Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence

Mark D. Adams, Anthony R. Kerlavage, Robert D. Fleischmann, Rebecca A. Fuldner, Carol J. Bult, Norman H. Lee, Ewen F. Kirkness, Keith G. Weinstock, Jeannine D. Gocayne, Owen White, Granger Sutton, Judith A. Blake, Rhonda C. Brandon, Man-Wai Chiu, Rebecca A. Clayton, Robin T. Cline, Matthew D. Cotton, Julie Earle-Hughes, Leah D. Fine, Lisa M. FitzGerald, William M. FitzHugh, Janice L. Fritchman, N. S. M. Geoghagen, Anna Glodek, Cheryl L. Gnehm, Michael C. Hanna, Eva Hedblom, Paul S. Hinkle Jr., Jenny M. Kelley, Karin M. Klimek, John C. Kelley, Li-ing Liu, Simos M. Marmaros, Joseph M. Merrick, Ruben F. Moreno-Palanques, Lisa A. McDonald, Dave T. Nguyen, Susan M. Pellegrino, Cheryl A. Phillips, Sean E. Ryder, John L. Scott, Deborah M. Saudek, Robert Shirley, Keith V. Small, Tracy A. Spriggs, Teresa R. Utterback, Janice F. Weidman, Yi Li*, Ray Barthlow, Daniel P. Bednarik*, Liang Cao*, Mario A. Cepeda*, Timothy A. Coleman*, Erin-Joi Collins*, Donna Dimke*, Ping Feng*, Andrew Ferrie*, Carrie Fischer*, Gregg A. Hastings*, Wei-Wu He*, Jing-Shan Hu*, Kathleen A. Huddleston, John M. Greene*, Joachim Gruber*, Peter Hudson*, Ann Kim*, Diane L. Kozak*, Charles Kunsch*, Hungjun Ji*, Haodong Li*, Paul S. Meissner*, Henrik Olsen*, Lisa Raymond*, Ying-Fei Wei*, John Wing*, Charlotte Xu*, Guo-Liang Yu*, Steven M. Ruben*, Patrick J. Dillon*, Michael R. Fannon*, Craig A. Rosen*, William A. Haseltine*, Chris Fields†, Claire M. Fraser & J. Craig Venter‡

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA
* Human Genome Sciences, Inc., 9410 Key West Avenue, Rockville, Maryland 20850, USA

In an effort to identify new genes and analyse their expression patterns, 174,472 partial complementary DNA sequences (expressed sequence tags (ESTs)), totalling more than 52 million nucleotides of human DNA sequence, have been generated from 300 cDNA libraries constructed from 37 distinct organs and tissues. These ESTs have been combined with an additional 118,406 ESTs from the database dbEST, for a total of 83 million nucleotides, and treated as a shotgun sequence assembly project. The assembly process yielded 29,599 distinct tentative human consensus (THC) sequences and 58,384 non-overlapping ESTs. Of these 87,983 distinct sequences, 10,214 further characterize previously known genes based on statistically significant similarity to sequences in the available databases; the remainder identify previously unknown genes. Thirty tissues were sampled by over 1,000 ESTs each; only eight genes were matched by ESTs from all 30 tissues, and 227 genes were represented in 20 or more of the tissues sampled with more than 1,000 ESTs. Approximately 40% of identified human genes appear to be associated with basic energy metabolism, cell structure, homeostasis and cell division, 22% with RNA and protein synthesis and processing, and 12% with cell signalling and communication.

A MAJOR objective of the human genome project is identification of the complete set of human genes. Single-pass, partial sequencing of cDNA clones from one or both ends to generate expressed sequence tags (ESTs)[1] provides a rapid method of gene discovery that has been widely applied in humans[1–14] and other species[15–24,67]. The EST strategy was developed to allow rapid identification of expressed genes by sequence analysis, while providing a key resource for gene mapping[1,25–27]. Before 1991 and the development of the EST method, sequence data existed for fewer than 3,000 human genes (GenBank release 68, June 1991)[28]; as of January 1995, annotated coding-sequence data were available for approximately 5,100 human genes (GenBank release 86), and up to 25,000 additional genes were represented by EST sequences (dbEST release 3.0)[29]. The combination of data on gene expression and putative gene functions inferred from sequence similarity provides a powerful means of assessing the transcriptional activity of the genome in the cells and tissues of an organism. The results presented in this paper summarize our preliminary characterization of 87,983 unique complementary DNA sequences expressed in thirty-seven human tissues at various stages of development.

To obtain an initial overview of gene diversity and expression patterns in human tissues and organs, we sequenced over 200,000 DNA templates selected from 248 primary and 52 screened or subtracted cDNA libraries. The use of screened or subtracted libraries was limited because of the value of obtaining a representation of the transcription level in primary libraries to as faithful an extent as possible. All but three libraries were constructed specifically for this study. These cDNA libraries represented all

† Present address: National Center for Genome Resources, Santa Fe, New Mexico 87505, USA
‡ To whom correspondence should be addressed.

major organs, several developmental stages and disease states, and many individual cell types (either primary or immortalized). We have defined 174,472 new, high quality ESTs that represent a broad array of genes, including many with transcripts that appear to have a limited range of expression.

We have taken advantage of the inherent redundancy of cDNA sampling to build assemblies of ESTs, essentially treating the expressed portion of the genome as a shotgun sequence assembly project. These assemblies (termed tentative human consensus sequences (THCs)) carry information on the source library and abundance of ESTs, and in many cases represent full-length transcripts. All human EST data from GenBank have been incorporated into THCs. Altogether, over 83 megabases (Mb) of cDNA sequence data have been analysed. This 'Human Gene Anatomy' project provides a large-scale view of differential gene expression that encompasses human anatomy and development.

Overall quality control

Large-scale sequencing demands attention to the quality of materials and to accurate performance of each step in the process, both to provide sequence data of the highest possible quality and to detect or avoid problems[30]. At each step of the EST methodology, both in the laboratory and during sequence analysis, a quality control and evaluation procedure was developed to assess the EST data quality (Table 1). The objective was to assure at each step that the material produced was of sufficient quality (for example cDNA libraries) or purity (for example, templates) to provide a high level of confidence that a high proportion of the sequencing reactions would produce useable data. Checks of the accuracy of ESTs and EST assemblies (THCs are described below) served to define confidence values for interpreting database matches.

TABLE 1 Quality control and evaluation

Procedure	Quality control and evaluation
Tissue procurement	Tissues snap frozen as quickly as possible *post mortem*; tissue samples that tested positive for HIV and/or hepatitis were not used
mRNA purification	500 μg total RNA to start; mRNA concentration determined by spot blot
cDNA synthesis	Tracer levels of ^{32}P included; agarose gel examination for degradation; size selection ⩾500 bp
cDNA library construction	Blue/white screen for inserts before and after *in vivo* excision; PCR to check insert size: must be 1.0–1.5 kb average; libraries must contain $\geqslant 0.5 \times 10^6$ recombinants
Sample sequencing	Check gene diversity and content, mitochondrial contamination, insert size, % full-length, Alu content, directionality
Library screening	Screening with total cDNA, mitochondrial genome, or specific abundant cDNA clones; rechecked by sample sequencing to confirm reduction of abundant cDNAs (also see Fig. 1)
Template preparation	Concentration checked by CytoFluor or agarose gels; some sets checked by sample sequencing; clone location tracking in ESTDB; success rate tracked in ESTDB
DNA sequencing	pGEM control plasmids on ~1/4 of sequencer runs; protocol/reagent tracking in ESTDB
EST sequence quality check	⩽3% Ns, ⩾100 bp; HBQCM for non-human contamination
Database searching	BYOB evaluation, HT dataset screening, nomenclature consistency checking; analysis of accuracy with respect to known sequences
Sequence assembly	Chimaera-detection algorithm, search databases with BYOB evaluation, manual evaluation of selected assemblies

Quality control procedures for each step in the EST process are listed with specific points of evaluation or standards to be met. The EST process is described in detail in the text.

cDNA libraries

Sources of tissue. Human tissue samples were obtained from the National Disease Research Interchange (Philadelphia, Pennsylvania) and The International Institute for the Advancement of Medicine (Philadelphia, Pennsylvania). When feasible, donors were adult male or female patients, aged 20–50 years, and free of any known major diseases and medications. Tissues were dissected using sterile techniques, snap frozen in dry ice/acetone less than 4 hours *post mortem*, and kept frozen at −80 °C until processed for library construction. Whole blood and blood components were obtained from the American Red Cross.

RNA preparation. Total RNA was prepared from whole frozen tissues or cultured cells using acid-phenol-thiocyanate[31]. The frozen tissues were pulverized in liquid nitrogen before the addition of the guanidinium solution, and the slurry was then immediately homogenized with a Polytron. The selection of poly(A)$^+$ RNA from total RNA isolation was performed using either oligotex-dT (Qiagen, California), which is oligo-dT covalently linked to latex particles[32] or oligo(dT)-coated magnetic beads (Dynabeads, Dynal, New York)[33]. When the quantities of total RNA were sufficient, two rounds of purification were performed.

cDNA library construction. For this type of study, a cDNA library should be: (1) representative, containing all sequences present in the initial poly(A)$^+$ RNA population in the same relative frequencies; (2) unidirectionally cloned so that the orientation of each cDNA is known, facilitating subsequent sequence analysis; (3) composed of a high proportion of long or full-length inserts; (4) uncontaminated with genomic, mitochondrial or ribosomal RNA inserts; and (5) composed of a large proportion of inserts with short poly (A) tails.

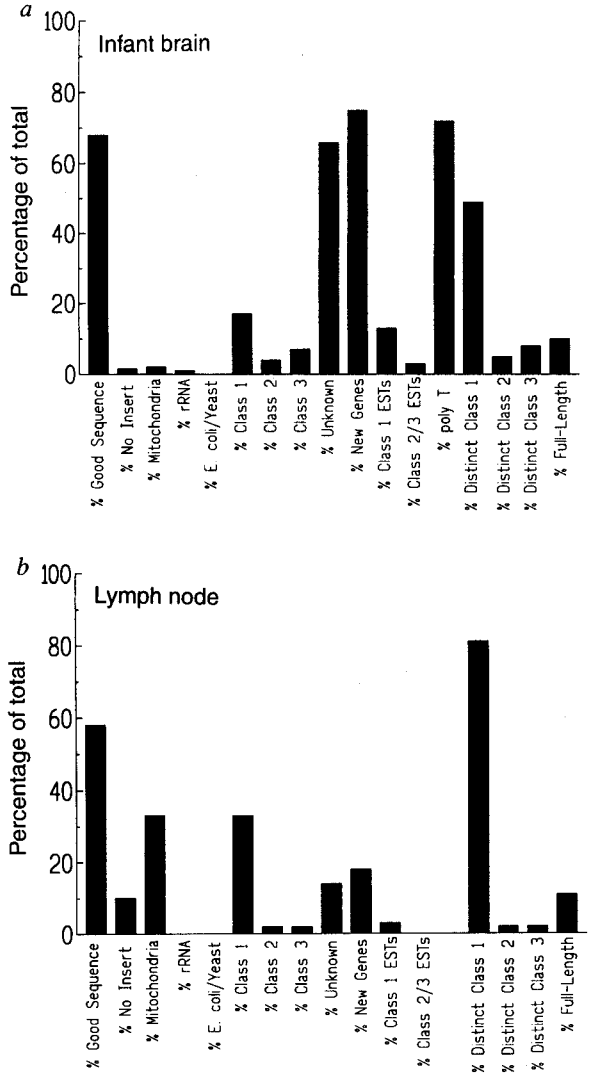
cDNA synthesis was performed and unidirectional cDNA libraries were constructed using the Lambda ZAP II vector[34] (Stratagene, La Jolla, California). This cDNA synthesis method is based primarily on the method of Gubler and Hoffman[35] with a few modifications. The first strand reaction was primed by an oligo-dT primer containing a *XhoI* site. To protect the cDNA from digestion by restriction enzymes, 5′-methyl dCTP was used in the first strand synthesis, thereby producing hemimethylated DNA. The efficiency of cDNA synthesis reactions was monitored with radioactive nucleotides and a portion of each reaction product was separated electrophoretically to determine the size range of the synthesized cDNA. We used Sephacryl-S400 columns (Life Technologies, Gaithersburg, Maryland) to select by size the cDNA and remove excess adapters. cDNA greater than 500 base pairs (bp) in size was pooled and ligated to the Lambda Zap II vector arms and subsequently packaged with Gigapack II Gold packaging extract (Stratagene). This vector incorporates the pBluescript SK- plasmid, which can be excised *in vivo* from the lambda vector as a phagemid carrying a cDNA insert[36].

After construction, cDNA libraries were evaluated for quality by determining the ratio of recombinants to non-recombinants and the average size of the cDNA inserts using polymerase chain reaction (PCR) analysis of approximately 50 individual clones. cDNA libraries that contained an average insert size of greater than 1000 bp were excised *en masse*; at least 10^6 primary packaged phage from the unamplified cDNA library were rescued as phagemid particles by coinfecting either *Escherichia coli* SURE or XL1-BLUE MRF′ cells (Stratagene) with the ExAssist helper phage (Stratagene). The excised pBluescript phagemids were used to infect SOLR cells (Stratagene), which lack the amber suppressor necessary for ExAssist phage replication. The infected SOLR cells were selected for ampicillin resistance on Luria-Bertani (LB) plates supplemented with X-gal and isopropyl-β-D-thiogalactoside (IPTG) and used for production of

double-stranded templates for automated sequencing of cDNAs.

For this study a total of 248 cDNA libraries were constructed; an additional 3 were purchased or obtained from other sources (see Table 2 following reference list). Libraries were grouped into general categories for comparison of tissue expression patterns. Libraries prepared from normal adult and/or fetal tissue samples, diseased tissues when available, and primary or immortalized cell lines derived from the respective tissues were grouped together. Note that tissue samples in most cases include multiple cell types, including vascular and immune-system cells.

Evaluation of cDNA libraries. Sequencing a small number (100–200) of clones proved to be an excellent way of assessing library quality in terms of gene content and determining problems that may have arisen during library construction. As summarized in Fig. 1, 17 different parameters were examined to evaluate library quality. Libraries selected for large-scale EST analysis based on quality control evaluation typically exhibited less than 50% exact matches to known human genes, a broad diversity of transcripts (no single gene or small group of genes dominating the distribution), a low percentage of clones with no inserts, mitochondrial transcripts and/or ribosomal RNA species, and no evidence of contamination with sequences from another organism. Certain libraries that did not meet these general criteria were either remade, screened before sequencing (52 libraries, described below), or sampled on a limited basis because they were derived from materials difficult to replace.



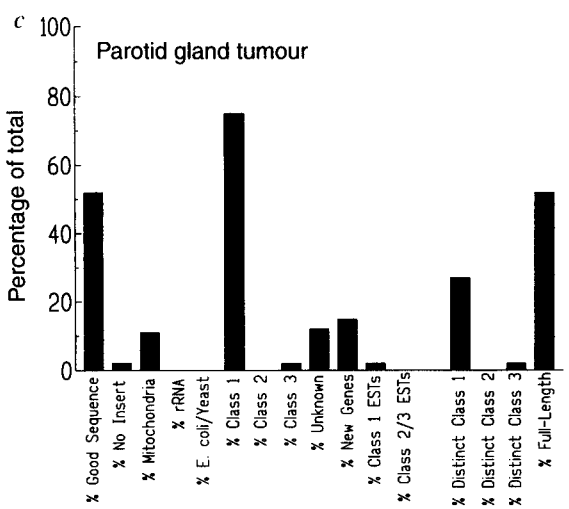


FIG. 1 cDNA library quality control parameters. Prior to large scale EST analysis, 100–200 clones from each cDNA library were sequenced to obtain quality control information. The following parameters were evaluated. % good sequence was included as a measure of the overall success rate for the quality control sequencing reactions. If the % good sequence obtained was not sufficient to obtain statistically significant data, the quality control reactions were repeated. % no insert; % mitochondria (% of ESTs with exact matches to sequences encoded by mitochondrial genes); % rRNA (% of ESTS with exact matches to sequences derived from ribosomal RNA species), % *E. coli*/yeast (% of ESTs with exact matches to sequences in public databases derived from *E. coli*, other bacteria, yeast, and bacteriophage lambda), % Class 1 (% of ESTS with a ≥95% nucleotide match to human sequences in the GenBank and dbEST); % Class 2 (% of ESTS with a <95% nucleotide match to human sequences in the public databases), % Class 3 (% of ESTS with a best match to a non-human sequence in the public databases); % unknown (% of ESTs with no significant matches to anysequences in the public databases); % new genes (the sum of all Class 2, Class 3 and unknown ESTs divided by the total number of ESTs); % Class 1 ESTs (% of ESTs with a ≥95% match to another human EST sequence in the public databases); % Class 2 or Class 3 ESTs (% of ESTs with a <95% match to another human EST sequence or a best match to a non-human EST sequence in the public databases); % poly(T) (% of ESTs with a stretch of poly(T) sequence at the beginning, indicative of a poly(A) tail). Because of the directional cloning strategy used in cDNA library construction, the M13 forward primer site in pBluescript allows for sequencing from the 3′ end of a cDNA insert, which theoretically contains a polyadenylation signal. Thus % poly(T) gives information on the success of directional cloning for each library. % distinct Class 1 (a measure of the diversity of the Class 1 sequences in each library). If a cDNA library contains one or more abundant Class 1 transcripts, the % of distinct Class 1 hits will be low. % distinct Class 2 or Class 3 (a measure of the diversity of the Class 2 or Class 3 sequences in each library as described). % full length (an estimate of the number of clones in each cDNA library that contain full-length coding region sequence). This number is calculated from 5′ EST sequencing where database matches indicate that the complete coding sequence of the cDNA is contained in the insert. To justify large-scale EST analysis, each cDNA library had to contain less than a total of 20% of no insert clones, mitochondrial and rRNA sequences, no contaminating sequences from another organism, at least 50% new genes and at least 80% distinct Class 1, 2 and 3 genes. *a*, Infant brain cDNA library. Based on quality control analysis, this library was deemed to be of sufficient quality for large scale EST analysis. In particular, the sum of % no insert, mitochondrial, rRNA and contaminating sequences was extremely low and the % of new genes was very high. *b*, Lymph node cDNA library. Based on quality control analysis, this library was not approved for large-scale EST analysis because the sum of the % no insert and % mitochondrial sequences was greater than 40% and the % of new genes was less than 50%. *c*, Parotid gland tumour. Based on quality control analysis, this library was not approved for large-scale EST analysis because the % of new genes was less than 50%, and the % of distinct Class 1 genes was less than 30%. Filter screening of cDNA libraries described in *B* and *C* was performed with the abundant transcripts and libraries were re-evaluated for content.

Several cDNA libraries contained one or more extremely abundant species (more than 5% of the sequenced clones). In these cases, the individual abundant cDNAs or total cDNA was labelled and used as a probe to screen gridded arrays of clones from the library. Non-hybridizing clones were chosen for sequencing. This procedure was applied to 52 libraries (indicated in Table 2); 15,521 ESTs in the dataset are from screened libraries. These data were not used to estimate quantitative differences between libraries; in other respects, these ESTs were treated in the same way as ESTs from non-screened libraries.

Template preparation and DNA sequencing

Over the course of the project, several different DNA template preparation methods were used, all producing double-stranded templates (Table 3). Most of the templates were made by the standard boiling method[37], 96-well boiling mini prep method (AGTC, Gaithersburg, Maryland), or PCR; the latter two methods are still in use at our laboratories: The Institute for Genomic Research (TIGR) uses AGTC prep, and Human Genome Sciences (HGS) uses AGTC and PCR preps. Both PCR and AGTC preps are performed in a 96-well format for all stages of template preparation, from bacterial growth through to final purification. The 96-well format is faster and more reliable than individual clone formats because fewer samples are handled manually. Template concentrations were not adjusted, although low-yielding templates were identified and not sequenced where possible.

Sequencing reactions were performed on either plasmid or PCR product templates using the Applied Biosystems (AB) Catalyst Lab station or Perkin-Elmer 9600 Thermocyclers with Applied Biosystems PRISM Ready Reaction Dye Primer Cycle Sequencing Kits for the M13 forward (-21M13) and the M13 reverse (RP1) primers. Catalyst software version 2.0.1 or an experimental software version 1.35 was used. Reaction products were precipitated with 95% ethanol using either sodium acetate (3 M) or glycogen as carrier and washed once with 70% ethanol before drying under vacuum. The dried reactions were stored at −20 °C in the dark.

The sequencing reaction products were analysed using AB 373 DNA sequencers and version 1.2 data collection and analysis software, as previously described[30]. Data were obtained using 24-cm well-to-read plates and 32 lanes with 6% acrylamide gels and 30 W constant power. The gels contained 8 M urea and 1X Tris Borate EDTA buffer, pH 8.3 (TBE), and 1X TBE buffer was used as the running buffer. Sample loading buffer was formamide-EDTA (5:1 v/v). Because the primary goal of this project was gene discovery, most of the sequencing was done from the 5′ end of inserts. The 5′ end of each clone is more likely to contain protein coding sequence than the 3′ end, which increases the likelihood that database searches will result in the assignment of putative identifications. Two to three portions of each sequenced cDNA clone were stored in separate locations.

Sequence quality analysis. Standards for sequence accuracy were implemented at two levels. First, all AB 373 Sequencer output obtained at TIGR was scanned visually to confirm overall quality of peak shape and correspondence with base calls. Approximately 700 bases are called by the AB 373A data analysis software, including several hundred at the end of the run that are beyond the limit of resolution of the sequencing gel and detection system. The end of a run frequently contains a high percentage of ambiguous base calls because the peaks broaden and overlap. Clean templates exhibit low background signals, hence sequence quality may be poor owing to peak broadening without a concomitant increase in ambiguous base calls. Therefore the 3′ end of each sequence was trimmed manually to provide consistent editing criteria that included visual analysis of peak shape as well as the number of ambiguous base calls. Second, software filters trimmed sequences with more than 3% ambiguous base calls.

A series of software programs was developed to transfer sequence and associated information into the expressed sequence tag database, ESTDB[30,38,39] (Fig. 2). Short descriptions of each program are given in Table 4. The software program MAP was used to load sample and gel data, and the program ESP was used to load edited sequences into ESTDB and to assign 'trash codes' to sequences that did not meet the minimum standards of accuracy. Trash codes were divided into 13 categories representing sequence failures caused by instrumentation, technician error, reaction or template quality (such as low concentration or mixed templates), and library quality (such as no-insert or long poly(A) tails). Trash codes allowed an ongoing and thorough evaluation of template quality and instrument performance, and rapid assessment of the impact of changes in protocols or reagents. Of a total of 165,787 sequencing reactions performed at TIGR, 64,559 sequences were rejected because they did not meet our minimum standards of sequence quality. Precise sequence accuracy was assessed by comparing ESTs to known sequences from GenBank. The average sequencing accuracy for ESTs generated for this project was >98%, consistent with that obtained in our previous EST projects[3].

Leading and trailing vector and polylinker sequence were removed by VERM, which identifies vector, polylinker, adaptor and poly(A/T) sequences by similarity searching. Coordinates of the removed sequence were recorded in ESTDB along with the edited sequence. VERM was designed to handle any combination of vector and cloning site, including both directional and non-directional libraries, by reading essential library-specific parameters from ESTDB, and vector sequence from an external file.

Given that organisms such as bacteria may be present in human tissue samples and in the laboratory environment, it is possible that low numbers of non-human sequences may be present as contaminants in some human cDNA libraries. Some contaminants are not found by sequence similarity searches because they do not have a corresponding protein or DNA sequence in the public databases. To prevent sequencing from a large number of contaminant cDNAs in this project, we developed an algorithm that detects gross differences in the composition of DNA from different organisms. The test takes advantage of the statistically distinct hexamer content of human DNA. Previously, hexamer analysis has been used to identify human cDNA libraries that were contaminated with prokaryotic and yeast sequences[40]. In the current study, the hexamer test showed that two libraries (fewer than 200 ESTs each) had large portions of non-human sequences. The same two libraries were also identified by BLAST[41] as having exact matches to either *Mycoplasma* or *E. coli* DNA; all sequences from these libraries were removed from the database. Hexamer analysis also identified non-random distributions of human cDNAs in libraries that were not contaminated with foreign DNA. Typically, less than 150 cDNAs were

TABLE 3 Template method summary

Preparation method	No. sequenced	No. successful	% successful
PCR	51,483	29,964	58.2
AGTC	47,913	32,486	67.8
Rapid Boil	42,650	27,680	64.9
Qiagen	3,939	2,459	62.4
Promega	1,531	872	56.9
Autogen	901	644	71.5
Other	1,792	940	52.4

Template preparation methods are listed with the number of sequencing reactions performed, the number of successful reactions and the overall success rate. PCR, AGTC (ref. 68), Rapid Boil[37], Qiagen[64], Promega Wizard Mini Preps (Promega), and an Autogen 540 DNA preparation instrument were used as described. Percentage successful was measured by the number of sequencing reactions giving at least 100 bp with fewer than 3% ambiguous base calls.

sequenced from these non-random libraries because preliminary screening suggested the occurrence of anomalous RNA isolation or cDNA construction.

Assembly of consensus sequences of ESTs

Messenger RNA species are present at different concentrations in cells, and these differences are reflected in the composition of cDNA libraries. Thus random-sampling strategies result in abundant mRNAs being represented by many ESTs. To quantify the extent of redundancy and to build longer, contiguous blocks of sequence, we treated the ESTs as shotgun fragments. cDNA contiguous, overlapping sets of DNA clones (contigs) were assembled based on stringent overlap criteria[42]. These contigs have been termed tentative human consensus sequences (THCs).

THCs present a data management challenge analogous to that of simultaneously assembling shotgun fragments from multiple independent clones, as they constantly grow (to the point of representing a full-length transcript)[42], requiring revision at each stage, and are subject to such biological variability as alternative splicing. An update procedure that uses the existing THCs as seeds for the assembly of new contigs has been developed to simplify incorporation of new ESTs.

The process of assembling THCs served to identify ESTs from the same gene to reduce redundancy, and aided in the assignment of putative identification, because the consensus of several sequences was generally longer than any of its constituents. Furthermore, assembly of ESTs into consensus sequences also facilitated the identification of distinct transcripts that differ only in splicing pattern. Despite efforts to minimize false assemblies, THCs are artificial constructs derived from independent cDNA clones, and therefore have the potential to be chimaeric. A reasonable degree of confidence in the assembled consensus sequence is obtained with sufficient redundancy.

A program, TIGR ASSEMBLER, was written to assemble large sets of sequence data; it has been successfully used in the assembly of the complete *Haemophilus influenzae* genome[43]. TIGR ASSEMBLER was used to assemble 185,420 non-mitochondrial ESTs from TIGR, HGS and dbEST v2.6. A further 4,554 sequences from a non-redundant set of human transcripts (HT sequences, see below) were included in the assemblies to analyse the expression profile of these genes. TIGR ASSEMBLER rapidly assesses sequence-level similarity by tabulating 10-bp oligonucleotide subsequences for each EST, and placing ESTs of similar oligonucleotide content onto a search list of possibly overlapping sequences. A seed sequence initiates an assembly that is extended by attempting to add the best matching fragment from the search list to the consensus sequence (this is termed a 'greedy algorithm'). An EST sequence was added to the consensus sequence only if it met stringent overlap criteria, as determined by performing a modified version of the Smith-Waterman algorithm for aligning the EST with the consensus sequence. The criteria included a minimum of 95% identical nucleotides in an overlap region of at least 40 bp with a maximum unmatched overhang of 20-bp. Each Smith–Waterman alignment that met the criteria was added to the full multiple alignment, which was then used to define the THC consensus sequence. Evaluation of TIGR ASSEMBLER was performed at this stage by spot checking the multiple alignments. Another program, LASSIE, loaded the consensus sequence and relative coordinates of contributing ESTs into ESTDB. TIGR ASSEMBLER produces separate THCs for ESTs that are the result of separate mRNA splice forms. Potentially chimaeric fragments and alternatively spliced or unspliced cDNAs are flagged based on partial mismatches at the ends of alignments and excluded from assemblies. As a result, a certain redundancy remains within the THC dataset, with alternative splice forms represented as distinct THCs. In addition, low-quality sequences that cannot be assembled with high confidence into a THC remain as single ESTs (singletons).

Several THCs matched two unrelated genes in the public databases and hence were likely to be chimaeric. In most cases, a single EST was identified as chimaeric and removed from the database; the remaining sequences were re-assembled to produce two non-chimaeric THCs. In other cases, two overlapping genes appear to be represented in a single contig[44].

THCs may also assemble together ESTs from distinct genes, where the level of sequence identity is high. Although some gene

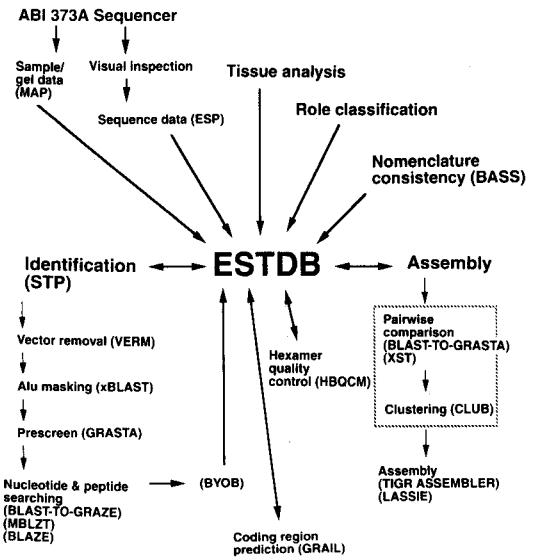


FIG. 2 Data flow for EST analysis. ESTDB is the Expressed Sequence Tag Database[38,39]. Software routinely used in the analysis is shown in brackets. Additional description of the EST data analysis procedures can be found in the text. Software developed at TIGR specifically for this project is listed in Table 4.

TABLE 4 Software tools developed at TIGR

Program	Platform/Language	Task
BASS	Unix/C	Automatic annotation based upon sequence identity
BLAST-TO-GRAZE	Unix/shell	Sequential BLAST prefilter and Smith–Waterman alignment
BTAB	Unix/C/lex/yacc	Parse search output into standard format[66]
BYOB	Unix/C++/Motif	Graphical sequence similarity search output browser
CLUB	Unix/C	Cluster building of matching ESTs based on XST analysis
ESP	Macintosh/MacApp	Sequence data entry into ESTDB
ESTDB	Sybase	Relational database structure for tracking EST analysis
GRASTA	Unix/C	Double stranded FASTA[50] search with BTAB output
GRAZE	Unix/C	Modified Smith–Waterman alignment tool with BTAB output
HBQCM	Unix/C	Hexamer-based quality control analysis for cDNA libraries[40]
LASSIE	Unix/C	Loading assemblies into ESTDB
MAP	Macintosh/MacApp	Sample and gel data entry into ESTDB
MBLZT	Unix/shell	6-frame translation search with BLAZE
NCOUNTER	Unix/C	Trim ESTs based on ambiguities
SEQV	Unix/C/C++/Motif	Graphical sequence manipulation
SGRASTA	Unix/C	Strict GRASTA alignment for vector identification at ends of ESTs
STP	Unix/C/Motif	Sequence extraction from ESTDB and process control
TIGR ASSEMBLER	Unix/C	Automated sequence assembly[69]
VERM	Unix/C	Vector and polynucleotide removal
XST	Unix/C	Pairwise GRASTA comparison and logging of EST overlap data

Description of the analysis programs shown in Fig. 2.

```
T57234
  74.3% identity in 148 nt overlap

pBlue   AATTCGATATCAAGCTTATCGATACCGTCGACCTCGAGGGGGGGCCCGGTACCCAATTCG
                                        ::  :::::::::::::.::: :::::::
T57234  ACTTTTGAACTTTTAAAAAAAAAAAAAAAAAACTCCGAGGGGGGGCCNGGT-CCCAATTTT
         100       110       120       130       140       150

pBlue   CCCTATAGTGAGTCGTA-TTACAATTCAC-TGGCCGTCGTTTTACAACGTCGTGACT---
         :::::::: ::::::: ::.::::::.:: : ::::::::::::   ::::::: :::
T57234  GCCTATAGTTAGTCGTATTTNCAATTNACTTTGCCGTCGTTTTTTCACGTCGTTACTTGG
          160       170       180       190       200       210

pBlue   GGGAAAACCCTGGCGTTACCCA-ACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGC
        :::::: :   :: :::.::::   ::: :.:: :: ::: :::: ::: ::: : ::
T57234  GGGAAACCTGGGGGGTTNCCCACTTTTATTNGCTTTTCAGGACATTCCCTTTTTGGCA
          220       230       240       250       260       270
```

FIG. 3 Detection of vector in an EST sequence with GRASTA. An alignment of EST T57234 with the pBluescript vector sequence from GenBank. EST T57234 apparently reads through the poly(A) tail at the end of a short cDNA insert and reads back into the pBluescript SK- vector sequence.

families are known to include two or more members that are over 99% identical over their entire lengths (see refs 45, 46), the members of most characterized gene families have less than 95% nucleotide identity (see refs 47, 48) and hence can readily be distinguished at the stringency used in our assemblies.

Incorporation of data from dbEST. After analysis of the initial dataset, it became clear that the quality control protocols, and EST assembly, identification and annotation tools we had developed for this project could be readily applied to all available human EST data. We have therefore incorporated all available data from dbEST (through to 28 April 1995) into the analysis presented here. In addition to reducing the redundancy of the world-wide EST dataset, this provided additional depth to sequence assemblies as well as additional transcript expression data. We perform this data analysis on an ongoing basis to make our dataset the most complete representation of EST data.

A semiautomated procedure was developed for incorporation of sequences from dbEST. Scripts automatically download daily updates from dbEST to our computers using FTP. A program parses fields from the dbEST flatfile (dbEST accession number, GenBank accession number, Genome Data Base accession number, library, primer, source laboratory, sequence) which are subsequently loaded into ESTDB. New sequences are automatically flagged for analysis. A script checks nightly for new sequences to be analysed and subjects them to the quality control routine. All sequences shorter than 100 bases after the trimming steps outlined below are flagged in the database and not used for further analysis. Sequences are next checked for the presence of ambiguities (Ns) using NCOUNTER, which trims sequences from the 3′ end until the total number of Ns does not exceed 3%. The sequences are then checked for vector, poly(A), poly(T) or other sequence artefacts such as poly(GA) or poly(CT). Because the sequences come from a variety of libraries and the vector used may not be annotated, we performed an exhaustive search against a database of vector sequences extracted from GenBank. This search was performed in three stages; the first vector search was done using BLAST-TO-GRASTA because it has greater sensitivity than BLAST alone. Second, the most common vectors were screened with SGRASTA, a more stringent version of GRASTA which looks for short matches to vector, specifically at the ends of ESTs. The match stringency of SGRASTA was increased to find only nearly exact matches (>95% nucleotide identity). A final search using GRASTA alone was evaluated manually to identify vectors that may have escaped detection in the first two steps. An example of the value of using GRASTA to identify vectors remaining on ESTs in GenBank is shown in Fig. 3. This method discovered 4,807 vector fragments remaining on sequences in dbEST. Finally, the dataset was checked for the presence of mitochondrial transcripts or ribosomal RNA. The results of this clean-up process for several large EST datasets in dbEST are shown in Table 5.

Sequences were next subjected to a clustering process. New sequences were checked for overlap with the existing set of THCs and singleton ESTs using BLAST-TO-GRASTA. Pairwise comparision results were loaded to ESTDB using XST. Those sequences completely contained within an existing THC were linked to assembly records in ESTDB for that THC. Those new sequences that were not completely contained in THCs were searched against the HT dataset for the purpose of assigning putative identifications. Finally, the set of remaining new sequences was analysed for overlaps within itself and with the existing THCs and ESTs. The results of all the clustering steps were loaded into ESTDB, where CLUB linked sets of overlapping ESTs and THCs into clusters. ESTs and THCs representing each cluster were assembled independently with TIGR ASSEMBLER, and the resulting assemblies were loaded into ESTDB with LASSIE.

A set of 130,238 sequences (41,066,788 bases) from dbEST were subjected to this analysis (Table 5). Of these, 11,832 (2,094,263 bases) were eliminated during the quality control steps (3,041 for length, 5,738 for percentage Ns, 3,052 for vector/poly(A)/mitochondrial, or other contamination). Of the 118,406 remaining sequences (36,972,525 bases), 54,198 (17,700,416 bases) clustered with existing THCs and single ESTs, and 64,567

```
1==================================THC69097==================================3266
-----------------------------------1------------------------------------->
-----2---->                                ---4--->          ---12-->
-----3----->                                 -----5---->       ---13-->
                                            -6-> ---11-->        ---14--->
                                            |--a->.7..|            -----15---->
                                            |----b--->|             ----16---->
                                             --8->
                                              ----9--->
                                               <-10---
```

No.	EST	GB	ATCC	left	right	library
1	E HT1002			1	3183	
2	B EST143092			21	464	Fetal heart II
3	B EST143393			21	467	Fetal heart II
4	B EST124352			1781	2100	Cerebellum II
5	D	R12095		1862	2284	Infant brain 1NIB
6	A EST31673			1870	2003	Embryo, 12 week I
7a	C EST07766		84742	1905	2078	Infant brain
7b	C EST06074	T08183	84742	1905	2263	Infant brain
8	D	F00027		1917	2089	Skeletal muscle
9	A EST56833			1978	2304	Infant brain
10	D	Z19202		2018	2258	Skeletal muscle
11	A EST70041			2094	2381	T-cell lymphoma
12	A EST77257			2504	2794	Pancreas tumor III
13	A EST55039	T28754	103497	2588	2884	Hippocampus II
14	D	R24627		2730	3038	Placenta
15	D	T90430		2807	3243	Stratagene lung (#937210)
16	D	T83396		2829	3266	Fetal liver+spleen 1NFLS

Sequence source codes: A = TIGR, B = HGS, C = NIH, D = GENBANK, E = EGAD

FIG. 4 THC report. An example THC is shown with the ESTs and HT sequence it contains drawn to scale. The coordinates of the overlap and source libraries are listed. THC reports such as this are returned by HCD software (see data availability).

TABLE 5 Quality control analysis of worldwide EST projects

	TIGR	HGS	dbEST	A	B	C	D	E	F	G	H	I	J
Total ESTs	92,830	55,476	130,238	77,922	25,461	7,316	3,367	3,321	1,946	1,875	1,554	1,153	1,120
Average length	292	313	327	356	285	321	338	223	243	213	212	311	306
% length <100	0	0	2.3	2.4	0	0	0	20.8	0	7.8	4.8	0	0.4
ESTs <100 nt	0	0	3,041	1,904	0	0	0	690	0	146	74	0	5
% lengths >200	90.6	88.9	87.7	90.1	92.2	94.4	94.1	47.4	64.7	56.0	53.7	94.9	91.1
ESTs >200 nt	84,140	49,304	114,218	70,194	23,481	6,910	3,169	1,573	1,259	1,049	834	1,094	1,020
Average %N	1.03	1.77	1.05	1.2	0.75	1.16	0.88	2.03	0.03	0.77	1.72	0.05	0.06
% w/%N >3	0	0	6.5	6.9	3.8	4.7	3.0	22.9	0.1	6.9	18.1	0.3	0.4
ESTs w/%N >3	0	0	8,438	5,405	972	344	100	761	2	130	282	3	5
ESTs w/vector	0	0	2,495	1,310	550	0	152	76	62	30	33	49	44
Mitochondrial RNA	0	0	124	32	1	4	21	8	20	0	2	18	13
Ribosomal RNA	0	0	76	8	1	5	17	1	2	0	1	11	7
Alu	6,914	857	6,148	4,358	761	543	161	15	69	0	20	30	62
% Alu	7.4	1.5	4.7	5.6	3.0	7.4	4.8	0.4	3.6	0	1.3	2.6	5.5
% unused			9.1	7.4	11.8	4.9	4.2	33.3	4.9	14.7	22.5	2.9	3.7
ESTs unused	0	0	11,832	5,752	3,004	359	142	1,105	96	275	349	33	41
ESTs in study	92,830	55,476	118,406	72,170	22,457	6,957	3,225	2,216	1,850	1,600	1,205	1,120	1,079

A comparison of several assessments of sequence quality among large EST datasets is shown. ESTs were extracted from dbEST and analysed by the procedures described in the text to examine the length and percent ambiguous (N) nucleotides (nt). All contributors with over 1000 ESTs in dbEST are listed. For TIGR and HGS, sequence trimming and vector removal were performed as part of the initial data acquisition. Sources of EST datasets: TIGR and HGS, this study; dbEST, all ESTs from dbEST as of 28 April 1995; A, Washington University-Merck EST project; B, Auffray, *et al.*[12]; C, Adams, *et al.*[1–4]; D, Khan, *et al.*[8]; E. Okubo, *et al.*[9,13]; F, Liew, *et al.*[7,14]; G, UK Human Genome Mapping Project Resource Centre Clinical Research Centre; H, Takeda *et al.*[10]; I, Soares, *et al.*[11]; J, Genzentrum Muenchen Laboratorium für molekulare Biologie.

(21,272,109 bases) only clustered within the set or remained as single ESTs.

After assembly there were 58,384 singletons and 29,599 THCs containing at least two ESTs. Clusters ranged in size from 2 to 2,780 ESTs (the largest was elongation factor-1α). The average length of non-HT THCs was 492 bases, compared with 288 bases for individual ESTs. An example of a THC is shown in Fig. 4.

Identification of EST and THC sequences

A standard protocol was developed for analysing EST and THC sequences and storing the results of each analysis in ESTDB (Fig. 2, Table 4). The analysis procedure was divided into two conceptually different parts: (1) automated processing of analysis results that could be unambiguously interpreted without human intervention; and (2) sequence comparison searches, the results of which were viewed and evaluated by scientists.

Construction of non-redundant human transcript set. Because GenBank is a repository for nucleotide sequence data, multiple entries have been deposited for many genes. This redundancy, and the inclusion of intron and non-transcribed sequence data, complicates the identification of cDNA (EST) sequences by sequence similarity. Therefore, we created a non-redundant human transcript (HT) database from the sequence information contained in GenBank for the purpose of accurately identifying ESTs from known human genes.

The non-redundant HT database contains nucleotide sequences that represent mature mRNAs. Sequences were either loaded directly from GenBank (cDNAs) or were derived from GenBank entries (genomic sequences). Where available, 5′ and 3′ non-coding regions were included. All sequences known to represent alternative splice forms of a gene, either through explicit GenBank annotation or through multiple sequence alignments of GenBank accessions, were loaded into the HT database and were linked internally. When possible, splice sites were verified to occur at known intron/exon boundaries or to have the standard splice-site nucleotides by comparing genomic and cDNA sequences.

GenBank records contain annotations that identify locations of numerous features within the nucleotide sequence, such as coding regions of cDNAs and intron/exon boundaries of genomic sequences. However, owing to the variety of sequence types within GenBank, not all features are used in all GenBank entries. This complicates consistent, automated retrieval of nucleotide data from GenBank. Even so, for construction of the HT database, many accessions were parsed and loaded automatically. Separate parsing programs were required for automated parsing and loading of genomic sequences, cDNAs and accessions that lacked certain annotations. Sequences that could not be parsed and loaded automatically as a result of incomplete or incorrect annotation were loaded manually, using the interactive graphical/manipulation program SEQV. This imports sequences from GenBank, checks for non-biological splice sites, truncated coding sequences, and mis-annotated translational starts or stops, and displays the readable features graphically. Features that are clearly misannotated can be corrected within SEQV and verified by checking the translation for premature stop codons. Out of 4,554 total sequences from 4,100 genes, 3,114 (68%) sequences were loaded automatically and 1,440 (32%) sequences were manually loaded.

Several assumptions were made when loading GenBank sequences into the HT database. Most accessions deposited in GenBank as cDNAs or mRNAs annotated only the coding region. For these accessions, all nucleotides beyond either end of the coding region were assumed to be 3′ or 5′ untranslated sequence. Many genomic sequences also did not contain features indicating the 3′ and 5′ ends of the primary transcript or mRNA. In these cases, only the sequence of the cDNA corresponding to the translated portion of that molecule was placed into the non-redundant database and the transcription unit for that cDNA was annotated as truncated. Similarly, GenBank contains gene fragments with undetermined or incomplete coding regions. In these cases, both the transcription and translation regions were annotated as truncated. When GenBank annotation did not allow for reliable distinction between cDNAs and genomic sequences that were single exons, the sequences were loaded as cDNAs without 5′ and 3′ untranslated regions. Problematic accessions that lacked critical annotations were either loaded manually or discarded. In some cases, cDNA sequences were constructed from overlapping fragments, or from individual exon sequences. SEQV provides tools for constructing and annotating complete sequences from such fragments. The predicted protein sequences were compared with the protein sequence databases to check the constructed sequences for accuracy.

Redundancy among GenBank sequences was addressed at multiple stages. Candidate sequences were first used in BLAST[41] queries against all human accessions in GenBank to identify, by sequence similarity, other entries likely to represent the same

gene. Sequences with matches longer than 100 nucleotides with greater than 98% identity over the entire match region were evaluated, and the longest cDNA was loaded. Coding sequences of less than 30 nucleotides were not loaded. Candidate sequences that were manually loaded were compared in multiple alignments. A single sequence from each set of related sequences that were considered to be from the same gene was loaded; remaining sequences were referenced in the database by accession numbers only. In a separate process, all annotated coding sequences for human genes were extracted from GenBank and compared using a method that detects matches of 100% identity. Coding sequences that were exact matches, greater than 40 nucleotides long, and were entirely contained in a previously loaded coding sequence were stored in the database as accession numbers only. All HT sequences, including alternative splice forms and related accessions from GenBank from each gene, were grouped under a single gene identifier (HG number). Cross-references between HG numbers and the GenBank and GDB accession numbers for the underlying HT sequences are listed in Table 6.

Database searching. Sequences were initially searched for Alu and other repetitive elements. If an Alu repeat was found, a masked copy of the sequence (with the Alu region represented as Ns) was created using XBLAST[49] and was used for subsequent analysis; if Line-1 or transposable human element (THE) elements were found, the sequence was not analysed further. Automatic searching and annotation using a modified FASTA[50] algorithm (GRASTA) identified previously known human genes and exact matches to THCs. This prescreen step used a minimum match length of 50 bp and a minimum percentage match of 97% over the entire aligned region. Each match was also evaluated automatically for evidence of an alternatively spliced or unspliced cDNA, or a chimaeric clone based on unaligned regions of the EST. The prescreen database was composed of EST assemblies (THCs) and the HT dataset. Comparison of the ESTs against this non-redundant set of cDNA sequences significantly reduced naming inconsistencies. Between 30 and 50% of all ESTs from most libraries were automatically annotated by this prescreen step, including more than 95% of those ESTs with exact matches to sequences in GenBank. We found that 2,947 THCs contained an HT sequence and therefore represent a known human gene. An X-windows graphical interface and control program, STP was developed for performing each of these searches in series with specific groups of sequences retrieved from ESTDB based on user-defined or standard queries[39].

THCs and singleton ESTs from TIGR with no matches after the automated annotation phase were searched against all nucleotide sequences in GenBank using BLAST-TO-GRAZE, a procedure that uses BLAST as a filter to find potentially good matches and GRAZE, a modification of the Smith-Waterman algorithm[51], to produce an optimal gapped alignment between two similar sequences. BLAST-TO-GRAZE is only slightly slower than BLAST alone, but produces a more accurate alignment. Peptide searches were performed with all six possible translations of THCs (or ESTs) against a composite database of sequences from GenPept, PIR and Swiss-Prot. Peptide searches using BLAZE (Intelligenetics[52]) with the BLOSUM60 substitution matrix[53] were performed on a MasPar MP-2 massively parallel computer with 4,096 microprocessors (MasPar). Results from each frame were combined in a single output file by MBLZT, which also rejected matches below a threshold of statistical significance based on the observed distribution of match scores.

Results of the nucleotide and peptide database searches were inspected individually using a custom graphic viewing program, BYOB, that interacts directly with ESTDB. If an EST or THC was determined to match a database sequence based on an evaluation of the alignment by an experienced scientist and estimates of statistical significance, the match information and putative identification were saved to ESTDB. Weak matches were given general putative identifications when several proteins in a family were matched in conserved regions (for example, zinc-finger proteins).

Several tools are provided in BYOB to assist in the assignment of consistent putative identifications to ESTs by more than two dozen users. Links are provided between GenBank and HT sequences so that THCs containing HT sequences are automatically associated with the HT name. Sequences from GenBank and other databases are represented in ESTDB by their unique accession numbers. When a match with a particular sequence in one of the databases is saved, BYOB extracts from ESTDB all names of previously assigned matches to that accession and ranks them by order of frequency. By convention, names used previously are preferred to new or less-used names. If the accession has not been matched previously, all names in ESTDB are returned to assist in assigning a gramatically consistent putative identification. THCs also served as an effective way of linking overlapping ESTs to the same name.

Analysis of the coding content of ESTs

Individual ESTs and THCs were analysed using the GRAIL neural network (v1.2)[54] to estimate the probability that the sequence encodes a protein. As well as providing a measure of library quality (the proportion of clones from a library containing coding sequence), predicted translations also allow further analysis of ESTs to search for new gene families, motifs and other structural features. Libraries varied greatly in coding content from 7 to 100% (a red-blood-cell library with nearly all clones α- or β-globin) (Table 7). We found that 69% of libraries contained between 30 and 60% predicted coding regions based on sequencing at the 5′ end of inserts. Interestingly, 48% of THCs were predicted to contain coding regions by GRAIL, whereas only 26% of singleton ESTs (matching neither GenBank sequences or THC sequences) appear as coding regions to

TABLE 7 Overview of the Human Gene Anatomy project

cDNA libraries	
cDNA libraries constructed (obtained)*	245 (3)
cDNA libraries screened/subtracted*	52
Distinct cDNA libraries*	300
Distinct cDNA libraries (dbEST ESTs)	43
Distinct organs/tissues	37
EST and THC sequence data	
Templates sequenced (TIGR)	150,069
Sequencing reactions (TIGR)	165,787
Number of ESTs (TIGR + HGS)	174,472
ESTs from screened libraries	15,521
Number of ESTs from dbEST	130,238
Total ESTs before editing	304,710
Total ESTs after editing†	266,714
Total nucleotides after editing†	83,601,121
Number of THCs	29,599
Total singleton ESTs and THCs	87,983
Total nucleotides in singleton ESTs + THCs‡	39,573,675
Putatively identified THCs	7,642
Putatively identified singleton ESTs	2,538
Average library percent coding ESTs*	48%
Range of predicted coding content in libraries*	7–100%
Average library Alu content*	7.4%
Range of Alu content in libraries*	0–23.5%
Alu-containing ESTs	13,919
Mitochondrial ESTs*	20,266
rRNA ESTs*	630

Several features and summary numbers from the EST project are summarized. All analysis results numbers apply to the combined dataset of 266,714 ESTs, except where noted explicitly.

* Denotes TIGR/HGS only.

† After editing includes after removal of mitochondrial and rRNA ESTs.

‡ Includes HT sequences, which can include additional nucleotides not present in EST sequences.

TABLE 8 Percentage of genes in seven cellular roles across 37 human tissues

Tissues	Cell signalling communication	Gene/protein expression	Cell division/ DNA synthesis	Cell structure/ motility	Cell/organism defence and homeostasis	Metabolism	Unclassified	Distinct genes per tissue
Adipose tissue	9.0	26.4	3.8	8.0	15.5	14.4	22.9	581
Adrenal gland	11.9	23.5	3.7	7.7	9.9	18.7	24.6	658
Bone	11.1	23.1	4.7	9.7	12.1	17.1	22.2	904
Brain	17.3	17.5	4.7	8.2	7.4	18.1	26.9	3,195
Breast	9.2	23.6	4.2	8.9	14.4	14.8	24.9	696
Colon	10.1	22.0	4.3	7.4	17.2	15.1	23.9	879
Embryo	14.3	21.1	4.7	10.4	8.0	16.7	24.9	1,989
Endothelial cells	13.3	25.1	5.0	6.7	8.5	16.0	25.4	1,031
Epididymis	10.6	26.1	4.0	8.4	14.5	13.7	22.7	370
Oesophagus	2.6	22.4	0.0	6.6	51.3	6.6	10.5	76
Eye	11.9	24.3	3.4	7.5	11.0	17.1	24.9	547
Gall bladder	8.9	23.2	3.9	7.6	13.9	17.4	25.1	768
Greater omentum	9.0	20.4	2.4	17.4	15.6	18.0	17.4	163
Heart	11.8	19.3	4.8	11.2	11.4	15.6	25.8	1,195
Kidney	13.5	22.3	3.3	7.6	10.6	17.2	25.5	712
Liver	12.4	19.0	4.9	6.7	12.4	19.4	25.3	2,091
Ovary	7.9	30.5	4.2	8.1	12.5	13.2	23.6	504
Pancreas	12.1	20.9	3.7	7.9	15.8	14.4	25.3	1,094
Parathyroid gland	6.1	34.7	2.0	12.2	8.2	20.4	16.3	46
Placenta	14.6	20.4	4.0	10.0	9.4	15.4	26.3	1,290
Platelet	17.4	34.8	4.3	4.3	13.0	4.3	21.7	22
Prostate gland	11.6	22.4	4.4	7.5	10.5	17.3	26.2	1,203
Red blood cell	12.5	12.5	0.0	0.0	37.5	0.0	37.5	8
Salivary gland	5.3	36.8	0.0	5.3	10.5	5.3	36.8	17
Skeletal muscle	11.8	21.2	3.5	12.5	12.0	17.8	21.2	735
Skin	9.5	26.2	4.0	8.2	12.7	14.3	25.1	629
Small intestine	9.8	18.9	2.9	6.2	24.1	20.2	17.9	297
Smooth muscle	9.1	34.8	1.5	14.4	8.3	10.6	21.2	127
Spleen	10.6	23.6	4.4	7.6	15.0	14.9	23.7	924
Synovial membrane	12.9	23.6	3.5	8.6	10.7	16.6	24.1	813
Testis	10.8	23.7	5.3	6.4	12.5	16.4	24.9	1,232
Thymus gland	11.2	23.1	2.6	7.1	29.9	10.8	15.3	261
Thyroid gland	11.6	27.9	4.5	6.0	8.6	17.0	24.5	584
Uterus	11.4	22.3	4.9	7.3	10.7	17.3	26.2	1,059
White blood cell	12.8	20.7	5.7	6.0	12.4	16.7	25.8	2,164
Average % of genes per role	12.4	21.9	4.4	8.1	11.9	16.4	24.8	

Cellular roles were assigned to the following categories by examination of database annotations and literature cited therein: **Cell signalling/cell communication**, includes receptors, protein modification, hormone/growth factors, intracellular transducers, effectors/modulators, metabolism, cell adhesion and channels/transport proteins. **Gene/protein expression**, includes protein synthesis, translation factors, ribosomal proteins, post-translational modification/targeting, protein degradation, tRNA synthesis/metabolism; RNA synthesis, transcription factors, RNA polymerase, RNA processing; RNA degradation. **Metabolism**, includes amino acids, nucleotides, sugars, lipids, cofactors, protein modification, energy, and carrier proteins/membrane transport. **Cell division/DNA synthesis**, includes cell cycle, apoptosis, DNA synthesis/replication and chromosomal structure. **Cell structure/motility**, includes cytoskeletal, microtubule-associated proteins/motors, and extracellular matrix. **Cell/organism defence and homeostasis**, includes immunology, homeostasis, carrier proteins/membrane transport, stress response, and DNA repair. Average percentage of genes per role do not add to 100% because some genes appear in more than one role or subrole.

GRAIL. This is in part due to the fact that consensus sequences are longer, on average, than the individual ESTs. However, the possibility remains that the more abundant transcripts match the neural network training set used for GRAIL better than the lower abundance mRNAs represented by unique ESTs.

Two measures of GRAIL's effectiveness were used. First, two estimations of the false negative rate were made based on the ability of GRAIL to recognize ESTs known to contain protein-coding regions. GRAIL predicted that 82.7% of ESTs with exact matches in coding regions of known human genes and 78.2% of ESTs with non-exact protein matches would be coding. Therefore, the false negative rate is likely to be about 20%. Second, the false positive rate was estimated by assessing GRAIL's tendency to identify 3′ untranslated sequence as coding. A set of 2,447 3′ untranslated regions from sequences in the HT database was constructed based on GenBank annotations. Of these 3′-UT sequences, 18% were identified as potentially protein coding by GRAIL.

Analysis of potential genomic contamination: Assessing the degree of contamination with intron and intergenic sequences is an important part of quality control for cDNA libraries. Although exceptions have been reported[55], Alu elements generally do not occur in protein-coding regions. Their presence in cDNA libraries may be a result of Alu elements in 3′ untranslated sequences (for which there are a large number of examples), transcription of independent Alu elements, unspliced precursor RNA, or genomic DNA contamination of the library. cDNA libraries differed markedly in their content of Alu sequences, containing from 0 to 23.5% Alu sequence-ESTs; however, 78% of libraries had less than 10% Alu-containing ESTs (Table 7). The Alu content of HT sequences was found to be 5.6%, compared with 7.4% in the EST dataset. The elevated presence of Alu repeats in some libraries may be due to a higher level of genomic or unspliced RNA contamination in these libraries. The Alu content of several large EST datasets is shown in Table 5.

As a separate measure of unprocessed RNA or genomic DNA contamination, ESTs were compared with a dataset of 1,733 annotated intron sequences associated with 344 HT sequences. Of the TIGR and HGS ESTs matching those genes, 2.6% matched in an intron (match criteria: >100 bp with 95% identity), indicating a relatively low overall level of ESTs from unprocessed RNAs in the dataset.

Overview

Of the 300 cDNA libraries sampled, 131 are represented by more than 200 ESTs, and 55 are represented by more than 1,000 ESTs.

TABLE 11 Widely expressed genes

HG/THC	N	Identification
HG942	20	adenylyl cyclase-associated protein
HG1541	20	ATP synthase, alpha chain, mitochondrial
HG777	20	CCAAT-box DNA-binding protein YB-1, HLA class II regulating
HG1126	20	cell surface protein TAPA-1, 26 kDa
HG2000	20	collagen, type III, alpha 1
HG2799	20	collagen, type IV, alpha 1
HG1332	20	cytochrome-c oxidase, Vb subunit
HG3245	20	DNA-binding protein B
HG2750	20	globin, alpha 1
HG2751	20	globin, alpha 2
HG1994	20	glutathione S-transferase pi
HG3099	20	H. sapiens hypothetical protein (GB:D26068)
HG2857	20	heat-shock protein, 27/28 kDa
HG1212	20	integrin, beta 1
HG3524	20	lactate dehydrogenase A
HG3598	20	major histocompatibility complex, class I, A11E
HG223	20	phospholipase A2 (GB:M86400)
HG2021	20	ribosomal protein L35a
HG2349	20	ribosomal protein S28 homologue
HG661	20	transcription factor IIB
HG3514	20	tropomyosin TM30nm, cytoskeletal
HG3516	20	ubiquitin (GB:X04803)
HG530	20	ubiquitin-activating enzyme E1
THC59782	20	
THC60068	20	
THC63271	20	
THC68451	20	
THC69013	20	
THC72733	20	
THC73105	20	
THC74780	20	
HG1151	21	ADP-ribosylation factor 2
HG3358	21	antigen ME491, melanoma-associated
HG3432	21	fibroblast growth factor receptor k-sam
Hg3044	21	fibronectin
HG1428	21	globin, beta
HG4141	21	H. sapiens hypothetical protein (GB:D30757)
HG1465	21	high mobility group protein 1, placenta
HG2806	21	histone H3.2
HG111	21	insulin-like growth factor binding protein 5
HG950	21	Mac2-binding protein
HG4604	21	nascent polypeptide-associate complex, alpha
HG3038	21	phosphoglycerate mutase, B
HG556	21	pigment epithelium-differentiation factor
HG2306	21	ribosomal protein L27
HG1272	21	ribosomal protein S15
HG1905	21	ribosomal protein S16
HG1298	21	ribosomal protein S28
HG302	21	set
HG1322	21	small nuclear ribonucleoprotein, polypeptide C
HG34	21	thyroid-binding protein
HG1980	21	tubulin, beta 2
THC67562	21	
HG402	22	alpha-2-macroglobulin
HG1860	22	ATP synthase, H+ transporting, mitochondrial F1 complex, beta chain
HG4504	22	calgizzarin
HG1833	22	calmodulin type II
HG3010	22	collagen, type I, alpha 1
HG737	22	cystatin C
HG463	22	cytochrome-c oxidase, IV subunit
HG3431	22	decorin
HG662	22	Epstein-Barr virus small RNA-associated protein
HG557	22	erythrocyte membrane protein 7.2
HG1712	22	H. sapiens hypothetical protein (GB:D14696)
HG3343	22	heterogeneous nuclear ribonucleoprotein E2
HG1283	22	lectin, beta-galactosidase-binding, 14 kDa
HG2917	22	major histocompatibility complex, class I, E (GB:M21533)
HG2990	22	phosphoglycerate kinase
HG1599	22	proliferation-associated gene
HG3640	22	protease, calcium dependent, small subunit
HG4319	22	ribosomal protein L5
HG849	22	ribosomal protein S21
HG324	22	smooth muscle protein SM22
HG172	22	spermidine/spermine N1-acetyltransferase
HG1598	22	tissue-specific protein
HG1515	22	transcription factor BTF3b
HG2083	22	translation elongation factor 1, delta
HG1693	22	translation initiation factor 4A
HG4130	22	translation initiation factor 4AII
HG1793	22	tubulin alpha k1
HG1642	22	vimentin
THC56917	22	
THC58167	22	
THC62616	22	
THC67258	22	
HG1532	23	14-3-3 protein
HG922	23	calnexin
HG1354	23	casein kinase 2, beta polypeptide
HG1764	23	collagen, type I, alpha 2
HG753	23	DNA-binding protein TAXREB67
HG1715	23	H. sapiens hypothetical protein (GB:D14812)
HG2455	23	heat shock protein, 86 kDa
HG2004	23	heat shock protein, 89 kDa alpha
HG2211	23	pyruvate kinase, M1/M2 type
HG311	23	ribosomal protein L30
HG3413	23	ribosomal protein L7
HG613	23	ribosomal protein S12
HG1938	23	ribosomal protein S19
HG1919	23	ribosomal protein S25
HG1506	23	RNA helicase p68
HG526	23	thymosin beta-10
HG1414	23	thymosin beta-10 (GB:S54005)
HG1981	23	tissue inhibitor of metalloproteinase 1
HG3485	23	translation initiation factor 4B
THC68993	23	
HG1474	24	actin, gamma, smooth muscle, aorta
HG2729	24	beta-2-microglobulin
HG1219	24	complement-associated protein SP-40
HG2030	24	glutamate ammonia ligase
HG3550	24	guanine-nucleotide-binding protein, G(s) alpha subunit
HG1705	24	H. sapiens hypothetical protein (GB:D14662)
HG2940	24	major histocompatibility complex, class II, DR, invariant region
HG760	24	osteonectin
HG2872	24	ribosomal protein L11 homologue
HG934	24	ribosomal protein L18
HG4320	24	ribosomal protein L21
HG1520	24	ribosomal protein L23
HG384	24	ribosomal protein L26
HG4323	24	ribosomal protein L27a
HG1494	24	ribosomal protein S2
HG3530	24	ribosomal protein S7
HG745	24	ribosomal protein S8
HG672	24	translation elongation factor 1, beta
HG1423	24	ubiquitin carboxyl-terminal extension protein
THC60316	24	
THC60908	24	
THC68882	24	
HG1307	25	aldolase A
HG1584	25	breast basic conserved protein 1
HG2788	25	calcyclin
HG730	25	high mobility group protein
HG1538	25	interferon-induced protein 1-8U

TABLE 11 continued

HG/THC	N	Identification	HG/THC	N	Identification
HG1773	25	laminin receptor 1	HG1738	27	tubulin, beta d1
HG1075	25	nucleolar phosphoprotein B23	THC70403	27	
HG2893	25	phosphatidylethanolamine-binding protein	HG772	28	ADP/ATP translocase
HG1627	25	polyadenylate-binding protein (GB:Y00345)	HG417	28	cathepsin B
HG1485	25	ribosomal protein L31	HG2020	28	cyclophilin A
HG3364	25	ribosomal protein L37	HG418	28	enolase, alpha
HG1058	25	ribosomal protein L7a	HG2855	28	heat shock protein, 70 kDa (GB:Y00371)
HG4326	25	ribosomal protein S10	HG449	28	ribosomal phosphoprotein P2
HG1310	25	ribosomal protein S17	HG1299	28	ribosomal protein L18a
HG4324	25	ribosomal protein S5	HG1571	28	ribosomal protein L19
HG447	25	thymosin beta-4	HG3191	28	ribosomal protein L4
HG741	25	ubiquitin-52 amino acid fusion protein	HG586	28	ribosomal protein S11
THC51021	25		HG3559	28	ribosomal protein S18
THC62444	25		HG1800	28	ribosomal protein S20
THC64762	25		HG3474	28	ribosomal protein S24
HG4281	26	Csa-19	HG874	28	ribosomal protein S6
HG1323	26	heat shock protein, 90 kDa	HG2037	28	translation elongation factor 1, gamma
HG1028	26	heterogeneous nuclear ribonucleoprotein protein A1	HG1809	29	actin, beta
			HG1832	29	actin, gamma 1
HG2815	25	myosin, light chain, alkali, smooth muscle (GB:U02629)	HG1777	29	glyceraldehyde-3-phosphate dehydrogenase
			HG2026	29	highly basic protein, 23 kDa
HG3424	26	prothymosin, alpha	HG4567	29	major histocompatibility complex, B0704
HG2749	26	proto-oncogene rhoA, multidrug resistance protein	HG3574	29	major histocompatibility complex, B61
			HG3583	29	major histocompatibility complex, Bw62.3
HG448	26	ribosomal phosphoprotein P1	HG2907	29	major histocompatibility complex, class I, B (GB:M16102)
HG1801	26	ribosomal protein L12			
HG3455	26	ribosomal protein L28	HG2914	29	major histocompatibility complex, class I, B (GB:M19757)
HG4718	26	ribosomal protein L35			
HG1531	26	ribosomal protein S3	HG2918	29	major histocompatibility complex, class I, B (GB:M24037)
HG4478	26	ribosomal S30/ubiquitin fusion protein			
HG1487	26	translationally-controlled tumor protein	HG3609	29	major histocompatibility complex, class I, B27
HG2279	26	triosephosphate isomerase	HG1644	29	major histocompatibility complex, class I, B35
HG1979	26	tubulin, alpha b1	HG3070	29	major histocompatibility complex, class I, Bw57.2
THC72731	26				
HG1017	27	annexin II	HG2904	29	major histocompatibility complex, class I, C (GB:M11886)
HG1654	27	cofilin			
HG3183	27	DNA-binding protein TAX	HG2920	29	major histocompatibility complex, class I, C (GB:M26430)
HG958	27	H. sapiens hypothetical protein, liver (GB:L13799)			
			HG2921	29	major histocompatibility complex, class I, C (GB:M28172)
HG3066	27	major histocompatibility complex, B homologue	HG658	29	major histocompatibility complex class I, C (GB:X58536)
HG3214	27	metallopanstimulin 1			
HG1524	27	myosin, light polypeptide 2, regulatory	HG1409	29	ribosomal protein L41 homologue
HG761	27	prosaposin	HG214	29	ribosomal protein S3a
HG4542	27	ribosomal protein L10	HG33	29	ribosomal protein S4, X-linked
HG1511	27	ribosomal protein L17	HG1859	29	ubiquitin (GB:M26880)
HG2873	27	ribosomal protein L30 homologue	HG1969	30	ferritin, heavy polypeptide
HG3511	27	ribosomal protein L32	HG1812	30	ferritin, light polypeptide
HG1303	27	ribosomal protein L37a	HG1829	30	ribosomal phosphoprotein P0, acidic
HG3349	27	ribosomal protein L8	HG2053	30	ribosomal protein L3, isoform 2
HG2350	27	ribosomal protein L9	HG2268	30	ribosomal protein S14
HG821	27	ribosomal protein S13	HG4325	30	ribosomal protein S9
HG3324	27	TEGT	HG1784	30	translation elongation factor 1, alpha
HG2265	27	translation initiation factor sui1iso1	HG3549	30	Wilms' tumour-related protein

HT sequences and THCs matched in at least 20 of the 30 tissues from which more than 1000 TIGR ESTs were obtained are listed with the number of tissues. Tissues with more than 1,000 ESTs are listed in Table 2.

Including ESTs from dbEST, a total of 304,710 ESTs were analysed for this study (Table 7). This number includes 101,344 ESTs sequenced at TIGR, an additional 73,128 ESTs obtained from clones sequenced at HGS which overlapped sequences obtained at TIGR or public sequence databases, and 130,238 ESTs from dbEST. After trimming low-quality sequences and removal of vector, mitochondrial and ribosomal RNA ESTs, 266,714 ESTs potentially derived from nuclear-encoded mRNAs remained. As noted above, these 266,714 sequences were reduced to 29,599 THCs and 58,384 singleton ESTs. Of the 29,599 THCs, 7,642 (25.8%) had matches in the public databases (including the 2,947 THCs containing HT sequences). Of the 19,667 singleton ESTs sequenced at TIGR, 2,538 (12.9%) have putative identifications. These are listed in Table 9 (following reference list).

Cellular roles. The Human Gene Anatomy project provides the first opportunity to classify systematically the majority of highly expressed human genes with respect to their roles in cellular biology. Genes matched by ESTs cover a wide range of structural and biochemical functions. To estimate the percentage of expressed genes involved in various cellular processes, proteins encoded by the genes matched by ESTs were grouped into the following six broad categories of biological roles: (1) Cell signalling/cell communication, (2) cell division, (3) gene/protein expression, (4) metabolism, (5) cell structure and motility, and (6) cell/organism defence. Each of the six categories was subdivided further into subrole categories (see index to Table 9). For example, cell division was divided into four subcategories: chromosome structure, cell cycle, DNA synthesis and replication, and apoptosis.

Role categories and subcategories were chosen to encompass a broad view of human cell biology. Although many categorization schemes might be considered equally valid, we have attempted to group together proteins that share similar functional characteristics or cellular roles, rather than by a strict biochemical classification. TIGR scientists and outside reviewers with expertise in biochemistry, genetics, molecular biology, molecular evolution, microbiology, cell biology and physiology worked to refine the categories and to make role assignments. Roles were assigned according to the known or putative involvement of a gene or protein in a cellular process or pathway as opposed to participation in a specific binding or catalysis function. A seventh broad category, unclassified, was used for proteins and genes where the role was not known or could not be assigned with confidence based on searches of the literature.

Where possible, proteins were assigned to subdivisions of the broad role categories (Table 9). Some genes and proteins were included in more than one role or subrole. For example, tyrosine kinase receptors were assigned to the receptor and protein modification subcategories of the broad cell signalling/communication category.

The number of genes putatively associated with each role was calculated for 37 tissues (Table 8) based on analysis of all singleton ESTs from TIGR and all THCs. The proportion of transcripts (averaged across all tissues) associated with each of the broad role categories are as follows: cell signalling/cell communication, 12%; RNA synthesis and processing, 6%; protein synthesis and processing 15%; metabolism, 16%; cell division and DNA synthesis, 4%; cell structure/motility, 8%; cell/organism defence and homeostasis, 12%; unclassified, 24%. Although this analysis is based on less than 7,000 human genes with known or putative functions out of a total of 50,000–80,000 genes[56], it provides an initial estimate of the number of genes within the genome involved in these basic biological processes. The classifications shown here should be considered preliminary; as more is learned about the cellular roles of genes and proteins, the number and kinds of categories in our current classification scheme as well as the classifications of individual genes and proteins, will probably change. We welcome the input of the scientific community on the curation and refinement of these data.

The putative identifications assigned to ESTs and THCs as described above are listed in Table 9. 10,180 sequences were putatively identified. Table 9 is organized by the cellular role and sub role categories described in the index at its head. With each putative identification is the number of times the sequence was matched and the tissues from which the ESTs were derived.

Perhaps the most interesting cDNA sequences described here are those for which putative identifications could not be made. The expression patterns of 21,957 THCs without database matches are listed in Table 10 (following reference list). Some of these THCs stand out as broadly or narrowly expressed given the number of ESTs represented by the THC (see below). There are no database matches for 17,129 new singleton ESTs. The expression patterns of these previously unobserved sequences, with the data on similarities between sequences in this set (see below), provide starting points for the experimental characterization of the functions of these new human genes.

Analysis of transcript abundance and tissue distribution. Comprehensive information on steady-state mRNA levels is not readily available for most known human transcripts. Northern blots are often used to demonstrate the presence of a limited number of mRNA species in a small number of tissues. In contrast, the EST approach allows for a global examination of steady-state mRNA levels based on cDNA sequence. The accuracy of the expression profile based on EST sequencing depends greatly on the depth of sampling. Although abundant messages should be thoroughly represented, only a limited fraction of the rare cDNAs in any library will be sequenced in a sample size of a few thousand ESTs. cDNA libraries were chosen for the most extensive sequencing based on a large diversity of transcripts (and low content of abundant cDNAs).

EST projects underway in our laboratory and others have provided information on gene expression in several tissues, notably brain[1–4,8], liver[9], heart[14], testis[5], and pancreatic islet cells[10]. The EST data presented here extend the profile to include most major human organs and tissues. Although the failure to observe a transcript by cDNA sequencing at this stage does not prove the lack of gene expression in a particular tissue, the observation that it is expressed in many tissues or in only a single tissue at high levels is informative.

When cDNA libraries are constructed under optimum conditions, the cDNA distribution in a library will broadly match the steady-state mRNA level from which the cDNA was made[9]. We have used cDNA abundance and tissue distribution, as measured by repeated sequencing of cDNA clones representing the same gene, to determine the most highly transcribed genes and to identify potential housekeeping genes. We have recently demonstrated that EST sampling from carefully constructed cDNA libraries can be an effective and quantitative measure of steady-state mRNA levels[67]. Although the EST approach will underestimate the true tissue distribution of genes until assemblies are completed covering coding and non-coding regions of all genes, several observations can be made based on this set of EST data.

A large number of genes were matched by ESTs that are present in only a limited number of tissues. Of the 10,180 different putative identifications assigned, over 4,300 genes were found in only one tissue. This is partly due to seven tissues being represented by only a small number of ESTs (<1,000) (Table 2).

Most of the tissues (30 of 37) examined in this study were sampled in some depth, with at least 1,000 ESTs generated per tissue (Table 2). Only eight genes were found in all 30 tissues (ferritin light and heavy chains, ribosomal phosphoprotein P0, ribosomal proteins L3, S9 and S14, Wilm's tumor-related protein, and translation elongation factor-1α); a further 219 genes were found in two-thirds or more of these tissues (Table 11). Together, these 227 transcripts represent only a small fraction of the total number of identified ESTs. The limited number of broadly expressed, relatively abundant transcripts described herein suggests that there may be much greater differences in gene expression among different cell types than was previously assumed[57].

The genes listed in Table 11 apparently encode moderately to abundantly expressed 'housekeeping' genes, including more than 30 ribosomal proteins, translation factors and major structural proteins of the cytoskeleton. Housekeeping genes have been defined[57] as genes that are essential for cell viability, and therefore are expressed in all cells. Transcripts encoding a much larger number of housekeeping genes, such as various enzymes involved in the central reactions of metabolism, are not represented on this list, presumably because these mRNA species are expressed in cells at low levels, and hence were not found in a large number of tissues at the level of this initial survey (see Table 9), or may be expressed for limited periods of time.

A small number of genes were found to be represented by a relatively large number of ESTs ($\geqslant$15) from cDNA libraries representing a single tissue (Table 12). Most of these genes have previously been sequenced in humans and are characteristic of the tissues from which they are derived, including the histatins from the salivary gland and the pancreatic digestive enzymes. Interestingly, 15 novel genes, represented by THCs, appear to be relatively highly expressed in a tissue-specific manner. The relatively small number of highly expressed, apparently tissue-specific genes is due in part to the fact that some libraries were screened to reduce abundant cDNAs and others were selected for extensive sequencing based on a low content of abundant cDNAs (Fig. 1). Based on our strategy of sampling only a few thousand clones from many different libraries, genes expressed at quite low levels in all cells or at higher levels but with restricted cell specificity are likely to be underrepresented in our database.

TABLE 12 Tissue-specific abundant genes

THC no.	Tissue	Gene name	N
THC59571	Brain		15
THC64082	Brain		15
THC31058	Testis		15
THC67429	Endothelial cells	endothelial leukocyte adhesion molecule 1	15
THC50423	Brain		16
THC53024	Brain		16
THC56705	Brain		16
THC72760	Brain		16
THC33408	Pancreas	carboxypeptidase A2	16
THC67842	Brain		17
THC50003	Brain	cerebellin	17
THC60714	Liver	enoyl-coenzyme A	17
THC59593	Brain	amyloid precursor-like protein	17
THC63647	Lung	pulmonary surfactant-associated protein C	17
THC52989	Brain		19
THC60936	Lung	endogenous retrovirus LTR	19
THC54522	Prostate gland	prostate-secreted seminal plasma protein	19
THC30605	Endothelial cells		20
THC54713	Salivary gland	proline-rich protein *Hae*III, subfamily 1	20
THC60205	Liver	inter-α-trypsin H3, heavy polypeptide	23
THC67374	Testis		24
THC71058	Prostate gland	prostate specific antigen	24
THC57645	Brain	calcium/calmodulin-dependent protein kinase, type II, β	24
THC58311	Placenta		26
THC49945	Brain	pro-opiomelanocortin	27
THC56298	Lung	pulmonary surfactant apoprotein	29
THC69189	Liver	haptoglobin, α-β	30
THC31431	Brain	amyloid precursor-like protein	30
THC59600	Brain		32
THC74799	Pancreas	carboxypeptidase B, pancreatic	40
THC75587	Pancreas	triacylglycerol lipase	41
THC65728	Lung	pulmonary surfactant-associated protein A	44
THC63404	Prostate gland	acid phosphatase, prostate	45
THC63646	Lung	pulmonary surfactant-associated protein C	53
THC58776	Lung		57
THC58616	Lung		58
THC60607	Pancreas	chymotrypsin B1	58
THC62945	Lung	pulmonary surfactant-associated protein B	59
THC63477	Salivary gland	histatin 3	64
THC56025	Pancreas	elastase III	82
THC56021	Pancreas	trypsinogen II	190

THCs matched by at least 15 ESTs from a single tissue, but not from any other tissues among the complete dataset of 266,714 ESTs are listed as potentially tissue-specific, abundant genes. *N*, number of ESTs.

Gene families: To obtain an estimate of the size and number of possible gene families represented by our EST data, proteins encoded by HT sequences were searched against the set of translated ESTs that were similar but not identical to proteins in the public databases. Of the HT sequences, 49% had non-exact matches to at least one EST, suggesting that at least half of the known human genes belong to families of related genes. Some matches revealed simple domain or motif similarities, such as the nucleotide binding fold; others defined full-length isologues that may provide a function similar to the known proteins but with different expression patterns or functional activities. In the human ubiquitin-conjugating enzyme E2 gene family, for example, at least 17 new, distinct gene family members have been assembled from approximately 100 ESTs (E. F. Kirkness, unpublished data).

Among the ESTs with no matches in the public protein databases, several new gene families were defined based on similarities between predicted translations of ESTs and EST consensus sequences. Over 9,000 potential protein-coding regions of ESTs and THCs without database matches were predicted using the GRAIL neural network[54]. These peptide sequences were compared to one another using GRASTA with a filter to eliminate simple sequence repeats (XNU[49]). Preliminary results indicate that about 15% of the new peptide sequences matched one another, so several hundred families may be represented among these ESTs and THCs. These new families may be specific (within the detection limits of conventional similarity searching) to mammals or to vertebrates, or may correspond to ubiquitous gene families that have not been observed previously[58]. Given that the EST dataset is likely to be underrepresented for low-abundance transcripts, these estimates of the number of families must be treated as lower limits.

Human sequence variation: The large EST dataset contains many sequences that differ from the human gene transcripts currently available from GenBank. This dataset allows examination of abundant sequences for sequence variation including alternative splicing, polymorphic repeats and point polymorphisms. As an example, several point polymorphisms were observed in the highly polymorphic mitochondrial D-loop region[59], confirming that libraries were derived from different individuals. Sequence differences in several nuclear-encoded genes are currently under investigation as potential polymorphisms. The protocols developed for assigning putative identifications to ESTs allowed the recognition of alternatively spliced transcripts and polymorphic repeats. Several ESTs were identified that matched alternatively spliced transcripts already documented in the literature, and others were new splice variants.

Conclusion

It is inherently difficult to calculate the proportion of human genes matched by the ESTs described here because neither of the two fundamental variables—the number of genes and the number of genes matched by ESTs—are known with any certainty[56]. ESTs were obtained for 2,947 of the 4,100 genes in the HT dataset (72%). The total percentage of human genes matched is likely to be less than 72% for two reasons, however: (1) the HT set is not exhaustive; and (2) the HT set is possibly biased towards abundantly transcribed genes that are more likely to have been matched by ESTs. We have defined a set of 87,983 distinct, non-overlapping ESTs/THCs (including ESTs from dbEST), but the uncertainties in calculating the undetected redundancy within this set, for genes with complex expression patterns, prevent a more precise estimate of the number of distinct genes. We can assume, however, that some undetected redundancy does exist, and that as many as half of the human genes have been matched by the ESTs described here.

Less than one-quarter of these genes have sequences detectably similar to each other or to sequences from other organisms. There is no reason to believe that the as yet undiscovered genes in the human genome will display a greater overall level of sequence similarity than the ones described here. Hence the common assumption that most genes will, in the end, be classifiable intofamilies by straightforward similarity methods may simply be wrong. The average rate of sequence divergence—and presumably functional specialization—during metazoan evolution may be higher than anticipated.

Mapping these new genes physically and genetically will be of enormous value, both to medical genetics and to the basic understanding of genome structure and function. The dataset presented here doubtless contains hundreds to thousands ofdisease genes; the three new genes associated with colon cancer identified in this data set[60,61] are examples. Several ESTs for

the Polycystic Kidney Disease gene[62] were present in the public databases well before the sequence of the disease-associated cDNA was published. Had these clones been mapped, the gene might well have been discovered much earlier. A focused effort to relate transcripts represented by ESTs or EST assemblies to a genome-spanning clone set such as that described in this volume[63] is clearly in order. TIGR has undertaken to develop a set of approximately 10,000 sequence-tagged sites (STSs) from a unique set of 3′ ends of transcripts identified using the THCs presented here. These STSs are being placed on several existing physical maps at collaborating laboratories worldwide. The resulting transcript map should greatly speed the search for disease genes.

Data, clone and software availability

The EST and THC sequence data and analysis results we have described are available electronically through several components of the TIGR Database. Sequences and associated information on tissue distribution are available through the Human cDNA Database (HCD). The information in HCD is available at two levels. Over 85% of the data, including all data incorporated into THCs from dbEST, are in Level 1. Access to these data is available to all researchers at universities, United States government laboratories, or other non-profit research institutions on the signing of a waiver of liability agreement. There are no intellectual property, confidentiality, or publication restrictions on these data. The remainder of the data are available at Level 2 of HCD. Access to these data is available to all researchers at universities, United States government laboratories, or other non-profit research institutions on the signing of an agreement granting Human Genome Sciences the option to negotiate a licence to commercialize potential products derived from use of these data. HCD has been tested by researchers at fifty-four institutions worldwide.

HCD offers users the ability to search the data by putative gene identification (Table 9) or by EST or THC number (Tables 9 and 10). Users at both levels may also submit a nucleotide or peptide sequence to search against the sequences in HCD. Searches are performed against six-frame translations of the HCD sequences using a modification of the Smith–Waterman algorithm[51] with frame-shift detection to provide highly sensitive cross-species matches. Access to Level 1 data is by e-mail or through the World Wide Web (WWW) at TIGR's WWW site (URL: http://www.tigr.org/). Access to Level 2 data is by e-mail only. HCD agreements and other information about the database are available from TIGR's WWW site or by contacting the TIGR database by e-mail at info@hcd.tigr.org.

Sequences and related biological information for HT sequences are available in TIGR's Expressed Gene Anatomy Database (EGAD). EGAD currently contains over 4,500 HT sequences along with links to function and cellular role classifications (Table 9), and can be searched by HT accession number, gene name or role. Where HT sequences have been incorporated into THCs, tissue distribution information from ESTs is also available in EGAD. EGAD is accessible at TIGR's WWW site.

Over 11,000 TIGR EST sequences have been deposited in the dbEST division of GenBank. These ESTs represent at least one EST from each of approximately 7,500 THCs. The GenBank records contain pointers in the comment field to the THCs to which they belong.

Clones corresponding to TIGR and HGS sequences described in this manuscript are available through the American Type Culture Collection (ATCC) as part of the TIGR/ATCC Human cDNA Special Collection. Clones corresponding to any of these sequences which have been deposited in GenBank have been deposited with the ATCC, and may be obtained by contacting the ATCC directly. Requests for all other clones should be directed to the TIGR database by e-mail at clones@tdb.tigr.org. Level 1 clones are available without intellectual property restrictions to academic scientists. Level 2 clones are available to Level 2 HCD users or to non-HCD users who complete a Clone Material Transfer Agreement.

Information on the relationships between particular clones and human transcripts is available through HCD. The positions of ESTs corresponding to each clone present in a THC are available as HCD reports, as shown in Fig. 4. These reports help to easily identify 5′ and 3′ clones for a given transcript. In cases where they have been assigned, ATCC numbers for the clones are also available in the HCD reports. The ATCC accession numbers in the HCD WWW reports are linked directly to ATCC's Gopher site of product descriptions. Clones corresponding to sequences which have been deposited in GenBank have GenBank accession numbers associated with them in the HCD reports.

Software tools described in Table 4 will be made available to academic researchers upon request. e-mail requests should be directed to tools@tdb.tigr.org.

The data in HCD and EGAD are constantly being updated and curated. For the latest information regarding the data, clones and software tools and all of the access procedures outlined above, please check TIGR's World Wide Web site (URL: http://www.tigr.org/) or send e-mail to tdbinfo@tdb.tigr.org (all TIGR databases) or info@hcd.tigr.org (HCD). □

Received 8 August 1994; revised 21 February 1995; accepted 13 July 1995.

1. Adams, M. D. *et al. Science* **252,** 1651–1656 (1991).
2. Adams, M. D. *et al. Nature* **355,** 632–634 (1992).
3. Adams, M. D., Kerlavage, A., Fields, C. & Venter, J. C. *Nature Genet.* **4,** 256–267 (1993).
4. Adams, M. D., Soares, M., Kerlavage, A., Fields, C. & Venter, J. C. *Nature Genet.* **4,** 373–380 (1993).
5. Affara, N. *et al. Genomics* **22,** 205–210 (1994).
6. Gieser, L. & Swaroop, A. *Genomics* **13,** 873–876 (1992).
7. Liew, C. J. *Molec. cell. Cardiol.* **25,** 891–894 (1993).
8. Khan, A. *et al. Nature Genet.* **2,** 180–185 (1992).
9. Okubo, K. *et al. Nature Genet.* **2,** 173–179 (1992).
10. Takeda, J., Yano, H., Eng, S., Zeng, Y. & Bell, G. *Human molec. Genet.* **2,** 1793–1798 (1993).
11. Soares, M. B. *et al. Proc. natn. Acad. Sci. U.S.A.* **91,** 9228–9232 (1994).
12. Auffray, C. *et al. C. rebd. Séanc. Acad. Sci.* **318,** 263–272 (1995).
13. Okubo, N., Yoshii, J., Yokouchi, H., Kameyama, M. & Matsubara, K. *DNA Res.* **1,** 37–45 (1994).
14. Liew, C. C. *et al. Proc. natn. Acad. Sci. U.S.A.* **91,** 10645–10649 (1994).
15. McCombie, W. R. *et al. Nature Genet.* **1,** 124–131 (1992).
16. Waterson, R. *et al. Nature Genet.* **1,** 114–123 (1992).
17. Collet, C. & Joseph, R. *Biochem. Genet.* **32,** 181–188 (1994).
18. Hofte, H. *et al. Pl. J.* **4,** 1051–1061 (1993).
19. Hoog, C. *Nucleic Acids Res.* **19,** 6123–6127 (1991).
20. Keith, C. *et al. Pl. Physiol.* **101,** 329–332 (1993).
21. Kim, C., Markiewicz, P., Lee, J., Schierle, C. & Miller, J. *J. molec. Biol.* **231,** 960–981 (1993).
22. Park, Y. *et al. Pl. Physiol.* **103,** 359–370 (1993).
23. Uchimiya, H. *et al. Pl. J.* **2,** 1005–1009 (1992).
24. Newman, T. *et al. Pl. Physiol.* **106,** 1241–1255 (1994).
25. Polymeropoulos, M. H. *et al. Nature Genet.* **4,** 381–386 (1993).
26. Durkin, A. S. *et al. Cytogenet. Cell Genet.* **65,** 86–91 (1994).
27. Boguski, M. S. & Schuler, G. D. *Nature Genet.* **10,** 369–370 (1995).
28. Benson, D., Lipman, D. J. & Ostell, J. *Nucleic Acids Res.* **21,** 2963–2965 (1993).
29. Boguski, M., Lowe, T. & Tolstochev, C. *Nature Genet.* **4,** 332–333 (1993).
30. Adams, M. D. *et al. Nature* **368,** 474–475 (1994).
31. Chomczynski, P. & Sacchi, N. *Analyt. Biochem.* **162,** 156–159 (1987).
32. Matsubara, K. & Okubo, K. *Gene* **135,** 265–274 (1993).
33. Hornes, E. & Korsnes, L. *Genet. Analyt. tech. Appl.* **7,** 145–150 (1990).
34. Short, J. M., Fernandez, J. M., Sorge, J. A. & Huse, W. D. *Nucleic. Acids Res.* **16,** 7583–7600 (1988).
35. Gubler, U. & Hoffman, B. J. *Gene* **25,** 263–269 (1983).
36. Hay, B. & Short, J. M. *Strategies* **5,** 16–18 (1992).
37. Holmes, D. S. & Quigley, M. *Analyt. Biochem.* **114,** 193, (1981).
38. Kerlavage, A. R. *et al. In Proc. 26th Hawaii Int. Symp. System Sciences* 585–594 (IEEE Computer Society Press, Los Alamitos, CA, 1993).
39. Kerlavage, A. R. *et al. Engineering in Medicine and Biology* Vol. 14 (ed. Lawrence, C.) (IEEE Computer Society Press, Los Alamitos, CA, in the press).
40. White, O. *et al. Nucleic Acids Res.* **21,** 3829–3838 (1993).
41. Altschul, S. *et al. J. molec. Biol.* **215,** 403–410 (1990).
42. Fields, C. & Adams, M. *Biochem. biophys. Res. Commun.* **198,** 288–291 (1994).
43. Fleischmann, R. D. *et al. Science* **269,** 496–512 (1995).
44. Tsai, J.-Y., Namin-Gonzalez, M. L. & Silver, L. M. *Nature Genet.* **8,** 321–322 (1994).
45. Slightom, J. L., Blechl, A. E. & Smithies, O. *Cell* **21,** 627–638 (1980).
46. Krause, M., Wild, M., Rosenzweig, B. & Hirsh, D. *J. molec. Biol.* **208,** 381–392 (1989).
47. Khan, W. *et al. Genomics* **12,** 780–787 (1992).
48. Wilkie, T. *et al. Nature Genet.* **1,** 85–91 (1992).
49. Claverie, J.-M. In *Automated DNA Sequencing and Analysis* (Academic, London, 1994).
50. Pearson, W. & Lipman, D. *Proc. natn. Acad. Sci. U.S.A.* **85,** 2444–2448 (1988).
51. Smith, T. & Waterman, M. *J. molec. Biol.* **147,** 195–197 (1981).
52. Brutlag, D. *et al. Comput. Chem.* **17,** 203–207 (1993).
53. Henikoff, S. & Henikoff, J. *Proc. natn. Acad. Sci. U.S.A.* **89,** 10915–10919 (1992).
54. Uberbacher, E. & Murai, R. *Proc. natn. Acad. Sci. U.S.A.* **88,** 11261–11265 (1991).
55. Claverie, J.-M. *Genomics* **12,** 838–841 (1992).

56. Fields, C., Adams, M., White, O. & Venter, J. C. *Nature Genet.* **7,** 345–346 (1994).
57. Watson, J. D. *et al. Molecular Biology of the Gene,* 4th edn. (Benjamin/Cummings, Menlo Park, CA, 1987).
58. Green, P. *et al. Science* **259,** 1711–1715 (1993).
59. Vigilant, L. *et al. Science* **25,** 1503–1507 (1991).
60. Papadopoulos, N. *et al. Science* **263,** 1625–1629 (1994).
61. Nicolaides, N. *et al. Nature* **371,** 75–80 (1994).
62. European Polycystic Kidney Disease Consortium. *Cell* **77,** 881–894 (1994).
63. Cohen, D. *et al.* (this volume).
64. McCombie, W. R. *et al. DNA Sequence* **2,** 289–296 (1993).
65. Cuticchia, A. J. *et al. Nucleic. Acids Res.* **21,** 3003–3006 (1993).
66. Dubnick, M. *Comput. Appl. Biosci.* **8,** 601–602 (1992).
67. Lee, N. H. *et al. Proc. natn. Acad. Sci. U.S.A.* (in the press).
68. Utterback, T. *et al. Genome Sci. Technol.* (in the press).
69. Sutton, G. G. *et al. Genome Sci. Technol.* (in the press).

ACKNOWLEDGEMENTS. The Human Gene Anatomy Project was supported entirely by private funds in the form of a research grant from Human Genome Sciences, Inc. to The Institute for Genomic Research, a not-for-profit research institute. Preliminary studies that led to the development of this project were funded by NINDS and the DOE. Tissue for this project was provided in part by the Cooperative Human Tissue Network, which is funded by the National Cancer Institute. RNA from a fetal lung fibroblast cell line was provided by B. Troen. We thank SmithKline Beecham and HGS for their support for equipment, software and personnel to build HCD. We thank Drs. D. Doyle, M. Rodbell, H. Smith and C. Woese for advice and reviews. We also thank the TIGR Chemistry, Computer, Template and Sequencing Core Facilities and the DNA Sequencing Core Facility at Human Genome Sciences, Inc. for technical assistance. We thank Beckman Instruments, Inc., Life Technologies, MasPar Computer Corporation, Applied BioSystems Division of Perkin-Elmer, and Forma Scientific, Inc. for helping to defray the cost of publication of this article. We thank SmithKline Beecham Pharmaceuticals for their co-sponsorship of the human cDNA database at TIGR. The authors dedicate this study to the memory of Wallace H. Steinberg, 1934–1995.