bernhard Korn,[1,11] Michaela Klein,[1]
and Annemarie Poustka[1]

[1]Molecular Genome Analysis, German Cancer Research Center, 69120 Heidelberg, Germany; [2]MIPS, GSF, 82152 Martinsried, Germany; [3]Biochemical Instrumentation, European Molecular Biology Laboratory, 69117 Heidelberg, Germany; [4]GBF–Genome Analysis, 38124 Braunschweig, Germany; [5]Genzentrum der LMU München, 81377 München, Germany; [6]QIAGEN GmbH, 40724 Hilden, Germany; [7]Biologisch-Medizinisches Forschungszentrum, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; [8]MediGenomix GmbH, 82152 Martinsried, Germany; [9]Fraunhofer Gesellschaft, 80636 München, Germany; [10]AGOWA GmbH, 12489 Berlin, Germany; [11]Resource Center of the German Genome Project, 69120 Heidelberg, Germany

With the complete human genomic sequence being unraveled, the focus will shift to gene identification and to the functional analysis of gene products. The generation of a set of cDNAs, both sequences and physical clones, which contains the complete and noninterrupted protein coding regions of all human genes will provide the indispensable tools for the systematic and comprehensive analysis of protein function to eventually understand the molecular basis of man. Here we report the sequencing and analysis of 500 novel human cDNAs containing the complete protein coding frame. Assignment to functional categories was possible for 52% (259) of the encoded proteins, the remaining fraction having no similarities with known proteins. By aligning the cDNA sequences with the sequences of the finished chromosomes 21 and 22 we identified a number of genes that either had been completely missed in the analysis of the genomic sequences or had been wrongly predicted. Three of these genes appear to be present in several copies. We conclude that full-length cDNA sequencing continues to be crucial also for the accurate identification of genes. The set of 500 novel cDNAs, and another 1000 full-coding cDNAs of known transcripts we have identified, adds up to cDNA representations covering 2%–5 % of all human genes. We thus substantially contribute to the generation of a gene catalog, consisting of both full-coding cDNA sequences and clones, which should be made freely available and will become an invaluable tool for detailed functional studies.

[The sequence data described in this paper have been submitted to the EMBL database under the accession nos. given in Table 2.]

The recent past has witnessed major advances in the determination of the sequence of the human genome (Dunham et al. 1999; Hattori et al. 2000). Although the whole genomic sequence will be completely unraveled in the near future (Collins et al. 1998), the identification of genes and the deciphering of gene structures will extend for a prolonged time, and cDNA sequences will continue to be invaluable tools for this adventure, especially in view of alternative splicing. The primary focus will shift to the functional analysis of the genes and their protein products to finally understand the molecular basis of human life. Current estimates vary between 29,000 and >70,000 genes to constitute the protein coding repertoire of the human genome (Fields et al. 1994; Ewing and Green 2000; Liang et al. 2000; Roest Crollius et al. 2000). However, thus far only some 11,000 cDNA sequences have been deposited in public databases, which are supposed to contain the complete

[12]**Corresponding author.**
E-MAIL s.wiemann@dkfz.de; FAX 49-6221-4252-4702.

Article published on-line before print: *Genome Res.*, 10.1101/gr.154701. Article and publication are at www.genome.org/cgi/doi/10.1101/gr.154701.

11:422–435 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

protein coding open reading frame (ORF). The majority of the respective cDNA clones are most likely not accessible. The generation of a physical clone set representing all human genes that should be made freely accessible is consequently regarded to have an extremely high impact (Schuler 1997; Pruitt et al. 2000). This would permit the establishment of a catalog of clones to provide the resources needed in the proteomics era where the functions of proteins, their action in pathways, and the possible disease relation are deciphered.

Until recently, the long-cDNA sequencing project carried out at the Kazusa Institute (Nomura et al. 1994; Nagase et al. 2000) Consortium had been the only systematic full-length cDNA sequencing project with a significant output of novel sequence information. The initiation of a new large-scale cDNA sequencing project has been announced lately that is coordinated by the National Institute of Health (Strausberg et al. 1999). We founded a cDNA Consortium in 1997 as part of the German Genome Project and aim at the characterization of the complete sequences of novel human transcripts at the cDNA level.

Here, we report the sequences and analysis of 500 novel human cDNAs that all contain the complete protein coding region. These cDNAs constitute the most valuable essence of 30,000 clones that have been EST sequenced and 3630 fully sequenced cDNAs. Over 1000 cDNAs that cover the complete coding sequence of already known transcripts have been identified in the EST-sequenced clone set. All clones are made available through the Resource Center of the German Genome Project (RZPD).

RESULTS

Libraries and Clones

To identify and sequence novel human cDNAs we have 5′-EST sequenced >30,000 independent cDNA clones. Bioinformatic evaluation of these sequences (Fig. 1) led to the identification of full-coding clones of already known proteins (>1000), and to cDNA clones lacking database hits, which are potential targets for full-length sequencing. Presumably novel cDNAs were 3′-EST sequenced and again analyzed for novelty. Out of the initial clones, 3630 cDNAs have been fully sequenced thus far, totaling 8.8 Mb. The sequence subset described here comprises 500 novel human cDNAs that are representations of the complete protein coding part of the original transcripts. Also the other fully sequenced cDNAs represent mostly genes that have not been fully sequenced elsewhere; however, the clones are not likely to contain the complete protein coding region of the respective transcripts, or they contain frame-shift mutations that have probably been introduced during reverse transcription in the cloning process. Therefore, these clones are only of reduced value for functional analysis. The number of bases reported for the 500 full-coding cDNAs is 1,264,620 bp; the average insert size of the clones is 2529 bp. The clones originate from five different cDNA libraries that have been sampled in varying numbers of clones (Table 1) to maximize the likelihood of identifying novel cDNAs.

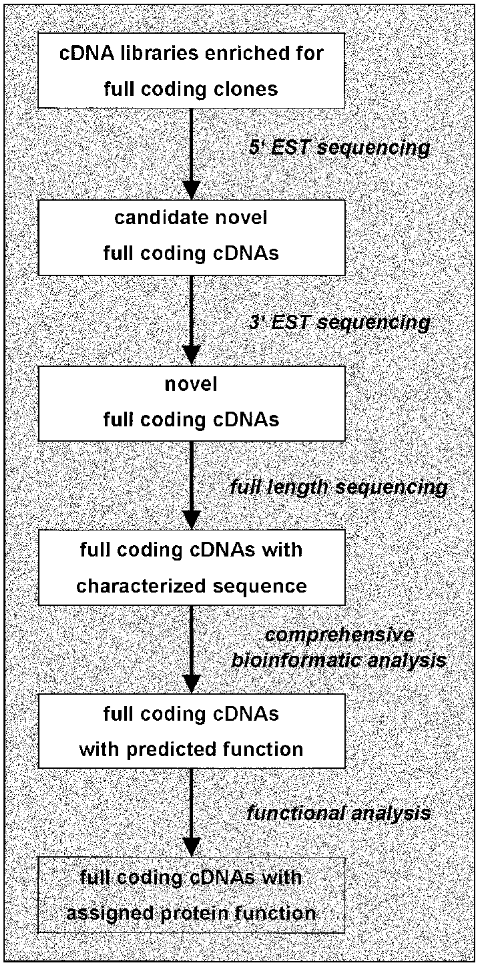
cDNA libraries enriched for full coding clones
5' EST sequencing
candidate novel full coding cDNAs
3' EST sequencing
novel full coding cDNAs
full length sequencing
full coding cDNAs with characterized sequence
comprehensive bioinformatic analysis
full coding cDNAs with predicted function
functional analysis
full coding cDNAs with assigned protein function

Figure 1 Flow of clones, sequences, and information in the German cDNA Consortium. 5′ EST sequences were systematically generated from the clones of 384-well microtiter plates and analyzed for hits in public databases. Clones with novel sequences were 3′-EST sequenced and these ESTs were analyzed again for novelty. Clones of uncharacterized transcripts were reported back to the sequencers who then did the full-length sequencing of cDNAs. The final sequence was analyzed comprehensively with bioinformatic tools and the outputs were evaluated manually. The clones feed functional analysis projects that take advantage of the clone resources generated.

The calculated average size of the encoded proteins was 470 amino acid residues, which equals the number that has been reported previously for some 1200 genes (Makałowski and Boguski 1998). There was, however, a wide variation between 66 and 1805 residues. The cDNA identifiers, the respective sequence ac-

Table 1. Library Distribution of cDNA Clones Analyzed

RZPD library identifier	Tissue	No. of clones	% of the clones reported	Average insert size (bp)	Average ORF size (aa)
DKFZp434	Testis	204	40.8	2766	562
DKFZp564	Fetal brain	142	28.4	2049	354
DKFZp566	Fetal kidney	43	8.6	2210	328
DKFZp586	Uterus	50	10.0	2506	492
DKFZp761	Amygdala (brain)	61	12.2	3055	506

cession numbers (EMBL/GenBank/DDBJ), cDNA sizes, the length of ORFs, the chromosomal location, and functional details for the individual cDNAs are broken down in Table 2. This table is available in its entirety at http://www.dkfz-heidelberg.de/abt0840/GCC.

Features of 5′- and 3′-Untranslated Regions

The 5′-untranslated regions (UTRs) averaged 148 nt, which is the same range as that reported previously (Pesole et al. 1996) but considerably shorter than the number (215 nt) calculated in the UTRdb (Pesole et al. 2000). There was a wide variation in size ranging up to >800 nt (e.g., DKFZp761F182). Even this long 5′-UTR was consistent with the scanning model for translational initiation (Kozak 1999) as there was no AUG codon in this stretch of sequence. In-frame stop codons upstream from the initiator ATG were present in 56.4% (282) of the cDNAs. This number is consistent with that observed with cDNAs isolated from oligonucleotide cap ligation libraries (Suzuki et al. 2000), where the cDNAs have been selected to contain the extreme 5′ ends of the respective transcripts. The overall GC content in the 5′-UTRs (56.3%) was considerably higher than that in the coding regions (52.6%) and the 3′-UTRs (45.7%). This is consistent with the finding that CpG islands frequently extend into the transcribed sequence (Cross and Bird 1995) whereas elements present in the 3′-UTR are often AU rich (Xu et al. 1997).

The average size of the 3′-UTRs was 926 nt [not including the poly(A) tail], which is considerably larger than the 388 nt and 820 nt reported by Makałowski and Boguski (1998) and Pesole et al. (1996), respectively. This discrepancy probably derives from the longer average size of the cDNAs described here, as compared with that observed in the previous studies. As with the 5′-UTR there was great variability with the size of the 3′-UTR. The translation terminator codon TAA could be part of the polyadenylation signal (e.g., in clone DKFZp564F2272) whereas in other cDNAs the 3′-UTR was found to be >4000 nucleotides (e.g., DKFZp486C1218).

We screened for the presence of repeat structures across the cDNA sequences. The *Alu* repeat family was most frequently contained in the cDNAs; 7.6 % (38) of the cDNA inserts carried this type of repeat. L1 repeats were present in two cDNAs; one cDNA contained both LTR2 and *Alu* repeats (DKFZp761G18121). The repeat structures were, without exception, located in the 3′-UTR of the respective cDNAs. However, in a number of other cDNAs we found repeats also in the presumed 5′-UTRs. All of these clones turned out to be not completely spliced and/or partial upon further analysis, and having intronic sequence at the 5′ ends. We therefore reason that the presence of repeat structures in 5′-UTRs of transcripts is rather rare. The lack of repeat structures in 5′ EST sequences has since been implemented as criterion in the selection process of cDNA clones that are targeted to full-insert sequencing to further increase the impact of the project.

Functional Classification

We grouped the cDNAs into functional classes according to homologies of their encoded proteins with already known proteins (Table 2 and Fig. 2): cell cycle, differentiation and development, membrane protein, metabolism, nucleic acid management, protein management, signaling and communication, structure and motility, transport and traffic, and unknown. Sequence annotations in databases sometimes were misleading, and the putative function of a protein could not be simply deduced by regarding the hit with the highest similarity as being the most significant. The integration of results from several search algorithms was necessary to draw relevant conclusions. For example, the deduced protein sequences were evaluated for the presence of specific (protein) sequence patterns necessary for the function/activity of a particular protein [e.g., the DFG/DWG and aPE motifs had to be present in a protein kinase, as reported by Hanks et al. (1988)]. The results of this functional classification are given in Table 2. The largest class constitutes proteins of unknown function (202 cDNAs, 41%). Considering that for another 39 cDNAs (8%) the only prediction that had been possible was that the deduced proteins would contain a putative transmembrane domain, no function could be inferred to a total of 241 cDNAs (48%) of the predicted proteins. But even if functional

Table 2. Functional Classification of Individual cDNAs[a]

Cell cycle											
cDNA data					Best database hit				Tissue specificity		
Clone ID	Accession no.	Contig size (bp)	ORF size (aa)	Chromosomal location	Description of best hit	Database accession no.	*P*-value	Gene family	Tissue	Score	# ESTs
DKFZp434A0530	AL136842	2768	254	2p22.1	gene: *Borg2;* product: "CRIB-containing BORG2 protein"; *Homo sapiens* CRIB-containing BORG2 protein (BORG2) mRNA, complete cds.	EMBL AF164118	2.1e-99				
DKFZp434A1135	AL122068	3010	670	5q13	*Homo sapiens* Rad 17-like protein (RAD17) mRNA, complete cds.	EMBL AF076838	0				
DKFZp434A1315	AL136755	1848	387	1q21.2	product: "F1N21.3"; The sequence of BAC F1N21 from *Arabidopsis thaliana* chromosome 1, complete sequence.	EMBL AC002130	5.7e-22				
DKFZp434B174	AL80146	1546	398	15q21.3	*Homo sapiens* mRNA for cyclin B2, complete cds.	EMBL AB020981	0		ear	6.38	6
DKFZp434G0514	AL136750	1503	379	4p16.2	cell growth regulating nucleolar protein LYAR—mouse	PIR A40683	2.7e-144				
DKFZp434H152	AL136840	4619	855	10p13	gene: *cdc23*; "SPBC1347.10"; product: "cell division cycle protein 23"; *S. pombe* chromosome II cosmid c1347.	EMBL AL035548	7e-21				
DKFZp434J037	AL136891	3443	628	1q32.1	gene: *KIAA0537;* product: "KIAA0537 protein"; *Homo sapiens* mRNA for KIAA0537 protein, complete cds.	EMBL AB011109	2.6e-148	protein kinase			
DKFZp434N0250	AL117525	1584	462	1q43-q44	product: "AKT3 protein kinase"; *Homo sapiens* AKT3 protein kinase mRNA, complete cds.	EMBL AF135794	2.1e-249	protein kinase			
DKFZp434P107	AL136894	2380	422	9q34	XPMC2 protein—African clawed frog	PIR S53818	5.9e-10				

Table 2. (*Continued*)

Cell cycle											
cDNA data					Best database hit				Tissue specificity		
Clone ID	Accession no.	Contig size (bp)	ORF size (aa)	Chromosomal location	Description of best hit	Database accession no.	p-value	Gene family	Tissue	Score	# ESTs
DKFZp434P2235	AL136860	2027	549	17q12	oncogene 1 (tre-2 locus) (clone 210)—human	PIR S22155	5.5e-226		testis	5.81	12
DKFZp564A0723	AL80116	2524	712	6q14.3-q16.1	gene: *ORC3L*; product: “origin recognition complex ORC3L subunit”; *Homo sapiens* origin recognition complex ORC3L subunit (ORC3L) mRNA, complete cds.	EMBL AF135044	0				
DKFZp564E2182	AL50261	2367	204	6q22.1-q22.33	*Homo sapiens* CGI-98 protein mRNA, complete cds.	EMBL AF151856	1.2e-265				
DKFZp564G1816	AL136599	4775	984	3q12.2-q12.3	gene: *KIAA0797;* product: “KIAA0797 protein”; *Homo sapiens* mRNA for KIAA0797 protein, partial cds.	EMBL AB018340	2.1e-50				
DKFZp564K142	AL136636	2241	335	17p11.2	*Rattus norvegicus* implantation-associated protein (IAG2) mRNA, partial cds.	EMBL AF008554	9.4e-184				
DKFZp564L0562	AL80090	941	185	4q31.21	*Homo sapiens* mRNA for APC10, complete cds.	EMBL AB012109	4.4e-178				
DKFZp564N0582	AL50264	1646	144	3p21.1	*Homo sapiens* DRR1 (DRR1) mRNA, complete cds.	EMBL AF089853	0		brain	5.16	50
DKFZp564N0582	AL50264	1646	144	3p21.1	*Homo sapiens* DRR1 (DRR1) mRNA, complete cds.	EMBL AF089853	0		retina	5.45	7
DKFZp566G0346	AL136719	4503	262	9q22.1	*Homo sapiens* spindlin mRNA, complete cds.	EMBL AF106682	0				

The cDNAs have been grouped into ten functional categories (see Statistics—Classification) based on sequence similarity data and have been grouped accordingly. The cDNA clones are available from the Resource Center of the German Genome project using the clone ID shown in the first column. The respective sequences are available at the EMBL/GenBank/DDBJ databases under the accession numbers shown in the second column. The third column provides the size of the individual cDNA inserts, and the fourth column shows the size of the encoded/predicted proteins. The chromosomal location of the respective genes is shown in the fifth column. Columns 6–8 describe database hits with the highest similarity: The accession number of the best hit (and the database where this hit was found), the description of the best hit, and the *P*-value of this hit is provided in these three columns, respectively. Similarities were predicted based on BLASTX and BLASTN2 analyses. Selection of the “representative = best” hit was done using the following criteria: (1) A BLASTX hit was judged better than a BLASTN hit. (2) In cases where the best BLASTX (only with TREMBL database) hit had been calculated from the same nucleotide sequence entry that was the best hit in the BLASTN analysis, the BLASTN hit is given, and (3) Only when no other hits were available, genomic sequence entries are given.
If classification of a protein to a major gene family was possible (based on similarity information), the respective family is shown in column 9. Based on the availability of EST information, tissue-specific expression of transcripts has been depicted in columns 10–13, showing the tissue, an arbitrary score (see WWW2001) and the absolute number of ESTs sequenced from that particular tissue (at the time of analysis), respectively.

[a] This section is excerpted from the full table, available on-line at http:www.dkfz-heidelberg.de/abt0840/GCC.

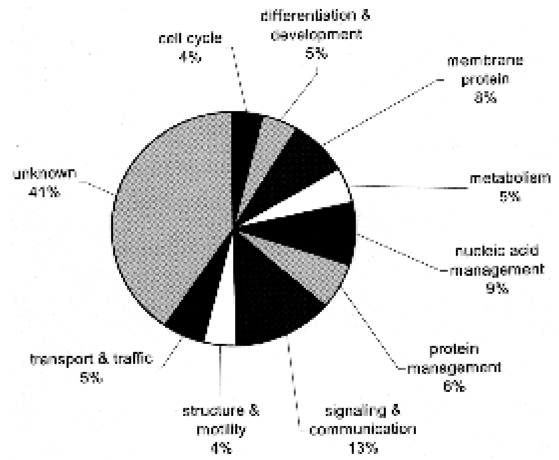

Figure 2 Functional classification of proteins encoded by the cDNAs. The deduced proteins were grouped into 10 functional categories based on sequence similarity with proteins of known function. The fraction of the 500 cDNAs grouped into the respective categories is indicated.

predictions were possible, the identification, for example, of a protein kinase, neither provides information on its substrates nor on the pathway(s) in which it is involved. Comprehensive functional analyses should be specifically indicated for a set of cDNAs encoding candidates for genes related to disease, such as putative GTP binding proteins, ion channels, and a cDNA encoding a protein that is highly similar to an oncogene.

We further analyzed the cDNAs for the presence of function-related sequence motifs to also identify novel members of gene families. We identified 41 potential leucine zipper proteins (Struhl 1989), 11 proteins with WD-domains (Neer et al. 1994), 11 proteins with predicted zinc finger domains (Parraga et al. 1988), 7 potential protein kinases, and 5 RNA helicases. The respective clones are indicated in Table 2 (column 9). Two cDNAs (DKFZp586I021 and DKFZp434O1826) contain both a WD-domain and a leucine zipper. A zinc-finger domain is predicted additionally for the deduced protein of the former cDNA.

Alternative Splicing

We found 39 (7.8%) cDNAs to represent putative splice variants of already known transcripts. This number is likely to represent the lower end of the fraction of transcripts that are alternatively spliced in vivo as any cDNAs representing already fully-known transcripts were excluded from further sequencing and alternative splice forms should therefore be under-represented in our set. We found ORFs with additional exons (e.g., DKFZp761B192), skipped exons (e.g., DKFZp564A032), and alternative exons including one containing the translational start codon and resulting in a different N terminus of the deduced peptide (e.g., DKFZp434J154). The percentage of alternatively spliced cDNAs appeared to be slightly higher in fetal brain, 40% of the alternatively spliced cDNAs originate from fetal brain whereas only 28% of all cDNAs analyzed originate from this tissue. This finding is consistent with reports by Sutcliffe and Milner (1988) and Hanke et al. (1999). The presence of intron sequences reminiscent in many cDNA sequences available in public databases, however, might lead to an overestimation of the extent of alternative splicing that is taking place in vivo. Experimental evidence will therefore be needed to confirm presumed alternative splice forms.

Representation of cDNAs in the UniGene Data Set

Depending on the true number of human genes, about 60%–90% have already been identified by partial sequencing of >2,000,000 cDNAs (EST sequencing). Overlapping EST sequences have been clustered to break down this large number of ESTs to comprehensive collections that should consist of nonredundant data sets having one representation (cluster) for every gene. The most widely accepted clustering data set is the UniGene (Schuler et al. 1996) resource at the NCBI (http://www.ncbi.nlm.nih.gov/UniGene/). This dataset currently consists of >90,000 clusters of mostly partial sequences. Consensus sequences of these clusters are available from http://www.rzpd.de. To investigate the representation of the novel cDNAs reported here in the UniGene data set and to evaluate the maximum number of genes that could be represented there, we aligned the full-length sequences with the UniGene database. The version of UniGene (Build 105) that was used in the analysis consisted of 92,931 clusters with 10,501 clusters containing known genes.

In total, 626 UniGene clusters matched with 472 out of the 500 full-coding cDNA sequences. The majority of cDNAs (342, 68%) was represented by one UniGene cluster. An additional 130 (26%) cDNAs were represented by 284 separate UniGene clusters (Fig. 3). Thus, a number of UniGene clusters could be linked by the full-length cDNA sequences. An example of three UniGene clusters that were joined with one cDNA is given in Figure 4. We analyzed the ESTs and clusters that were placed internal to the cDNAs reported here and found that most of the EST clones making up these clusters had originated from internal priming events (mostly in reminiscent intron sequences) and not from alternative polyadenylation. The number of 640 clusters that was hit with 472 cDNA sequences implies that there is ~35% redundancy in UniGene. As the average size of the human transcripts in general has been estimated to be in the same range as the average size of the cDNAs reported here (by quantification of Northern blots that had been hybridized with a labeled oligonucleotide dT probe; N. Nomura, pers. comm.), our finding should be representative. However, the true

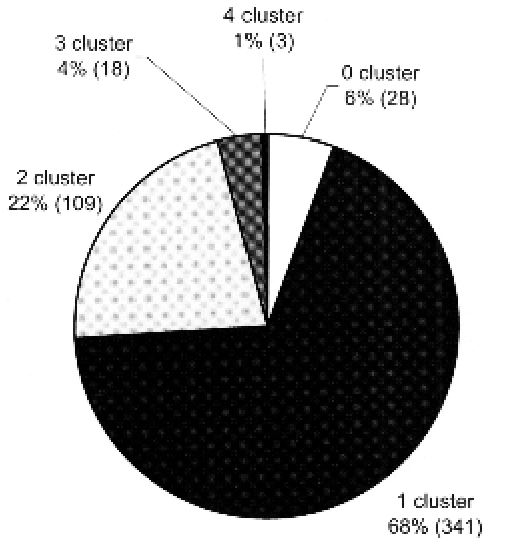

Figure 3 Representation of cDNAs in the UniGene data set (Build 105). Every cDNA was aligned with the UniGene data set to identify the number of EST clusters that was hit/joined with a given cDNA. The fraction and the total number (in parentheses) of the cDNAs are given for the varying numbers of clusters being hit.

number of genes represented in UniGene will further condense as a considerable fraction of the UniGene clusters are singletons (~39%), which are clusters made up by only one cDNA, and several of these will eventually turn out to be artifacts. Consequently, we estimate the number of independent genes that are represented in UniGene to be 50,000 at most.

A fraction of 6% (28 cDNAs) did not have hits in the UniGene database (cutoff, sequence identity >95% in 50 bp). The low number of the novel cDNAs without UniGene matches might in turn imply that >90% of all human genes were already represented in this database. However, we would rather assume that an unknown number of genes has escaped cloning and/or identification so far as the respective transcripts might be expressed only at extremely low levels or in very specialized cell types or differentiation stages. A proper selection of tissues or even single cell types for cDNA library production will be a critical issue for the detection and cloning also of these rarely expressed transcripts. For example, fetal brain, although very complex in expression, has been so deeply sampled in EST projects [especially the IMAGE 1NIB library (Soares et al. 1994)] but also in full-length cDNA sequencing (Nagase et al. 2000) that the novelty rate (3 of 142 cDNAs, 2%) is rather low in this tissue. In contrast, testis currently appears to have a higher potential for identifying transcripts not yet covered by ESTs (19 of 204 cDNAs, 9%).

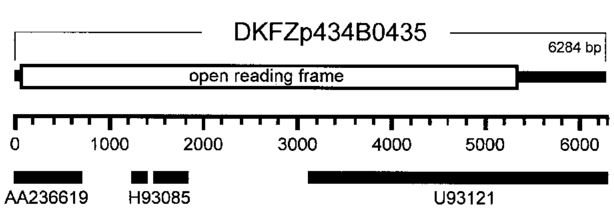

Figure 4 Three UniGene clusters are joined when aligned with the cDNA sequence DKFZp434B0435. The bar on top of the scale represents the cDNA with the open reading frame drawn as an open box. The bars below the scale represent the position and size (in bp) of the three UniGene clusters that are joined by the cDNA sequence. The accession nos. of representative sequences of the respective UniGene clusters are given below the bars.

Tissue Specificity of Expression

To analyze for a possible tissue specificity of expression we aligned the cDNA sequences with the EST database dbEST. ESTs originating from pooled tissues and tissues with unclear origin were excluded. Each cDNA received a score indicating the degree of tissue specificity. The higher this score, the higher the likelihood that expression of the particular transcript should be restricted to that tissue. A ubiquitously expressed transcript would have had a score of one. Only cDNAs with scores of five or higher are indicated in Table 2 (columns 10–12). In total, the expression of 22 transcripts appeared to be restricted to only one tissue with matching tissues of our cDNA and the ESTs (Table 2). Six brain-derived cDNAs only matched ESTs that had derived from brain tissues. Most of the cDNAs encode proteins that are either involved in the cell cycle or signaling pathways, for example, a stathmin-like protein and a protein similar to a calmodulin-binding protein. Only one of the six cDNAs encodes a protein of unknown function. Another 15 testis cDNAs had hits only with ESTs from testis/male genital tract. Although predictions could be made for three of the encoded proteins (a predicted sperm flagellar protein, a putative neurotransmitter transporter, and a possible nuclear pore protein), the other 12 cDNAs encode proteins of unknown function. The only uterus cDNA predicted to be specifically expressed in uterus/ovary encodes a putative chaperone-associated protease, which could indicate that this protein might be involved in the differentiation of the egg or embryo. The expression of several testis-derived transcripts appeared to be very selective as the scores calculated for these cDNAs were rather high, compared with scores obtained with other cDNAs and tissues (Table 2). This also matches the observation that the novelty rate, counting cDNAs without EST hits, was highest in the testis library (see above).

cDNAs Mapping to Human Chromosomes 21 and 22

To demonstrate the power of mapping genes by aligning cDNA with genomic sequences we downloaded the sequences of the first two completely sequenced human chromosomes 21 (Hattori et al. 2000) and 22

(Dunham et al. 1999) and aligned them with those novel cDNAs mapping to the respective chromosomes (Table 3). Clone identifiers of the respective cDNAs and the insert and ORF sizes are provided in the first three columns. For ORF sizes (column 3) the predicted number of amino acid residues is given first, followed by the number of the residues deduced from the cDNA sequence; a dash (-) is inserted for proteins that were not predicted. The predicted localization as based on mainly STS data is given in the fourth column, followed by the exact localization of the genes (gene locus in bp as defined in the published sequences of chromosome 21, http://hgp.gsc.riken.go.jp, and chromosome 22, http://www.sanger.ac.uk/cgi-bin/cwa/22cwa.pl). The accession numbers of the genomic clone(s) covering the genes, identifiers of predicted transcripts (if available; dashes indicate nonpredicted genes), the number of predicted exons out of the number of identified exons (based on cDNA sequence), and the number of UniGene clusters that were hit with the respective cDNAs are given in columns 6–9.

Whereas 13 of the novel cDNAs map to chromosome 22, only two cDNAs map to chromosome 21. This could either be a reflection of the generally higher gene content of chromosome 22 (554 compared with the 225 predicted genes on chromosome 21) or be a result of the fact that the percentage of genes that had been known previously is higher for chromosome 21 (this chromosome had long been carefully investigated because of its clinical implications, e.g., in Down syndrome). A third explanation could be a correlation between chromosomal location and global expression levels of the individual genes, as has been proposed by Ewing and Green (2000), with genes mapping to chromosome 21 in general possibly being expressed at lower levels compared with genes located on chromosome 22.

By combining the genomic and cDNA data, the exact gene structures of all 15 cDNAs could be determined. Although all cDNAs were covered by UniGene clusters, only 8 of the 15 genes had been predicted from the genomic sequence. Most of these gene predictions were precise, identifying the majority or all exons. The number of amino acid residues varied in most cases only marginally from the number deduced from the cDNA sequence. However, one cDNA (DKFZp564B212) merged three predicted transcripts to only one gene and overlapped another gene (bK445C9.C22.3) predicted on the opposite strand. In total, seven genes had completely failed to be predicted, some of which encode rather large ORFs and consist of several exons.

The mapping information that is based on genomic sequence not only gives the exact localization of individual genes but also provides information on the context of these genes in view of neighboring genes (e.g., DKFZp434B194 and DKFZp564B212 are only 13 kb apart) and the presence of probable additional gene copies. For example, the genes of cDNAs DKFZp434N035 and DKFZp434P211 appear to be present on chromosome 22 in 2 and 9 highly similar copies (>85% sequence identity on nucleotide level), respectively. DKFZp434P211 could indicate a cluster of highly similar POM121 related genes (Fig. 5), the first of which was described by Kawasaki et al. (1997). Two copies (2850458 and 2871777) seem to be ancient and inactive as they are incomplete, contain several frame shifts, and share only 89% and 87% sequence identity with the cDNA sequence in exon 1, respectively. The other copies are highly similar (>95% identity on nucleotide level). Further experiments will be necessary to investigate how many of the gene copies are expressed and to explain the presence of the stop codon at position 429 in three of the gene copies (and in the cDNA) but a sense codon in this position in four other gene copies, possibly leading to an extended protein product. EST evidence is available for transcripts of both types of genes (e.g., for copies 5055694 and 8220566).

DISCUSSION

The considerable fraction of genes that were not predicted in the analysis of the chromosome 21 and 22 sequences was somewhat surprising, as EST data and UniGene clusters (Table 3) were also available for these genes. Three of the genes that were not predicted even appear to be present in more than one copy on the same chromosome, namely, within 6 Mb on chromosome 22. But even if all genes could be identified via bioinformatic procedures, the alternative use of exons and promoters (alternative splicing) constitutes a problem that cannot currently be solved with knowledge of the genomic sequence alone. Consequently, only the availability of cDNA sequences enables us to define the precise protein coding parts of the genome and, in conjunction with the genomic counterpart, to also define the composition of exons in alternatively spliced transcripts of the same gene. Both the sequence and the chromosomal location of genes are important pieces of information supportive also in the process of defining and analyzing candidate disease genes.

Most of the genome has been unraveled as draft sequence, where sequence submissions of individual genomic clones are released in several contigs of varying length. These contigs are usually not ordered relative to one another. However, automated assembly and annotation tools like GoldenPath (http://genome.ucsc.edu/goldenPath/hgTracks.html) try to overcome this problem and prove to be extremely helpful for the mapping of cDNAs. The availability of cDNA sequences in turn immediately helps to identify the genes that are located on the respective

Table 3. Analysis of Gene Structures of cDNAs Mapping to Human Chromosomes 21 and 22

Clone D	Contig size (bp)	ORF size (aa)	Predicted chromosomal location	Chromosomal location relative to the published genomic sequences	Accession no. of genomic sequence entry	Annotation of predicted transcript	No. predicted/ No. true exons	No. of UniGene clusters
CHROMOSOME 21								
DKFZp434N0650	1095	−/186	21q	2914115–2890884	HS21C104N	—	0/4	1
DKFZp566A221	568	108/108	10	12684965–12674544	AL023494 M37104	M37104	4/4	2
CHROMOSOME 22								
DKFZp434B194	2876	838/837	22q12.1	10481146–10460626	Z95115	bK1048E9.C22.2	15/15	3
DKFZp434F0116	2377	−/478	22q12.3–13.1	23317974–23333542	AL022312N	—	0/7	2
				26764656–26674550	AL049758	—	0/10	2
DKFZp434H1130	3176	−/445	22q13.2–q13.33					
DKFZp434N035	1978	−/160	22q11.2	4629050–4632539	AC007308N AP000557N	—	0/8	2
				5394302–5397804		AP000557.C22.1	(1/8)	
DKFZp434P211	5024	−/428	22q11.2	2699820–2701392	AC008132/8103	—	0/8	4
				2850458–2713469	AC007326/8103	—		
				2871777–2870723	AC007326	—		
				4617790–4619361	AC007050	—		
				5055694–5042046	AC002472	—		
				5210585–5224231	AP000552	—		
				6560307–6558737	D87013/D8700	—		
				8220566–8234258	AP000354	—		
				8627661–8626091	AP000356	—		
DKFZp564B212	1915	−/377	22q12.1	10558798–10494469	Z95115	bK1048E9.C22.2 bK445C9.C22.7 bK445C9.C22.4 (bK445C9.C22.3)	7/7	1
DKFZp564F1978	1129	127/126	22q12–13	22474505–22491683	AL021707	dJ508I15.C22.1	5/5	1
DKFZp564G1978	1662	422/423	22q13.1–13.2	26895167–26845385	AL022476	dJ323M22.C22.2.a	11/11	2
DKFZp564K2478	1874	−/372	22q11	2498885–2525670	AC008079	—	0/11	2
				[2612231–2611609]		—		
				[2716855–2717477]		—		
				(4308401–5120712]		—		
				[5227647–5228269]		—		
DKFZp586H2219	1971	476/475	22q11.2-qter	29510375–29682479	Z95331 Z93784 Z84478	bK941F9.C22.2	12/12	1
DKFZp586K0922	3477	639/617	22q12.2	15147736–15179066	AC002073	AC002073.C22.1c	16/15	2
DKFZp761I141	3071	588/617	22q13.31–13.33	25023143–25048485	AL035658	dJ756G23.3	16/16	4
DKFZp761O17121	2690	−/212	22q13.31	28333768–28331093	dJ1033E15.1	—	0/2	1

```
               1                                                                                                         100
DKFZp434P211   MDSLWGPGAG SHPFGVHNTR LSPDLCPGKI VLRALKESG- -AGMPEQDKD PRVQENPGDQ RRVPEVTGDA RPAFRPLRDN GGLSPFVPGP --GPLQTDLH
      D87002   .......... .......... .......... .........- -......... .......... .......... PS........ .......... --........
     2699820 + .......... .......... .......... .........- -......... .......... .......... PS........ R......... --........
     2850458 - K........R .Q...A.... ....S..E.. ......D.R- -......... .G.....D.. ....QG.... PS.....W.. .......SR. --...ER...
     2871777 - K......... .....AHNIQ LCPRLLSRED RVEGPPEEGG RDAAGQGPQS PRESSEKGSG HRGCTVCIAP VGQWRPLSLR AQAWASAERP PCPEVRSRIK
     4617790 + .......... ........S. .......... .........- -......... .......... .......... PS........ .......... --........
     5055694 - .......... .......... .......... .........- -......... .......... .......... PS........ .......... --........
     5210585 + .......... ........S. .......... .........- -......... .......... .......... PS........ R......... --........
     6560307 - .......... .......... .......... .........- -......... .......... .......... PS........ .......... --........
     8220566 + .......... .......... .......... .........- -......... .......... .......... PS........ .......... --........
     8627661 - .......... .......... .......... .........- -......... .......... .......... PS........ ........R. --........

               101                                                                                                       200
DKFZp434P211   AQRSEIRYNQ TSQTSWTSSC TNRNAISSSY SSTGGLPGLK RRRGPASSHC QLTLSSSKTV SEDRPQAVSS GHTQCEKAAD IAPGQTLTLR NDSSTSEASR
      D87002   .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
     2699820 + .......... .......... .......... .......... .......... .......... .......... .......... .......A.. ..........
     2850458 - .....VT... R..S..M..F PK...*
     2871777 - PEIPDLLDEF VHQTKCHREL LQFHGRLPVT KEEGASLIPL SAAPQVLKDS EGPASGCLLG SHPVKGSRCS TRADTCPQEW LPHIPGLAPM QVSSAATQAR
     4617790 + ........D. S......... .......... .......... .......... .......... .......... .......V.E K......A.. ....R.....
     5055694 - .......... .......... .......... .......... .......... .......... .......... .......... .......A.. ..........
     5210585 + .......... .......... .......... .......... .......... .......... .......... .......... ....-..A.. ..........
     6560307 - .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
     8220566 + .........P .......... .......... .......... .......... .......... .......... .......... .......... ..........
     8627661 - .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........

               201                                                                                                       300
DKFZp434P211   PSTHKFPLLP RRRGEPLMLP PPLELGYRVT VEDLDREKEA AFQRINSALQ VEDKAISDCR PSRPSHTLSS LATGASGLPA VSKAPSMDAQ QETHKSQDCL
      D87002   .......... C......... .......... .......... .......... .......... .......... .......... .......... ..........
     2699820 + .......... .......... .......... .......... .......... .......... .......... ....T..... I.......V. ..........
     2850458 -
     2871777 - GASDAATSLT AGVWGHCRPG PGEEGGTLVQ KRTAGGQGHL GLQTLTAFPH FVLTCNRDFL SACCMSTQHG CTAGETQAPR *
     4617790 + ...R...... H......... ..V....... A....W.... ...C.K.... .......... .......... .......... .......... ..R.......
     5055694 - .......... .......... .......... .......... .......... .......... .......... .........P .......... ..........
     5210585 + .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
     6560307 - .......... C......... .......... .......... .......... .......... .......... .......... .......... ..........
     8220566 + .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
     8627661 - .......... .......... .......... .......... .......... .......... ..W....... .......... ....S..... ..........

               301                                                                                                       400
DKFZp434P211   GLLDPLASAA GVPSTAPMSG KKHRPPGPLF SSSDPLPATS SDSQDSAQVT SLIPAPFPAA SMDAGMRRTR HGTSAPAAAA AAPPRSTLNP TLGSLLEWME
      D87002   ...A...... E......... .......... .......... .H........ .......... .......... R......... ....P.A... ..........
     2699820 + ...A...... .......... ...K...... .......... .H........ .......... ...V...... C......... ....P..... ..........
     2850458 -
     2871777 -
     4617790 + ..VA.....T E......... E......... .......... .H........ .......... .......... P......... ....P...R. ..........
     5055694 - ...A...... .......... .......... .......... .H........ .......... .......... C......... ....P..... ..........
     5210585 + ...A...... .......... .......... .......... .H........ .......... .......... C......... ....P..... ..........
     6560307 - ...A...... E......... .......... .......... .H........ .......... .......... R......... ....P.A... ..........
     8220566 + .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
     8627661 - ...A...... E......V.. .......... .......... .......... .......... .......... R........T ....P..... ..........

               401
DKFZp434P211   ALHISGPQPQ LQQVPRGQNQ RSQTSWTSSC PK*
      D87002   .......... .......... .....R.... ..*
     2699820 + .......... .......... .....R.... ..RNAISSPY RSTGGLPERK RRRGPASSHC QLNLSS*
     2850458 -
     2871777 -
     4617790 + .......... .......... .....R.... ..RNAISSSC SSTGDLPGRK RKRRQPHPTA S*
     5055694 - .......... .......... .....R.... ..RNAISSPY RSTGGLPERK RRRGPASSHC QLTLSS*
     5210585 + .......... .......... .....R.... ..RNAISSPY RSTGGLPERK RRRGPASSHC QLTLSS*
     6560307 - .......... .......... .....R.... ..*
     8220566 + .......... .......... .......... ..*
     8627661 - .......... .......... .....R.... ..*
```

Figure 5 Multiple sequence alignment of cDNA DKFZp434P211 with POM121-related 1 (accession no. D87002) and sequences from chromosome 22 demonstrate the presence of a cluster of POM121-related genes. The individual genomic sequences were named after the start of the first exon relative to the cDNA: The open reading frame (ORF) was defined according to the predicted protein of the cDNA and of POM121-related 1. Genes located on the plus and minus strands of chromosome 22 are indicated with + and −, respectively. The cDNA sequence of DKFZp434P211 was taken as reference; identical residues in other sequences are indicated with a dot, residues deviating from the consensus are printed. Asterisks (*) indicate stop codons. The genomic sequences 2850458 and 2871777 are in italics because these copies deviate from the other copies by a premature stop or frame shifts and a large insertion, respectively, and are probably not expressed. In these two gene copies the initiator ATG is mutated. Dashes (-) were inserted by the software (CLUSTAL) to optimize the alignment.

genomic clones, to support the ordering of the draft sequence contigs, and to narrow down the regions where putative regulatory elements should reside. Thus, cDNA and genomic sequences are complementary and synergistically add information. The BLAST analysis of cDNAs and matching genomic sequences showed that only 32 cDNAs did not have corresponding genomic matches (not covered, NC in Table 2, column 5), which is the number expected because >91% of the genomic sequence are reported to be unraveled. The chromosomal localization could be approximated for 449 cDNAs using the GoldenPath web browser; 21 BACs had not been mapped (NM). The accession numbers of these BACs are provided in column 5 of Table 2. The combination of genomic and cDNA sequence provides the gene structures with precise exon–intron boundaries and defined intron sequences.

Furthermore, it will become increasingly important to not only have the human genes identified but rather to characterize the precise functions of the encoded proteins and also the functions of those transcripts that are not translated. To this end, full-coding cDNA representations are indispensable tools, for example, for the subcloning of exactly defined ORFs into expression vectors. However, currently only ~11,000 nonredundant cDNA sequences have been deposited in public databases which are supposed to contain the complete protein coding ORF. An even lower number of these full-coding ORFs can be obtained as cDNA clones through commercial or noncommercial providers (e.g., ATCC, Genome Systems, Research Genetics, HGMP, Resource Center of the German Genome Project) and would thus be available for functional research.

Recently, the range of estimates given for the number of human genes has evolved to the lower end, because in two calculations only ~35,000 human genes have been predicted (Ewing and Green 2000; Roest Crollius et al. 2000). Our data would also hint at a lower than previously expected number, as we would estimate the number of genes currently represented in UniGene to be 50,000 at most. Still, the real number of human genes needs to be established by further cDNA and also by comparative genomic sequencing (e.g., of the mouse). If it should hold true, however, that the number of genes in human was indeed only about twofold higher than the ~18,000 genes that have been predicted for *Caenorhabditis elegans* by The *C. elegans* Sequencing Consortium (1998) the question would arise as to where the difference in complexity between these two life forms originated. Because the sheer doubling of gene number would not be likely to account for all differences, the comprehensive analysis of gene and protein function(s) would become an even greater problem. This is because one solution to this apparent paradox could be the acquisition of multiple functions by many of the proteins expressed in human. This would add another order of complexity to the line starting with the genome and continuing through the transcriptome with alternative splicing, the proteome with post-translational modifications, and finally (?) to a ‘functiome,’ which would cover the acquisition of diverse functions by the same protein depending on its cellular and subcellular environment. Several examples of such multiple usages of proteins have already been described (Jeffery 1999).

In the set of 500 novel cDNAs described here, only about half of the deduced proteins could be functionally classified, while identification, for example, of a protein kinase does not provide information on substrates or pathways in which this protein is involved. Additionally, half of the predicted proteins remain without any hint as to their possible function. With this in mind, the establishment of a gene catalog which will eventually contain a nonredundant set of full-coding cDNA sequences and clones covering every human gene, is prerequisite to carry out the experiments needed to precisely identify the protein function(s). This catalog should be the result of a global enterprise integrating the data and clones from as many projects and researchers as possible and could be an extension of already existing databases such as GeneCards (Rebhan et al. 1998) and RefSeq (Pruitt et al. 2000) with, for example, links to the clone providers mentioned above. In addition to the novel full-coding cDNA sequences and clones described here, we have identified over 1000 cDNAs which comprise full-coding representations of previously known genes. In combination, these cDNAs represent 2%–5% of all human genes and will thus be a substantial part of the catalog and be ideal tools to carry out functional analyses. Although the 500 novel cDNAs have been fully sequenced and can be directly used in functional analysis, the cDNAs representing known genes need further characterization because these are not fully sequenced. To this end, we amplify the ORFs from these cDNAs and verify the predicted size. These ORFs are then cloned into a bacterial expression vector which contains a N-terminal fusion with the GFP. As the Gateway system (Life Technologies) is employed in the cloning process, the ORFs can be shuttled into any expression vector (Simpson et al. 2000). Only intact reading frames (no PCR frame shifts, no introns, no frame shifts in the clone) lead to fluorescent colonies as the ORF extends uninterrupted into the GFP. The Gateway entry clones of the verified genes are also made available through the Resource Center.

To address the systematic functional analysis of the novel proteins, a large-scale project dealing with the subcellular localization and functional analysis of the proteins encoded by newly identified cDNAs reported here is underway (Simpson et al. 2000). Thus, the gene catalog in upcoming years will form the basis for the large-scale and comprehensive functional analysis of human genes and proteins, which is crucial to understand the basis of human life, disease, and death.

METHODS

Library Construction

SMART Libraries

The DKFZp564 (human fetal brain) and DKFZp566 (human fetal kidney) libraries were generated using the SMART kit (Clontech). PCR amplification of the cDNA was necessary to obtain enough cDNA for cloning. The first-strand primer did contain the KS sequence of the pBluescript vector (Stratagene) and any base but T (IUB code = V) in the 3′-terminal position of the primer [TCGAGGTCGACGGTATCGATAAG$(T)_{19}$V]. Amplification of the primary cDNA with Amplitaq (Perkin Elmer) and Pfu (Stratagene) DNA polymerases in a ratio of 19/1 (vol/vol) was carried out with primers that contained uracil residues (3′ primer: CAUCAUCAUCAUCGAGGTCGAC GGTATCGATAAG; 5′ primer: CUACUACUACUATACGCT GCGAGAAGACGACAGAA) and that were compatible with the pAMP1 (Life Technologies) cloning sites for directional cloning. Prior to cloning, the cDNA was size fractionated on an agarose gel. Fragments >2 kb were excised and extracted from the gel using GELase (Epicentre). Cloning was done using uracil deglycosilase (UDG, LifeTechnologies) and chemically competent bacterial cells (XL-2 Blue, Stratagene).

Conventional Libraries

The DKFZp434 (human adult testis), DKFZp586 (human adult uterus), and DKFZp761 (human adult amygdala) libraries were generated using conventional approaches (Gubler and Hoffman 1983), employing a *Not*I-dT V primer for first-strand synthesis [GAGCGGCCGC$(T)_{19}$V]. After second-strand synthesis, *Sal*I adapters were ligated to the blunted cDNA. Then the cDNA was cut with *Not*I to generate *Sal*I–*Not*I-compatible ends at the 5′ and 3′ ends of the cDNA, respectively, to allow directional cloning. The cDNAs were then size-selected on agarose gels in two dimensions and cloned into pSPORT1 precut with *Sal*I and *Not*I (Life Technologies).

Availability of cDNA Libraries and Clones

All libraries have been arrayed into 384-well microtiter plates and spotted on high-density nylon membranes. Each library consists of 27,000 clones or multiples thereof. High-density clone filters and individual clones are available through the Resource Center of the German Genome Project (http://www.RZPD.de; clone@pzpd.de).

Selection of Clones for Sequencing

First, 5′ ESTs were systematically generated from all clones of 384-well microtiter plates. The sequences were analyzed with BLASTN (Altschul et al. 1990) and BLASTX (Gish and States 1993) against EMBL, PIR, SWISSPROT, and TREMBL databases for the lack of identical (>95% identity over 50 bp) matches with known cDNAs, and for the presence of ORFs.

Clones with novel sequences were 3′ end sequenced. These 3′ ESTs were checked for the lack of matches with known genes in public databases, for repeat structures, and for the presence of polyadenylation signals. Clones matching the selection criteria were subjected to full-length sequencing.

Sequencing Methodology and Strategy

Sequencing was done preferentially using dye terminator chemistry (Applied Biosystems or Amersham) on ABI 377 automated DNA sequencers; one partner used EMBL prototype instruments (Wiemann et al. 1995) mainly with dye primer chemistry. Primer walking (Strauss et al. 1986) was the preferred sequencing strategy for the full-length sequencing of cDNAs. Design of walking primers was done preferentially using software (e.g., Schwager et al. 1995; Haas et al. 1998) that permitted the complete automation of this usually-time-consuming process and thus helped in the parallel processing of large numbers of clones.

Bioinformatic Analysis

Every complete cDNA sequence was compared with the sequences in EMBL, EMBL-EST, EMBL-STS using BLASTN (Altschul et al. 1990). Searches against EMBL were done to determine whether the cDNAs were already known and to identify any genomic sequence information available that would cover the respective genes. Searches against EMBL-EST were performed to analyze for the abundance of transcripts, to obtain information on a possible tissue specificity of expression, and to identify putative alternative splice forms or alternative use of polyadenylation signals. The annotations on the source tissue of the respective EST clones were parsed from the database entries to calculate the real ratio versus the expected ratio of expression according to the equation: (# hits tissue/total # hits)/(# ESTs tissue/total # ESTs). A gene that was transcribed at a constant level in many tissues would have a ratio of one. Significant higher or lower ratios would indicate increased or decreased levels of transcription in the tissue, respectively. To identify tissue-specific expression, the parameters were set to >4 ESTs matching the respective cDNA that needed to have been sequenced from a given tissue, and the cutoff for the ratio of overexpression was set to five. ESTs originating from pooled tissues or that were of unspecified origin were disregarded in this analysis. To obtain chromosomal mapping information, the sequences were aligned with the EMBL-STS database.

The potential protein-sequences were identified by a search for the longest ORF in each of the three forward frames with a minimum length of 90 codons. The deduced protein sequences were searched against the nonredundant protein data set of PIR, SWISSPROT, and TREMBL [BLASTP, using the SEG-filter by Wootton (1994)]. Any cDNAs without ORF >90 codons were analyzed with BLASTN against TREMBL to identify even shorter ORFs present.

BLASTX searches were performed against a nonredundant protein database comprising PIR, SWISSPROT, and TREMBL. The SEG-filter was used to screen for potential frame shifts in the coding sequences of the cDNAs and to identify cDNAS that were not fully spliced or were alternatively spliced. The protein sequence was then transferred to PEDANT (Frishman and Mewes 1997). PEDANT performed automated database searches: psiBLAST (Altschul et al. 1997), an iterated profile search procedure; HMMER (Sonnhammer et al. 1997), a Hidden Markov model software which uses statistical descriptions of a sequence family's consensus; and BLIMPS (Wallace and Henikoff 1992) for similarity searches against the BLOCKS (Henikoff et al. 2000) database. PROSITE protein sequence patterns were identified by ProSearch (Kolakowski et al. 1992). CLUSTAL-W (Thompson et al. 1994) was used for multiple sequence alignments of DNA and proteins. Transmembrane regions were identified by ALOM2 (Klein et al. 1984), and signal peptides in secreted proteins by SIGNALP (Nielsen et al. 1997). SEG (Wootton and Federhen 1993) has been employed to detect low-complexity regions in protein sequences and COILS (Lupas et al. 1991) for the detection of coiled coils. For the functional classification of the cDNAs sequence, identities with E-values $<10E-30$ (BLASTN) and $<10E-10$ (BLASTX) were accepted to be significant. The comprehensive bioinformatic data on all cDNAs analyzed by the Consortium are accessible at http://www2.mips.biochem.mpg.de/proj/cDNA/index.html. Mapping of the cDNAs to chromosomes was done first by BLAST analysis of the cDNA sequences against the human genomic sequence (NCBI–htgs database), followed by identifying the mapping position with help of the GoldenPath (Jim Kent, UCSC) browser (http://genome.ucsc.edu/goldenPath/hgTracks.html).

Availability of Clones and Further Information

All clones described here, and the other clones analyzed by the German cDNA Consortium, are available from the Resource Center of the German Genome Project(http://

www.rzpd.de; clone@rzpd.de). The comprehensive bioinformatic data on all cDNAs analyzed by the Consortium are accessible at http://www2.mips.biochem.mpg.de/proj/cDNA/index.html. Additional information about the analysis of the described set of cDNAs is available at http://www.dkfz-heidelberg.de/abt0840/GCC. The full version of Table 2 can be obtained at this location in Excel, tab-delineated text, and pdf formats.

ACKNOWLEDGMENTS

We thank Christian Gruber, Oliver Heil, Lars Ebert, Daniel Bongartz, and Antje Krause for support with the bioinformatic analysis and presentation of the data, and Andreas Weller for encouraging discussions and support. This work has been supported by the Federal German Ministry of Education and Research (BMBF) through Projektträger DLR, within the framework of the German Human Genome Project (FKZ 01KW9705/07/10–16), and in part by the European Union (BIOMED 2 – BMH4-CT97–2284).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. 1998. New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282:** 682–689.

Cross, S.H. and Bird, A.P. 1995. CpG islands and genes. *Curr. Opin. Genet. Dev.* **5:** 309–314.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25:** 232–234.

Fields, C., Adams, M.D., White, O., and Venter, J.C. 1994. How many genes in the human genome? *Nat. Genet.* **7:** 345–346.

Frishman, D. and Mewes, H.-W. 1997. PEDANTic genome analysis. *Trends Genet.* **13:** 415–416.

Gish, W. and States, D.J. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3:** 266–272.

Gubler, U. and Hoffman, B.J. 1983. A simple and very efficient method for generating cDNA libraries. *Gene* **25:** 263–269.

Haas, S., Vingron, M., Poustka, A., and Wiemann, S. 1998. Primer design for large scale sequencing. *Nucleic Acids Res.* **26:** 3006–3012.

Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann, G., Luft, F., Reich, J., and Bork, P. 1999. Alternative splicing of human genes: More the rule than the exception? *Trends Genet.* **15:** 389–390.

Hanks, S.K., Quinn, A.M., and Hunter, T. 1988. The protein kinase family: Conserved features and deduced phylogeny of the catalytic domains. *Science* **241:** 42–52.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., et al. 2000. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405:** 311–319.

Henikoff, J.G., Greene, E.A., Pietrokovski, S., and Henikoff, S. 2000. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* **28:** 228–230.

Jeffery, C.J. 1999. Moonlighting proteins. *Trends Biochem. Sci.* **24:** 8–11.

Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Schmeits, J.L., Wang, J., and Shimizu, N. 1997. One-megabase sequence analysis of the human immunoglobulin λ gene locus. *Genome Res.* **7:** 250–261.

Klein, P., Kanehisa, M., and DeLisi, C. 1984. Prediction of protein function from sequence properties. Discriminant analysis of a data base. *Biochim. Biophys. Acta* **787:** 221–226.

Kolakowski, L.F., Jr., Leunissen, J.A., and Smith, J.E. 1992. ProSearch: Fast searching of protein sequences with regular expression patterns related to protein structure and function. *Biotechniques* **13:** 919–921.

Kozak, M. 1999. Initiation of translation in prokaryotes and eukaryotes. *Gene* **234:** 187-208.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. 2000. Gene index analysis of the human genome estimates approximately 120, 000 genes. *Nat. Genet.* **25:** 239–240.

Lupas, A., Van Dyke, M., and Stock, J. 1991. Predicting coiled coils from protein sequences. *Science* **252:** 1162–1164.

Makałowski, W. and Boguski, M.S. 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci.* **95:** 9407–9412.

Nagase, T., Kikuno, R., Ishikawa, K., Hirosawa, M., and Ohara, O. 2000. Prediction of the coding sequences of unidentified human genes. XVII. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. *DNA Res.* **7:** 143–150.

Neer, E.J., Schmidt, C.J., Nambudripad, R., and Smith, T.F. 1994. The ancient regulatory-protein family of WD-repeat proteins. *Nature* **371:** 297–300.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10:** 1–6.

Nomura, N., Miyajima, N., Sazuka, T., Tanaka, A., Kawarabayasi, Y., Sato, S., Nagase, T., Seki, N., Ishikawa, K., and Tabata, S. 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:** 47–56.

Parraga, G., Horvath, S.J., Eisen, A., Taylor, W.E., Hood, L., Young, E.T., and Klevit, R.E. 1988. Zinc-dependent structure of a single-finger domain of yeast ADR1. *Science* **241:** 1489–1492.

Pesole, G., Grillo, G., and Liuni, S. 1996. Databases of mRNA untranslated regions for metazoa. *Comput. Chem.* **20:** 141–144.

Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Larizza, A., Makalowski, W., and Saccone, C. 2000. UTRdb and UTRsite: Specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.* **28:** 193–196.

Pruitt, K.D., Katz, K.S., Sicotte, H., and Maglott, D.R. 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16:** 44–47.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. 1998. GeneCards: A novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* **14:** 656–664.

Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet.* **25:** 235–238.

Schuler, G.D. 1997. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75:** 694–698.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.

Schwager, C., Wiemann, S., and Ansorge, W. 1995. GeneSkipper:

Integrated software environment for DNA sequence assembly and alignment. *HUGO Genome Digest* **2:** 8–9.

Simpson, J., Wellenreuther, R., Poustka, A., Pepperkok, R., and Wiemann, S. 2000. Systematic subcellular localization of novel proteins identified by large scale cDNA sequencing. *EMBO Rep.* **1:** 287–292.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.

Sonnhammer, E.L., Eddy, S.R., and Durbin, R. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* **28:** 405–420.

Strausberg, R.L., Feingold, E.A., Klausner, R.D., and Collins, F.S. 1999. The mammalian gene collection. *Science* **286:** 455–457.

Strauss, E.C., Kobori, J.A., Siu, G., and Hood, L.E. 1986. Specific-primer-directed DNA sequencing. *Anal. Biochem.* **154:** 353–360.

Struhl, K. 1989. Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. *Trends Biochem. Sci.* **14:** 137–140.

Sutcliffe, J.G. and Milner, R.J. 1988. Alternative mRNA splicing: The Shaker gene. *Trends Genet.* **4:** 297–299.

Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H., Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., Isogai, T., et al. 2000. Statistical analysis of the 5′ untranslated region of human mRNA using "Oligo-Capped" cDNA libraries. *Genomics* **64:** 286–297.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Wallace, J.C. and Henikoff, S. 1992. PATMAT: A searching and extraction program for sequence, pattern and block queries and databases. *Comput. Appl. Biosci.* **8:** 249–254.

Wiemann, S., Stegemann, J., Grothues, D., Bosch, A., Estivill, X., Schwager, C., Zimmermann, J., Voss, H., and Ansorge, W. 1995. Simultaneous on-line DNA sequencing on both strands with two fluorescent dyes. *Anal. Biochem.* **224:** 117–121.

Wootton, J.C. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.* **18:** 269–285.

Wootton, J.C. and Federhen, S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17:** 149–163.

Xu, N., Chen, C.Y., and Shyu, A.B. 1997. Modulation of the fate of cytoplasmic mRNA by AU-rich elements: Key sequence features controlling mRNA deadenylation and decay. *Mol. Cell. Biol.* **17:** 4611–4621.

Received July 6, 2000; accepted in revised form December 29, 2000.