Generation and Analysis of 280,000 Human Expressed Sequence Tags

LaDeana Hillier,[1,4] Greg Lennon,[2] Michael Becker,[1] M. Fatima Bonaldo,[3] Brandi Chiapelli,[1] Stephanie Chissoe,[1] Nicole Dietrich,[1] Treasa DuBuque,[1] Anthony Favello,[1] Warren Gish,[1] Maria Hawkins,[1] Monica Hultman,[1] Tamara Kucaba,[1] Michelle Lacy,[1] Maithao Le,[1] Nha Le,[1] Elaine Mardis,[1] Bradley Moore,[1] Matthew Morris,[1] Jeremy Parsons,[1] Christa Prange,[3] Lisa Rifkin,[1] Theresa Rohlfing,[1] Kurt Schellenberg,[1] M. Bento Soares,[2] Fang Tan,[1] Jean Thierry-Meg,[1] Evanne Trevaskis,[1] Karen Underwood,[1] Patricia Wohldman,[1] Robert Waterston,[1] Richard Wilson,[1] and Marco Marra[1]

[1]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108; [2]Human Genome Center, Lawrence Livermore National Laboratories, Livermore, California 94550; [3]Department of Psychiatry, College of Physicians and Surgeons of Columbia University, and the New York State Psychiatric Institute, New York, New York 10032

We report the generation of 319,311 single-pass sequencing reactions (known as expressed sequence tags, or ESTs) obtained from the 5′ and 3′ ends of 194,031 human cDNA clones. Our goal has been to obtain tag sequences from many different genes and to deposit these in the publicly accessible Data Base for Expressed Sequence Tags. Highly efficient automatic screening of the data allows deposition of the annotated sequences without delay. Sequences have been generated from 26 oligo(dT) primed directionally cloned libraries, of which 18 were normalized. The libraries were constructed using mRNA isolated from 17 different tissues representing three developmental states. Comparisons of a subset of our data with nonredundant human mRNA and protein data bases show that the ESTs represent many known sequences and contain many that are novel. Analysis of protein families using Hidden Markov Models confirms this observation and supports the contention that although normalization reduces significantly the relative abundance of redundant cDNA clones, it does not result in the complete removal of members of gene families.

The recovery of single-pass sequences (known as expressed sequence tags, or ESTs) from random cDNA clones has been pursued as a relatively inexpensive and rapid means to access many of the expressed genes of an organism (Milner and Sutcliffe 1983; Putney et al. 1983). With the advent of high-throughput sequencing technology and an increased interest in genome-wide studies, it became clear that ESTs could be generated in sufficient numbers to provide a rapid means of gene discovery (Adams et al. 1991, 1995; Khan et al. 1992; McCombie et al. 1992; Waterson et al. 1992; Newman 1994; Sasaki et al. 1994; Houlgatte et al. 1995), especially for those searching for human disease genes or constructing physical maps of the human genome. ESTs also have proven valuable for studying temporal and spatial expression patterns (Matsubara and Okubo 1993) and could be used to construct a genome-wide STS based transcript map (Wilcox et al. 1991). In spite of their many uses, the number of ESTs in the public data bases totaled only 38,594 by December 1994.

ESTs have been used extensively in genomic sequencing projects. For example, the comparison of end sequences from *Caenorhabditis elegans* cDNAs (Y. Kohara, unpubl.; McCombie et al. 1992; Waterston et al. 1992) with genomic sequence has enhanced substantially our ability to

[4]Corresponding author.
E-MAIL lhillier@watson.wustl.edu; FAX (314) 286-1810.

identify genes (R. Waterston et al., in prep.; Wilson et al. 1994). ESTs both verify that predicted genes are transcribed and identify splice sites. Further, ESTs have been essential in the analysis of complex gene structures such as long introns, alternative splice sites, operons, and overlapping transcription units. In addition, it has become clear that ESTs need not be lengthy or of high accuracy to be of use; they need only be long enough to specify a unique sequence in the genome, and accurate enough to allow recognition of similarity by commonly used computer programs [e.g., BLAST (Altschul et al. 1990) or FASTA (Pearson and Lipman 1988)]. With the initiation of a coordinated effort to obtain the sequence of the human genome, human ESTs will assume a similar role in enhancing the interpretation and annotation of the genomic sequence. ESTs will also contribute to the development of sequence-ready maps.

The potential for early gene discovery combined with the long-term value of ESTs for analysis of the human genome prompted us, in cooperation with Merck (Aaronson et al., this issue) and the IMAGE Consortium (Lennon et al. 1996), to initiate a large-scale human EST sequencing effort to expand the scope of the publicly available EST data and thereby enhance its utility.

RESULTS

We have submitted to the Data Base for Expressed Sequence Tags (dbEST) 319,311 ESTs from 444,692 attempted sequences (194,031 cDNA clones) for an overall success rate of 72% calculated after removal of poor quality, bacterial, mitochondrial, and vector sequences. The analyses presented here were performed on 280,223 ESTs generated from the 173,620 clones sequenced as of April 1, 1996. These sequences were from 22 libraries (15 normalized and 7 non-normalized) representing 12 different tissues and 3 developmental states (fetus, infant, and adult). Two of the normalized libraries were prepared from RNA isolated from diseased tissue. The details of the construction of the normalized libraries and the advantages offered by them are presented by Bonaldo et al. (this issue). All of the sequences were obtained from oligo(dT) primed directionally cloned cDNAs. An approximately equal number of 5′ and 3′ ends (140,532 and 139,691 sequences, respectively) were sequenced. A list of the libraries sampled and the numbers of sequences obtained from each library are presented in Table 1.

The decision to generate ESTs preferentially from certain libraries was based on measures of library quality and library complexity (see below) as well as the general suitability of the library for large-scale sequencing. Suitability for sequencing initially was determined largely by trial and error. We found that cloning vectors that had M13 primer annealing sites tended to yield higher-quality sequence. Furthermore, M13 primers yielded favorable results over a broad range of DNA concentrations, which allowed sequencing without quantitation of individual samples. This was in sharp contrast to other dye primers that yielded high-quality sequence over narrower ranges in DNA concentrations. An additional factor in choice of libraries was the bacterial cell type in which the library was propagated. Best results were achieved with DH10B (Life Technologies).

Identification of Contaminating Sequences

To maximize the number of genes surveyed we monitored closely the data from each library, sampling extensively from only those libraries that were amenable to sequencing, of high quality, and continuing to yield a high fraction (greater than 40%) of novel sequences. To assay library quality, EST sequences were screened against data bases of bacterial sequences, mitochondrial sequences, and vector sequences. The levels of these contaminating sequences (or "nontechnical failures") are shown for each library in Figure 1. Libraries containing >2% non-recombinant plasmids, 4% bacterial sequences, or 7% mitochondrial sequences were not selected for extensive sampling. In general, the normalized libraries (Bonaldo et al., this issue) showed comparatively low levels of nontechnical failures, with essentially no nonrecombinant plasmids and low levels of bacterial sequences. All of the libraries contained mitochondrial sequences, ranging from a high of 16% of ESTs (N2b5HB55Y) to a low of less than 1% of ESTs (2NbHP8to9W). Sequences considered to be non-technical failures were not submitted to dbEST but are available upon request.

Assaying Library Complexity

Since all libraries were oligo(dT) primed from

Table 1. cDNA Libraries: Attempted and Successful Sequences as of April 1, 1996

Library[a]	Total attempts[b]	5′ attempts	5′ successes (%)[c]	3′ attempts	3′ successes (%)[d]	Clones represented
N. Fetal Liver/Spleen (1NFLS)	112,618	53,940	40,769 (76)	58,678	41,540 (71)	49,979
N. Infant Brain (1NIB)	56,829	28,607	23,566 (82)	28,222	21,042 (75)	25,701
N. Placenta (1NHP)	43,248	21,712	18,931 (87)	21,536	14,822 (69)	19,642
N. Melanocyte (2NbHM)	25,237	10,950	9,957 (90)	14,287	11,786 (82)	14,552
N. Multiple Sclerosis Plaques (2NbHMSP)	17,116	6,978	5,324 (76)	10,138	7,568 (75)	8,801
N. Breast (3NbHBst)	17,075	8,408	5,597 (67)	8,667	3,933 (45)	6,280
N. Adult Brain (N2b5HB55Y)	15,396	7,573	4,285 (57)	7,823	3,758 (48)	5,154
Adult Liver	12,049	5,971	4,446 (74)	6,078	3,971 (65)	4,801
N. Breast (2NbHBst)	11,580	5,784	4,305 (74)	5,796	3,440 (59)	4,755
Adult Lung	11,367	5,672	4,278 (75)	5,695	3,893 (68)	4,707
N. Adult Brain (N2b4HB55y)	9,506	4,849	2,553 (53)	4,657	1,853 (40)	3,169
N. Placenta (2NbHP8to9W)	8,181	4,121	2,993 (73)	4,060	2,753 (69)	3,669
Fetal Spleen	7,955	3,949	3,241 (82)	4,006	3,018 (75)	3,559
Placenta	6,864	3,415	1,922 (56)	3,449	1,718 (50)	2,173
N. Retina (N2b4HR)	6,650	3,327	2,137 (64)	3,323	1,887 (57)	2,639
Fetal Cochlea	6,191	1,663	1,219 (73)	4,528	3,157 (70)	3,435
Ovary	5,895	2,875	1,586 (55)	3,020	1,560 (52)	1,868
N. Fetal Lung (NbHL19W)[e]	4,819	0	NA	4,819	3,956 (82)	3,956
N. Retina (N2b5HR)	4,107	2,029	1,236 (61)	2,078	1,067 (51)	1,458
Olfactory Epithelium	3,930	1,466	951 (65)	2,464	1,647 (67)	1,802
N. Pineal Gland (N3HPG)	3,008	1,568	821 (52)	1,440	640 (44)	1,021
N. Ovary Tumor (NbHOT)	1,533	768	415 (54)	765	380 (50)	499
Total	391,154	185,625	140,532 (76)	205,529	139,691	173,620 (68)

[a]Libraries sequenced. The N denotes normalized libraries.
[b]Total number of attempted sequencing reactions.
[c]5′ ESTs submitted to dbEST after removal of contaminants (see text).
[d]3′ ESTs submitted to dbEST after removal of contaminants.
[e]5′ ESTs had not yet been obtained.

poly(A) tracts, 3′ ESTs were considered likely to provide the 3′ termini of genes. To examine library complexity we performed intralibrary comparisons of all successful 3′ ESTs. ESTs that had significant matches to other ESTs within the library were grouped into families, and those that could not be grouped into families were considered "singletons" (see Fig. 2 for examples). Those libraries exhibiting higher proportions of singletons were considered to be of higher complexity, thus warranting more extensive sampling.

To generate an estimate of the diversity of the entire set of ESTs we performed a similar analysis, first combining the highest quality 3′ EST data from all libraries and then grouping the ESTs as described in Methods. This analysis placed 81,884 of a possible 111,189 3′ ESTs into 14,850 groups. The remaining 29,485 3′ ESTs (27% of the total) could not be grouped and therefore were classified as singletons (Table 2). Thus, there are 44,335 (14,850 + 29,485) distinct 3′ EST groups. Others using different clustering

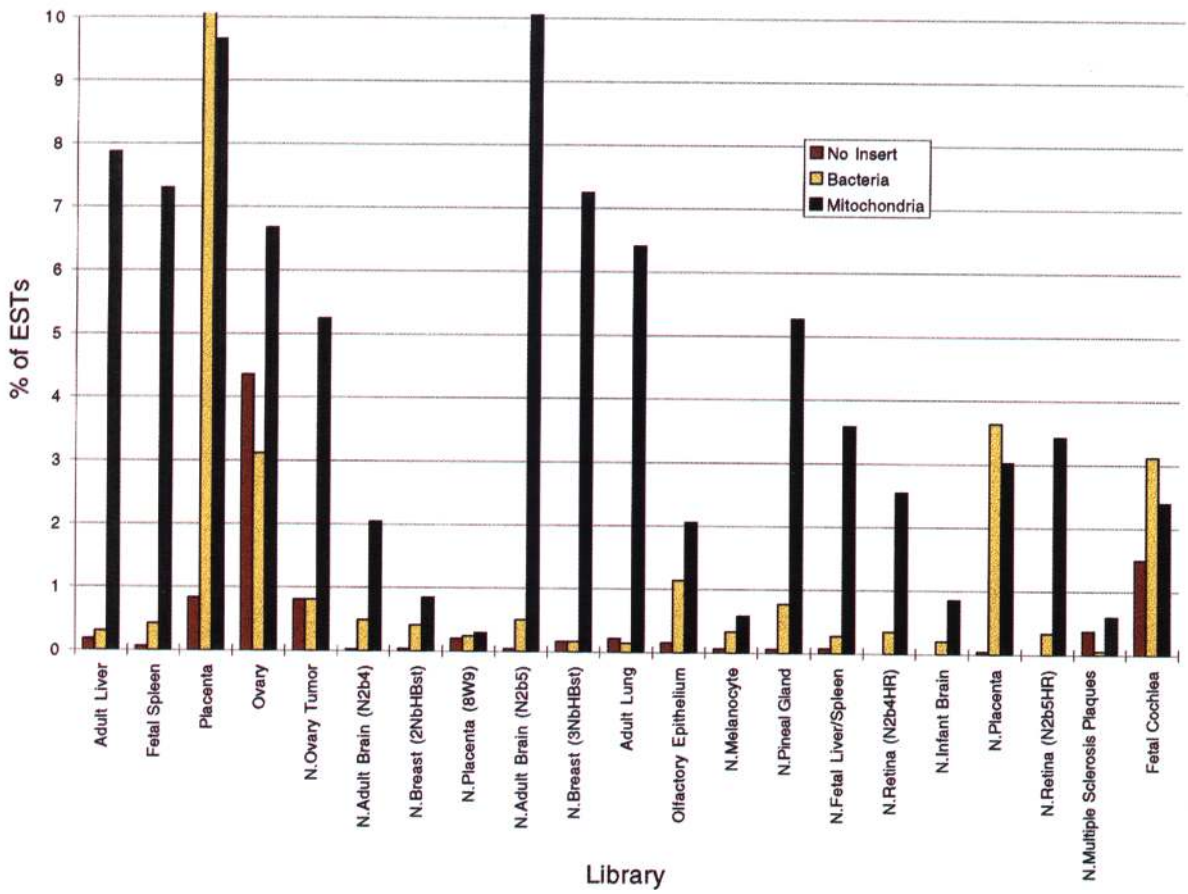

Figure 1 Proportion of ESTs classified as vector, mitochondrial, and bacterial from the normalized (N) and non-normalized libraries. For the normalized retina libraries and the normalized infant brain library, the percentage of nonrecombinant plasmids was less than 0.1%. For the N. Fetal Lung (NbHL19W) library (not shown), the percentages of nonrecombinant plasmid, bacterial, and mitochondrial clones were less than 0.3%. The mitochondrial contamination in the N2b5 adult brain library (16%) and the bacterial contamination in the placenta library (20%) exceeded the range plotted on this graph. The sequences determined to be from nonrecombinant plasmids or of bacterial or mitochondrial origin were not submitted to the public database but are available on request.

algorithms have generated similar estimates (G. Schuler and M. Boguski, in prep.; Aaronson et al., this issue). This number is a rough estimate of the number of distinct genes identified by our EST data. However, this figure could be considered a maximal estimate because it depends heavily on the clustering parameters as well as the frequency of reversed clones, internal or nonspecific priming, and alternatively spliced 3′ exons. To estimate the fractions of these artifacts in our data set the following analyses were performed.

Estimating the Frequency of Nonspecific Priming

To generate estimates of the frequency of nonspecific priming we compared 3′ ESTs with the nonredundant human mRNA data base. Highly significant EST matches were evaluated according to the location of the matched bases in the cognate mRNA and whether another EST was found to match at that position. Examining the entire data set, we found that 27,919 3′ ESTs matched sequences in the human mRNA data base in the correct orientation. Of these, 21,487 (77%) of the ESTs matched the annotated 3′ ends of mRNAs. An additional 21.5% (6009/27,919) of the ESTs matched sequences internal to mRNAs and were confirmed by at least one other EST matching at the same position. The remaining 423 ESTs (1.5%) matched sequences internal to mRNAs but were not confirmed by another EST hit to the same region. These latter ESTs represent

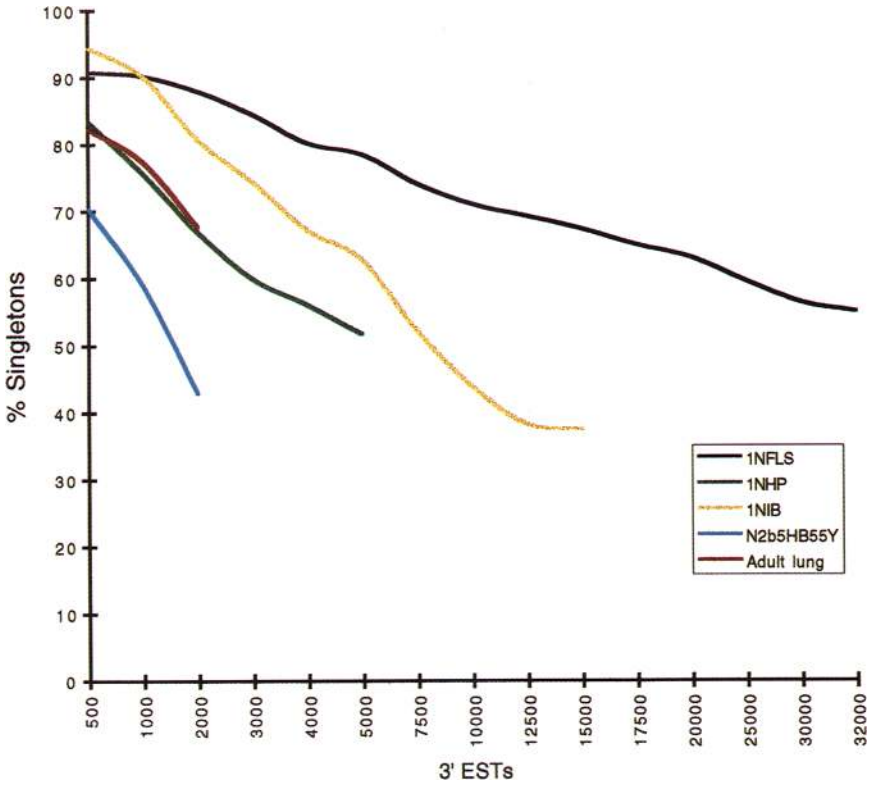

Figure 2 The percent of novel intralibrary 3′ ESTs (singletons) as a function of the number of successful 3′ ESTs is shown. The slopes of the curves indicate the rate at which the fraction of singletons within individual libraries declined. For example, a shallow slope (as seen for the 1NFLS library) indicates that the library continued to be a rich source of new information even after extensive sampling. These data show that even among normalized libraries there are different rates of discovery of novel sequences. The "best" non-normalized library (adult lung) is shown for comparison.

the maximal proportion of nonspecific priming events and may well include infrequent splicing events that have resulted in alternative 3′ exons in the mature mRNA. Data are given for each of the libraries in Table 3.

Estimating the Frequency of Inverted Clones

A substantial fraction of inverted clones would inflate estimates of library complexity. The cDNA inserts could be inverted due to failures in the directional cloning procedure. Alternatively, inserts could appear to be inverted as a result of human-induced errors in EST nomenclature during the sample handling, data tracking, or analysis phases of the sequencing process. To generate an estimate of the frequency of reversed clones the 3′ EST data set was compared with the human mRNA data base. In addition to the 27,919 3′ ESTs that matched in the correct orientation, there were 1,863 3′ ESTs that matched human mRNA data-base sequences in inverted orientation. This represents 6.25% of the total number of ESTs matching a human mRNA sequence in either orientation. Additional experimentation will be required to determine whether these represent artifacts or overlapping convergent transcripts.

To determine the fraction of clones perceived to be reversed because of human error, we searched for redundant EST hits on the wrong strand. Multiple independently derived hits to the same region of an mRNA were unlikely to reflect human error that would be nonsystematic in our data flow paradigm. Further, such hits would support the notion that the perceived reversal was not attributable to a failure in the directional cloning procedure. Of the 1,863 ESTs that hit on the wrong strand, 1327 (71%) were "confirmed" by at least one other hit to the same region of their cognate mRNA, and 536 ESTs (29%) were not (Table 4). Thus the clone reversal rate due to human error can be estimated to be no more than the number of unconfirmed ESTs divided by the number of ESTs found to match the mRNA in either orientation (1.8%).

Estimating the Frequency of Chimeric Clones

As an additional metric of library quality the incidence of chimeric cDNA clones was estimated. Chimeric clones could arise during the cDNA cloning procedure as a result of the artifactual fusion of cDNAs derived from genes unlinked in the genome. Alternatively, clones perceived as chimeric could arise because of errors in sample handling. For example, an error in postelectro-

Table 2. Clustering of 3′ ESTs

Library	3′ ESTs[a]	Singletons[b]	% singletons
N. Fetal Liver/Spleen (1NFLS)	34,156	12,578	36.8
N. Infant Brain (1NIB)	17,282	3,003	17.4
N. Placenta (1NHP)	11,751	2,061	17.5
N. Melanocyte (2NbHM)	10,563	2,113	20.0
N. Multiple Sclerosis Plaques (2NbHMSP)	6,148	1,714	27.9
N. Breast (3NbHBst)	2,652	462	17.4
N. Adult Brain (N2b5HB55Y)	2,715	414	15.3
Adult Liver	2,736	583	21.3
N. Breast (2NbHBst)	2,540	517	20.4
Adult Lung	2,527	840	33.2
N. Adult Brain (N2b4HB55Y)	1,338	231	17.3
N. Placenta (2NbHP8to9W)	2,526	563	22.3
Fetal Spleen	2,143	644	30.1
Placenta	1,199	254	21.2
N. Retina (N2b4HR)	1,601	779	48.7
Fetal Cochlea	2,581	696	27.0
Ovary	916	577	63.0
N. Fetal Lung (NbHL19W)	3,184	578	18.2
N. Retina (N2b5HR)	776	265	34.2
Olfactory Epithelium	1,079	293	27.2
N. Pineal Gland (N3HPG)	415	196	47.2
N. Ovary Tumor (NbHOT)	361	124	34.4
Total	111,189	29,485	26.5

[a]Number of high-quality ESTs included in the clustering analysis (see Methods).
[b]Number of ESTs that were unique among the combined set of high-quality 3′ ESTs from all libraries.

phoresis lane tracking could result in transposition of EST identities for all ESTs on the gel distal to the incorrectly tracked lane. In this case, the clones corresponding to these ESTs would be perceived as chimeric, since their 5′ and 3′ ESTs would derive from different mRNAs.

We compared our entire EST data set with the human mRNA data base and identified those cDNAs that had 5′ and 3′ ESTs that matched the same mRNA. We also identified those cDNAs that had 3′ EST hits to a human mRNA sequence but failed to exhibit a 5′ EST hit to that sequence as candidate chimeric clones. Alternatively, these 5′ ESTs might be derived from cDNAs that are longer than the human mRNA data-base sequence. Another possibility is that these 5′ ESTs identify alternatively spliced mRNA variants not included in the human mRNA data base. Elucidation of the exact nature of the candidate chimeric clones must await the collection of additional cDNA and genomic sequence data. However, the frequency of these events can provide a maximal estimate of the frequency of chimeras. We identified 5399 cDNA clones that had both 5′ and 3′ EST hits to a human mRNA sequence. An additional 57 cDNAs had 3′ EST hits to a human mRNA sequence but no 5′ EST hit to that mRNA, suggesting a frequency of chimeras no greater than 1.04% for the entire data set. Data for each library are presented in Table 5.

Estimating the Frequency of ESTs Derived from Intronic and Intergenic Sequences

To obtain an estimate of the number of cDNA clones derived from intronic or intergenic sequences we compared our 5′ and 3′ EST data with 8.2 megabases of human genomic DNA sequence culled from GenBank. We found 755 cDNA clones that had both 5′ and 3′ EST matches to the genomic sequence. To identify all the cDNAs deriving from mature mRNAs we performed the following analysis on these clones. First, we determined whether there existed other ESTs that confirmed the 5′ or 3′ EST match. We assumed that redundant cDNAs derived from nonspecifically

Table 3. Non-specific Priming

Library	Matches 3′ end or confirmed[a]	Internal and unconfirmed[b]	% unconfirmed
N. Fetal Liver/Spleen (1NFLS)	7429	151	2.00
N. Infant Brain (1NIB)	2851	45	1.60
N. Placenta (1NHP)	3051	26	0.90
N. Melanocyte (2NbHM)	2337	43	1.80
N. Multiple Sclerosis Plaques (2NbHMSP)	944	20	2.10
N. Breast (3NbHBst)	1149	11	1.00
N. Adult Brain (N2b5HB55Y)	834	9	1.08
Adult Liver	1818	17	0.93
N. Breast (2NbHBst)	872	13	1.49
Adult Lung	863	12	1.39
N. Adult Brain (N2b4HB55Y)	362	3	0.83
N. Placenta (2NbHP8to9W)	591	8	1.35
Fetal Spleen	850	9	1.06
Placenta	590	5	0.85
N. Retina (N2b4HR)	227	8	3.50
Fetal Cochlea	717	4	0.56
Ovary	300	6	2.00
N. Fetal Lung (NbHL19W)	962	19	1.98
N. Retina (N2b5HR)	112	5	4.46
Olfactory Epithelium	438	3	0.68
N. Pineal Gland (N3HPG)	85	4	4.70
N. Ovary Tumor (NbHOT)	114	2	1.75
Total	27,496	423	1.54

[a]The 3′ EST matches the annotated 3′ end of the mRNA or a second 3′ EST hits the same region of the mRNA (see text for details).
[b]The 3′ EST matches a region internal to the hu mRNA sequence and lacks a second, confirmatory 3′ EST. These represent the maximal number of nonspecific priming events.

transcribed genomic DNA would be present at low frequencies. Therefore, two or more ESTs that matched the same nucleotides in the genomic DNA were considered to identify an mRNA as were ESTs that had a match that indicated the presence of an intron in the genomic sequence. Finally, we identified those ESTs that matched genomic sequence annotated as coding. This left 17 (2.25%) of the original 755 cDNAs that were potentially derived from intronic or intergenic sequence. Of these 17 clones, 13 had ESTs that matched incompletely annotated GenBank sequences, which made verification of the origin of these clones impossible. Each of the remaining four clones (0.53%) matched a region of an annotated GenBank sequence that was not labeled as coding. Any of these 17 cDNAs could be derived from unannotated exons or nonintron-spanning untranslated regions (UTRs) of mRNAs. This low frequency of potential contaminants would not significantly alter our estimates of library complexity or have generally negative implications for the prediction of transcription units in genomic sequence.

Measuring the Frequency of Discrepancies

ESTs are unedited single-pass sequences and are thus prone to error. This error has different components, including base miscalling and errors in lane tracking. Errors in called bases are observed in sequences considered to be of high quality, suggesting that automated base calling itself is an error-prone process. Estimates of the frequency of errors introduced by the ABI base-calling software have been reported (Sulston et al. 1992).

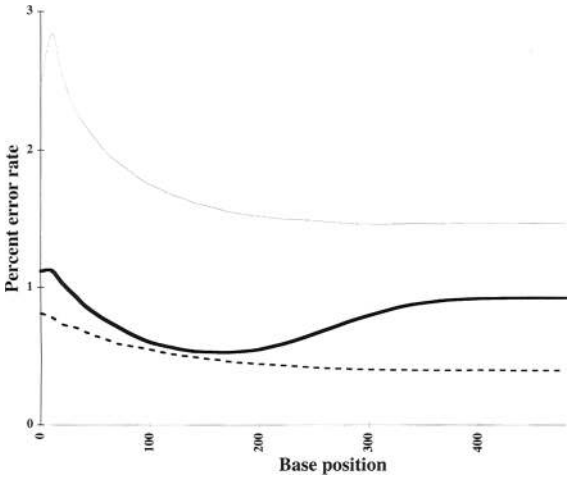

Figure 3 The percent substitution (thin line), insertion (heavy line), and deletion (dotted line) rates for 6000 5′ and 3′ EST sequences with respect to their corresponding human mRNAs. Errors were not examined individually and therefore do not account for sequence polymorphisms, alternative splicing events, or errors in the human mRNA data-base sequences. Since the comparisons were performed on the portion of the submitted sequences annotated as "high quality," the error rates do not increase significantly toward the end of the sequences. These data indicate that the highest quality portion of the EST sequence is between bases 100 and 300.

Thus, ESTs are best suited for purposes that do not rely on error-free data.

We estimated the frequency of errors in our EST data by performing alignments of 5′ and 3′ ESTs to sequences from a data base of nonredundant human mRNAs and identifying discrepancies between the sequences. The discrepancies were categorized as substitutions [mismatches including those involving ambiguous base calls ("Ns")], deletions (a base missing in the EST but present in the mRNA sequence), or insertions (a base present in the EST but not in the mRNA). We then calculated the average frequency of each of the discrepancies per human mRNA–EST alignment for 5′ ESTs and 3′ ESTs for regions annotated as high-quality in the dbEST entry. Data resulting from these calculations are shown in Table 6. These frequencies represent maximal estimates of error in the EST data, because they do not account for sequence polymorphisms, alternative splicing events, or errors in the human mRNA data-base sequences. Error rates for the entire length of the submitted EST that contained additional lower-quality sequences increase by about 1%. Distributions of those error rates across sequences show that for regions of the sequence annotated as high-quality, 94% of the 5′ ESTs had a substitution rate of 5% or less and 83% had a substitution rate of 2% or less. For the high-quality region of the 3′ ESTs, 93% had a substitution rate of 5% or less and 73% had a substitution rate of 2% or less. Average read lengths were 242 bases for the high-quality portions of the read and 364 for the entire submitted read. As shown in Figure 3, substitution error rates are highest at the beginning of the read; because the reads are trimmed based on quality, the error rates do not deteriorate significantly near the end of the read. Insertion errors are also higher at the 5′ end of the read, reach a minimum around 150 bases, and then increase after moving past about 250 bases. The region of the EST between 100 and 300 bases is the most accurate portion of the sequence.

What Fraction of Public Domain Sequences Are Hit by ESTs?

To assay the extent of overlap between our ESTs and other public sequence data we used BLAST programs to compare the ESTs with a number of data bases. These included the human mRNA data base described above, a version of SWISS-PROT (Bairoch and Boeckman 1994; release 32) containing only human protein sequences, and a data base of all publicly accessible yeast proteins. In addition, we used a sensitive Hidden Markov Model (HMM) analysis (Krogh et al. 1994; Eddy et al. 1995) to examine the representation of 171 different protein families in the EST data. We consider each of these analyses below.

The proportion of human mRNA data-base sequences hit by an EST was calculated for BLASTN2 P values ranging from 10^{-9} to 10^{-209}. With decreasing P values, the fraction of the sequences in the data base exhibiting a match of at most that P value decreased as expected (Fig. 4). Of interest is the large fraction of mRNA sequences represented in the ESTs. For example, a P value of 10^{-109} in this analysis corresponds to a BLASTN2 score of approximately 1000. Using our selected parameters where a match receives a +5 score and a mismatch receives a −11 score, a cumulative score of 1000 can be achieved by two sequences having 100% identity over a span of 200 bases. Approximately 72% of the 4169 human mRNA sequences were similar to 35,968 dif-

Table 4. Estimating the Frequency of Inverted Clones

Library	Total 3′ EST hits[a]	Coding strand hits (inverted clones)[b]	% coding strand hits	Confirmed coding strand hits (%)[c]	% unconfirmed coding strand hits[d]
N. Fetal Liver/Spleen (1NFLS)	7949	351	4.42	222 (63)	1.6
N. Infant Brain (1NIB)	2966	58	1.96	38 (66)	0.67
N. Placenta (1NHP)	3429	340	9.92	251 (74)	2.6
N. Melanocyte (2NbHM)	2640	255	9.66	226 (89)	1.1
N. Multiple Sclerosis Plaques (2NbHMSP)	1025	60	5.85	39 (65)	2.05
N. Breast (3NbHBst)	1287	122	9.48	102 (84)	1.55
N. Adult Brain (N2b5HB55Y)	961	114	11.86	74 (65)	4.16
Adult Liver	1850	12	0.65	7 (58)	0.27
N. Breast (2NbHBst)	981	95	9.68	75 (79)	2.04
Adult Lung	922	47	5.1	19 (40)	3.04
N. Adult Brain (N2b4HB55Y)	434	67	15.44	58 (87)	2.07
N. Placenta (2NbHP8to9W)	668	67	10.03	46 (69)	3.14
Fetal Spleen	870	9	1.03	2 (22)	0.8
Placenta	606	11	1.82	4 (36)	1.15
N. Retina (N2b4HR)	294	59	20.07	44 (75)	5.1
Fetal Cochlea	742	21	2.83	7 (33)	1.89
Ovary	327	21	6.42	8 (38)	3.98
N. Fetal Lung (NbHL19W)	1074	92	8.57	62 (67)	2.79
N. Retina (N2b5HR)	151	34	22.52	25 (74)	5.96
Olfactory Epithelium	448	7	1.56	4 (57)	0.67
N. Pineal Gland (N3HPG)	105	16	15.24	9 (56)	6.67
N. Ovary Tumor (NbHOT)	121	5	4.13	5 (100)	0
Total	29,850	1863	6.24	1327 (71)	1.8

[a]3′ ESTs that match either the coding or noncoding strand of an mRNA.
[b]3′ ESTs that match the coding strand of an mRNA ("inverted" clones).
[c]The number of inverted clones confirmed by a second 3′ EST hit to the same region of the mRNA.
[d]The number of unconfirmed inverted clones divided by the total number of 3′ EST hits (column 2). This is an estimate of the maximal frequency of clones reversed due to human error.

ferent ESTs at P values of 10^{-109} or better, indicating substantial overlap between our EST data and the human mRNA data base. In a similar fashion we compared our data with a human-only version of the SWISS-PROT (release 32) protein data base of 3281 sequences using TBLASTN (Fig. 4; Methods). Again, we found evidence for substantial overlap between the EST data and the human-only version of SWISS-PROT. Approximately 51% of the 3281 sequences in this data base were hit by ESTs at P values of 10^{-59} or of higher significance.

Similarly, we compared our EST data with a set of all of the publicly available 8269 yeast proteins. While the yeast data set is reportedly nonredundant, some redundancy clearly remained given that there are approximately 6000 genes in the yeast genome (Dujon 1996; Johnston 1996). Strongly conserved sequences were identified between the yeast and EST data sets (Fig. 4) in spite of the limited length and accuracy of the ESTs. We found that 89 yeast proteins were similar to 37 different translated ESTs (considering only the EST with the highest degree of similarity) at TBLASTN P values of 10^{-79} or better. Closer inspection of the hits revealed 24 ribosomal proteins, 12 heat shock proteins, 5 tubulins, and 2 actins (Table 7). These proteins are among the most strongly conserved across evolution; their inclusion here is not surprising. The presence of highly similar genes in model organisms like yeast offers the possibility of detailed functional characterization (see, e.g., Tugendreich et al. 1993).

Table 5. Estimating the Frequency of Chimeric Clones

Library	5′/3′ EST pairs[a]	Potential chimeras[b]	% chimeras
N. Fetal Liver/Spleen (1NFLS)	1528	12	0.78
N. Infant Brain (1NIB)	408	9	2.16
N. Placenta (1NHP)	734	4	0.54
N. Melanocyte (2NbHM)	270	7	2.53
N. Multiple Sclerosis Plaques (2NbHMSP)	94	6	6
N. Breast (3NbHBst)	269	0	0
N. Adult Brain (N2b5HB55Y)	189	0	0
Adult Liver	647	0	0
N. Breast (2NbHBst)	189	1	0.53
Adult Lung	263	3	1.13
N. Adult Brain (N2b4HB55Y)	91	1	1.09
N. Placenta (2NbHP8to9W)	93	3	3.13
Fetal Spleen	186	5	2.62
Placenta	141	1	0.7
N. Retina (N2b4HR)	43	1	2.27
Fetal Cochlea	65	0	0
Ovary	68	0	0
N. Fetal Lung (NbHL19W)[c]	–	–	–
N. Retina (N2b5HR)	25	2	7.41
Olfactory Epithelium	51	0	0
N. Pineal Gland (N3HPG)	20	0	0
N. Ovary Tumor (NbHOT)	25	2	7.41
Total	5399	57	1.04

[a]The number of 5′/3′ EST pairs hitting the same mRNA.
[b]The number of 5′/3′ EST pairs in which the 3′ EST hit the mRNA but the 5′ EST did not. This number approximates the maximal number of chimeric clones.
[c]5′ EST data did not exist for this library at time of manuscript preparation.

To examine the representation of different protein families in the EST data we performed a HMM analysis of the translated 5′ EST data as described in Methods. This involves comparison of the query sequence with a "model" constructed from a sequence alignment of multiple members of a protein family. We chose to examine only the 5′ ESTs because the HMM analysis is computationally intensive and because 5′ ESTs are more likely to contain sequence coding for protein. The method is sensitive so that even highly divergent members of well-characterized protein families can be recognized. The method is suited to analysis of ESTs because the sequence of the entire protein need not be known, and the method is somewhat tolerant of frame shifts.

HMM analysis of 139,418 translated 5′ ESTs allowed us to assign 10,899 ESTs to all but 39 of a possible 171 protein families for which models could be constructed. Some families had many

Table 6. Discrepancies between ESTs and Human mRNA Sequences

High quality	mRNA matches	% substitutions	% deletions	% insertions
5′ ESTs	3000	1.38	1.14	0.74
3′ ESTs	3000	1.6	0.94	0.42

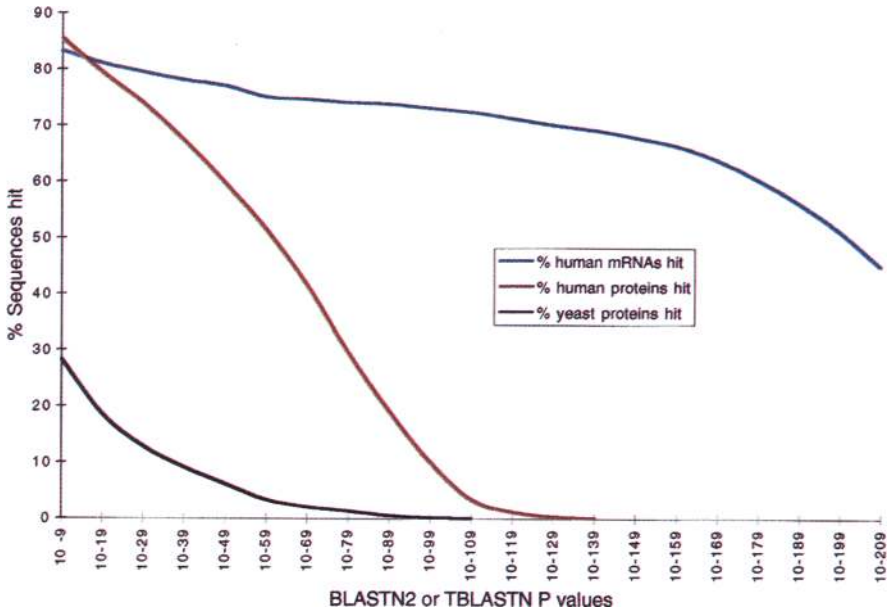

Figure 4 The proportion of sequences from a yeast protein data base, nonredundant human mRNA data base, and human protein data base (all human proteins in SWISS-PROT) showing similarity to at least one human EST. These proportions are shown as a function of *P* value (smaller *P* values indicate matches of higher significance), where the *P* value is calculated by BLASTN2 (nonredundant human mRNA dat abase) or TBLASTN (protein data bases). These data show that a majority of the sequences in the human data bases were represented in our EST data, and that more than 15% of the yeast proteins were significantly similar to one or more of the human ESTs.

representatives in the EST data. For example, sequences classified as globins by HMM analysis comprised 33% (3601/10,899) of all ESTs that could be assigned to a family. The majority of the globins were contributed by the normalized fetal liver/spleen (1NFLS; 2095 ESTs) and placenta libraries (1NHP; 557 ESTs), and the non-normalized fetal spleen library (731 ESTs). Although the number of globin sequences contributed by the 1NFLS library was large, globins comprised only 5% (2095/40,618) of all 5′ ESTs from the library. In contrast, globins comprised 24% (737/3020) of all 5′ ESTs obtained from the non-normalized fetal spleen library. This large reduction in the proportion of globin sequences could be due to the effects of normalization, although this cannot be stated with certainty as these two libraries were derived independently from different tissue sources and were not treated identically prior to normalization.

There were other large families that were apparent as a result of the HMM analysis. These included the glyceraldehyde-3-phosphate dehydrogenases (215 ESTs), protein kinases (363 ESTs), ras proteins (363 ESTs), serine protease inhibitors (643 ESTs), actins (191 ESTs), tubulins (310 ESTs), and intermediate filament proteins (173 ESTs). A complete list of the protein families we used for HMM analysis and our results per library are available at http://genome.wustl.edu/est/esthmpg.html/hmm_results.

There were 39 protein families for which representatives could not be identified in the ESTs. To understand why these families were not represented we examined the proteins used in the HMM multiple sequence alignments. For 25 of the 39 nonrepresented families, the HMM alignments were constructed exclusively of bacterial proteins (10/39), mitochondrial proteins (6/39), plant proteins (6/39), viral proteins (2/39), or snake venoms (1/39). Certain alignments contained bacterial, plant, and fungal sequences (8/39). Given that these 33 alignments contained only prokaryotic or plant proteins, or proteins that we had removed intentionally as contaminants (i.e., mitochondrial proteins) their absence in the EST data is not surprising. Indeed, at least three of these families are specific for photosynthesis (ribulose bisphosphate carboxylase large and small subunits and photosynthetic reaction center proteins).

The six remaining families for which representatives were not found in the EST data include the laminin Bs, the C-type lysozymes, the Zn proteases, the protein hormones, the interferons, and the peroxidases. There are several possible explanations for the failure of the EST data to contain representatives of these families. These genes may not be expressed in the tissues from which we have obtained cDNA libraries, or they may be expressed at only very low levels in these tissues. Alternatively, there may be a technical reason for their absence. For example, 3′ UTRs could be so long that even 5′ ESTs would fall

Table 7. Yeast Proteins Matching Human ESTs

Reference[a]	EST	*P* value	Accession no.	Description
GP-3435	yu56e04.r1	9.00E-82	V01296	beta-tubulin
GP-3328	yw71d04.r1	2.20E-118	V01289	actin
GP-4396	yx16a10.r1	1.30E-80	X17204	L4 protein
SW-H3_YEAST	yx41c04.s1	4.20E-80	P02303	histone h3
GP-472523	yj93f04.r1	2.80E-102	M27070	protein phosphatase 1
SW-TBB_YEAST	yu56e04.r1	7.10E-82	P02557	tubulin beta chain
GP-313261	yv70e08.s1	1.30E-109	X73532	TEF1 gene product
GP-914973	yw82b03.r1	6.70E-85	U32445	60S ribosomal protein L16
GP-172444	yx16a10.r1	2.80E-83	M88608	ribosomal protein L4
GP-172713	yx25f09.r1	4.70E-101	M17583	70-kD heat shock protein
GP-407521	yx25f09.r1	5.50E-92	Z26879	chaperone
GP-312352	yx25f09.r1	6.70E-92	X12926	SSA1 protein
PIR-HSBY3	yx41c04.s1	4.50E-81	HSBY3	histone H3
GP-671634	yx60a10.r1	5.20E-83	X66206	pid:g671634
SW-TSA_YEAST	yx71e09.r1	2.50E-80	P34760	thiol-specific antioxidant protein (prp)
GP-468426	yx74d12.r1	4.90E-88	L31405	ribosomal protein S3
SW-RS3_YEAST	yx74d12.r1	6.10E-89	P05750	40s ribosomal protein s3
SW-RS6_YEAST	yx86d02.r1	7.90E-81	P02365	40s ribosomal protein s6
GP-173058	yx99h01.r1	1.70E-87	M28429	alpha tubulin
PIR-B22696	ya02a07.r3	2.50E-104	B22696	polyubiquitin 6
PIR-D29456	ya02a07.r3	7.70E-92	D29456	ubiquitin precursor UBI4
SW-EF1A_YEAST	ya02a12.r3	1.50E-96	P02994	elongation factor 1-alpha
SW-RHO1_YEAST	yg93g02.r1	2.20E-80	P06780	rho1 protein
SW-PP12_YEAST	yj93f04.r1	5.10E-103	P32598	serine/threonine protein phosphatase pp1-2
SW-DHSA_YEAST	yl95a06.r1	6.50E-81	Q00711	succinate dehydrogenase (ubiquinone) flavoprotein subunit precursor
SW-P2A2_YEAST	yl95b06.r1	1.70E-97	P23595	serine-threonine protein phosphatase pp2a-2 catalytic subunit
SW-P2A1_YEAST	yl95b06.r1	5.60E-97	P23594	serine/threonine protein phosphatase pp2a-1 catalytic subunit
SW-UBC4_YEAST	yp57e01.r1	4.90E-86	P15731	ubiquitin-conjugating enzyme e2-16 kd
SW-UBC5_YEAST	yp57e01.r1	6.20E-86	P15732	ubiquitin-conjugating enzyme e2-16 kd
SW-GSP2_YEAST	yv41b06.r1	1.80E-98	P32836	gtp-binding nuclear protein gsp2/cnr2
SW-GSP1_YEAST	yv41b06.r1	3.20E-99	P32835	gtp-binding nuclear protein gsp1/cnr1
SW-CC42_YEAST	yv46d01.r1	1.80E-85	P19073	cell division control protein 42
PIR-S51452	yv46d01.r1	1.80E-85	S51452	cell division control protein CDC 42
SW-TBA3_YEAST	yv70d06.r1	8.70E-84	P09734	tubulin alpha-3 chain
PIR-S50272	yw70f12.r1	1.50E-86	S50272	hypothetical protein YBR1317
SW-RS11_YEAST	yw70f12.r1	3.40E-87	P05755	40s ribosomal protein ys11 (yp28) (s13)
PIR-S16822	yw70f12.r1	4.70E-86	S16822	ribosomal protein S9.e.A
SW-ACT_YEAST	yw71d04.r1	3.00E-118	P02579	actin
SW-YPT7_YEAST	yw76g03.s1	5.30E-84	P32939	gtp-binding protein ypt7
SW-RL16_YEAST	yw82b03.r1	4.70E-85	P06380	60s ribosomal protein l16
PIR-R5BY16	yw82b03.r1	9.10E-86	R5BY16	ribosomal protein L11.e
SW-RS41_YEAST	yw95d08.s1	1.60E-83	P26781	40s ribosomal protein rp41
PIR-S41784	yw95d08.s1	4.20E-82	S41784	ribosomal protein S11.e
SW-RL4B_YEAST	yx16a10.r1	2.40E-82	P29453	60s ribosomal protein l7a-1
PIR-S16810	yx16a10.r1	3.10E-82	S16810	ribosomal protein L7a.e.B
SW-RL4A_YEAST	yx16a10.r1	3.40E-81	P17076	60s ribosomal protein l7a-1
PIR-R5BY7A	yx16a10.r1	4.40E-81	R5BY7A	ribosomal protein L7a.e.A
PIR-S58785	yx19b12.r1	2.10E-104	S58785	ADP-ribosylation factor 2
PIR-A36367	yx19b12.r1	3.00E-104	A36367	ADP-ribosylation factor 2
SW-ARF2_YEAST	yx19b12.r1	5.60E-104	P19146	adp-ribosylation factor 2
SW-ARF1_YEAST	yx19b12.r1	7.80E-104	P11076	adp-ribosylation factor 1

Table 7. *(Continued)*

Reference[a]	EST	*P* value	Accession no.	Description
SW-HS75_YEAST	yx25f09.r1	1.10E-80	P11484	heat shock protein ssb1
PIR-S20149	yx25f09.r1	1.30E-80	S20149	heat shock cognate protein SSB1
SW-HS76-YEAST	yx25f09.r1	1.50E-80	P40150	heat shock protein ssb2
PIR-S49859	yx25f09.r1	1.80E-80	S49859	probable heat shock protein N1333
SW-GR78_YEAST	yx25f09.r1	2.00E-80	P16474	78 kd glucose regulated protein (grp78)
SW-HS72_YEAST	yx25f09.r1	2.00E-91	P10592	heat shock protein ssa2
PIR-S20139	yx25f09.r1	2.30E-91	S20139	heat shock protein SSA2
SW-HS74_YEAST	yx25f09.r1	4.10E-91	P22202	heat shock protein ssa4
PIR-B36590	yx25f09.r1	4.70E-91	B36590	heat shock protein SSA4
SW-HS71_YEAST	yx25f09.r1	5.80E-92	P10591	heat shock protein ssa1
SW-HS73_YEAST	yx25f09.r1	7.10E-92	P09435	heat shock protein ssa3
PIR-HHBYA1	yx25f09.r1	7.80E-92	HHBYA1	heat shock protein SSA1
PIR-S48413	yx26a09.r1	3.10E-92	S48413	ribonucleotide reductase large chain 3 homolog
SW-RIR1_YEAST	yx26a09.r1	4.20E-86	P21524	ribonucleoside-diphosphate reductase large chain 1
SW-RIR3_YEAST	yx26a09.r1	7.80E-91	P21672	ribonucleoside-diphosphate reductase large chain 2
PIR-S54490	yx53d11.r1	5.10E-87	S54490	ribosomal protein L15
PIR-S48502	yx53d11.r1	6.00E-87	S48502	ribosomal protein L15
SW-RL15_YEAST	yx53d11.r1	6.60E-86	P05748	60s ribosomal protein yl10 (l13) (rp15r)
SW-BMH2_YEAST	yx60a10.r1	1.00E-83	P34730	bmh2 protein
PIR-S57278	yx60a10.r1	1.40E-83	S57278	14-3-3 protein homolog Bmh2p
PIR-S30863	yx60a10.r1	6.90E-84	S30863	BMH1 protein
PIR-S56166	yx60a10.r1	7.20E-82	S56166	BMH2 protein
SW-BMH1_YEAST	yx60a10.r1	9.50E-84	P29311	bmh1 protein
SW-CC28_YEAST	yx64a08.r1	5.00E-82	P00546	cell division control protein 28
SW-RS4E_YEAST	yx66c10.r1	6.80E-93	P05753	40s ribosomal protein s4 (s7) (ys6)
PIR-S20054	yx66c10.r1	9.00E-93	S20054	ribosomal protein S4
PIR-S46695	yx66c10.r1	9.80E-92	S46695	ribosomal protein S4
PIR-A47362	yx71e09.r1	3.30E-80	A47362	thiol-specific antioxidant
SW-QSR1_YEAST	yx71f07.r1	2.40E-80	P41805	ubiquinol-cytochrome c reductase complex
PIR-S48510	yx74d12.r1	1.10E-89	S48510	ribosomal protein S3.e
PIR-S48401	yx76b11.r1	4.30E-80	S48401	ribosomal protein rp22
GP-1151236	yx86d02.r1	3.50E-80	U43281	Lpg18p
PIR-S53911	yx89c10.r1	1.50E-80	S53911	hypothetical protein N2377
SW-FBRL_YEAST	yx93a11.r1	3.40E-80	P15646	fibrillarin (nucleolar protein 1)
PIR-A29456	yx98c09.r1	1.50E-89	A29456	ubiquitin/ribosomal protein CEP 52
SW-TBA1_YEAST	yx99h01.r1	2.00E-87	P09733	tubulin alpha-1 chain
PIR-S57550	yy86e11.r1	6.90E-80	S57550	hypothetical protein YP9531.10c

[a]Data base from which the yeast sequence was obtained. GP: GenPept; SW: SWISS_PROT; PIR: Protein information resource.

short of sequence that encoded protein, or the 5′ ends of the mRNAs may fall short of domains that would identify them as members of these families.

We tested the possibility that the normalization procedure was eliminating members of gene families, focusing our attention on ESTs identified as tubulins. We first recovered the 3′ ESTs corresponding to those cDNAs of which the 5′ ESTs had been identified as tubulins by HMM analysis and compared these with each other using the DNA sequence assembly engine PHRAP

using stringent parameters (P. Green, unpubl.). Only those ESTs derived from the same gene and containing overlapping sequences would be assembled into a single cluster.

This analysis was performed for the normalized fetal spleen (1NFLS) and normalized infant brain (1NIB) libraries. The 1NFLS library yielded 80 3′ ESTs whose corresponding 5′ ESTs had been identified as tubulins. PHRAP assembly of these resulted in 9 clusters, each containing at least 2 ESTs. The largest cluster contained 31 ESTs. Six ESTs could not be assembled into clusters because of insufficient similarity to other tubulin sequences and were therefore considered singletons. This yielded 15 different tubulin EST types, six of which contained only single ESTs. Analysis of ESTs from the 1NIB library yielded a similar result; there we found 11 clusters and 12 singlets for a total of 23 tubulin EST types. Taken together, these observations provide strong support for the contention that normalization does not result in the complete removal of gene family members from cDNA libraries.

DISCUSSION

A key resource for any EST project is the cDNA libraries from which sequence reads are to be derived. Early in the project, only commercially prepared non-normalized libraries were available. We commenced full-scale production sequencing when normalized libraries became available to us in mid-January of 1995 through the IMAGE consortium (Lennon et al. 1996). Since then, these libraries have been the focus of our efforts. In a project whose primary aim is the identification of as many different genes as possible, normalized cDNA libraries have the distinct advantage of having a more uniform representation than non-normalized libraries, where cDNAs are present in proportion to their expression level. Normalized libraries use a hybridization step to reduce relative representation of abundantly expressed genes (Soares et al. 1994; Bonaldo et al., this issue), thereby increasing the proportion of different genes sampled.

In monitoring bacterial sequences and non-recombinant plasmids, we found that most normalized libraries (Bonaldo et al., this issue) contained relatively few bacterial sequences and virtually no nonrecombinant clones. By clustering the 3′ ESTs, which should provide a gene-specific anchored tag for each cDNA, we monitored the level of redundant representation. Because many cDNAs are not full-length, the 5′ ESTs are not similarly anchored and cannot provide a reliable estimate of redundancy or gene representation. For different libraries an unacceptable level of redundancy was attained after different numbers of sequences. For example, the 1NFLS library remains, even after extensive sampling, a relatively rich source of new information. Our analysis of various potential artifacts has shown that this complexity is not a result of problems in library construction, but is likely a reflection of the tissue from which the library was prepared. In contrast, the normalized adult brain libraries were largely redundant after minimal sampling. The process of normalization itself is apparently insufficient to guarantee library complexity, which must also depend on the tissue source.

Comparison of the 3′ ESTs from all 22 libraries and grouping of identical sequences allowed us to estimate the number of unique cDNAs in the data set to be 44,377, a rough estimate of the number of genes sampled. With the problems created by the limited accuracy of single-sequence reads, alternative splicing, alternative polyadenylation sites, and similarity between gene family members, a more accurate estimation awaits interactive analysis of the data and directed collection of additional sequences. Comparison of the results of this cluster analysis with earlier studies (Adams et al. 1995) is also problematic, because most of the sequences generated in that study were derived from unanchored 5′ ends or from unoriented libraries and are not publicly available.

Comparisons of the EST data to other publicly available data show that the ESTs provide an estimate of the fraction of all human genes represented. In the case of the human mRNA data base, ESTs represented over 70% of the 4169 sequences at P values of 10^{-109} or better. An ongoing comparison conducted at the National Center for Biotechnology Information (NCBI; Bassett et al. 1996) shows that, as of April 1996, our data included tags for 33/60 (55%) of the positionally cloned human disease loci (see http://www.ncbi.nlm.nih.gov/dbEST/dbEST_genes). HMM analysis has shown that the EST data contains representatives of 77% of the protein families for which multiple sequence alignments have been constructed (Sonnhammer et al. 1996). When those protein families containing only bacterial, plant, fungus, or viral sequences (or combinations thereof) are removed from consideration, this percentage increases to 96%. Although these

numbers could be interpreted collectively to mean that we have identified in excess of 50% of human genes this must be an overestimate, because all of these data are biased toward genes expressed abundantly.

The identification of over 44,000 different EST clusters from 111,000 different clones is fewer than predicted by Poisson statistics for a population with uniform representation of transcripts arising from approximately 100,000 different genes. Of course, the mRNA population from which these cDNAs are derived is not uniform. The diversity we observe in our data set is presumably enhanced by the use of normalized libraries. In order to increase the number of genes represented by our set, we will sample from additional tissues. Still, rare transcripts (e.g., those present as only a single copy in a cell with hundreds of thousands of mRNA molecules) will be difficult to obtain, and common genes will be sampled repeatedly even in new tissues. Subtraction methods now under development (Bonaldo et al., this issue) may help to increase the representation of rare transcripts and extend the utility of current libraries.

The EST data presented here have proven useful for gene-based mapping strategies (Hawthorn and Cowell 1996; McGuire et al. 1996), with more than 13,600 already placed on the human physical map (M. Boguski, unpubl.). The redundancy of data for the more abundant cDNAs from multiple sources has proven useful for the identification of single-base polymorphisms (P. Kwok, pers. comm.). For genomic sequence annotation purposes, however, it is desirable to obtain the complete, accurate sequence from a nonredundant set of cDNA clones. We are developing such a set based on cluster analysis. In the meantime, ESTs are proving useful in sequence annotation. Groups currently annotate human genomic sequence by relying primarily on the similarities to protein and EST data bases and only secondarily on gene prediction programs, especially for information about the splicing and joining of exons to form genes (G. Miklem and A. King, unpubl.). Similarly, information about UTRs and starting/ending positions of messages depends largely on the additional information provided by the EST data. The data also has been used by individual investigators looking for a human representative of their gene of interest. In fact, half of all clones requested in the early phases of the project represented new human homologs of genes known previously only in other species (G. Lennon, unpubl.). Each cDNA clone is available and can be used to expedite isolation and sequencing of the gene of interest and as a probe to examine tissue-specific or developmentally regulated expression or the cloning of related genes from different species. This growing set of human ESTs, available in the public data bases, should immediately and positively affect the rate of disease gene discovery and the advancement of human biology.

METHODS

Sources of Material

Although we have sequenced primarily from normalized libraries, we have generated a relatively small number of sequences from non-normalized libraries. These included the following commercial libraries: Stratagene (La Jolla, CA) lung (#937210), Stratagene placenta (#937225), Stratagene fetal spleen (#937205), Stratagene liver (#937224), and Stratagene ovary (#937217). Other non-normalized libraries were kindly donated by N. Robertson and C. Morton (Harvard Medical School; Brigham and Women's Hospital; Boston, MA) (fetal cochlea; Robertson et al. 1994) and N. Walker and D. Lancet (an olfactory epithelium library constructed at the Weizmann Institute of Science). All normalized libraries were constructed in the lab of B. Soares (Bonaldo et al., this issue) and were provided by the IMAGE Consortium (Lennon et al. 1996).

Preparation of DNA

Frozen glycerol stocks of cDNA clones to be end-sequenced were received in 384-well format and thawed at room temperature. 5 μl of each glycerol stock were added, using a 12-channel pipettor, to 1 ml of terrific broth (DIFCO) containing the appropriate antibiotic in a 96-well block. Bacterial cell cultures were incubated at 37°C for 24 hr with shaking at 310 rpm in a incubator shaker (Labline) and then processed. For the majority of the ESTs, high-quality DNA was prepared using the 96-Well Miniprep Kit obtained from Advanced Genetic Technologies Corp. (AGTC, Rockville, MD) following the manufacturer's recommendations, except that bacterial pellets were initially resuspended in 50 ml of sterile distilled water. This AGTC kit uses a boiling step in the DNA isolation process. More recently, high-quality DNA has been prepared using the AGTC 96-Well Alkaline Lysis Miniprep Kit. We find that DNA yield and sample throughput are somewhat improved at no detectable loss in DNA quality as measured by sequencing success. DNA preparations were simultaneously performed on four 96-well culture-containing blocks, providing sufficiently high throughput. The final DNA pellet was resuspended in 70 μl or 140 μl of 10 mM Tris pH 8.0, 0.1 mM EDTA. Final DNA concentrations were ~100–400 ng/μl. The success of the procedure was verified by electrophoresis, on 0.7% agarose, of 96 randomly selected DNAs from each set of four 96-well blocks. This

material was sequenced directly without additional quantitation.

Restriction Enzyme Digestion of cDNA Clones

For libraries constructed in pT7T3Pac and the normalized infant brain library constructed in the Lafmid BA vector, 1 μl of DNA (50–200 ng) was incubated with 7.75 μl sterile ddH20, 1 μl of 10× SuRE/Cut Buffer "B" (Boehringer Mannheim), and 0.125 μl (5 units) each of *Hin*dIII and *Eco*RI. Recognition sites for these enzymes occur close to the cDNA insert in the polylinker of the vector. Double digestion with these enzymes releases the cDNA insert from the vector. Typically, "master mixes" containing all of the reaction components except the DNA were assembled on ice, and 9 μl of the master mix were aliquoted into 96-well Cycleplates (Robbins Scientific, Sunnyvale, CA) using a Robbins Scientific Hydra 96-channel mechanical pipetting device (Robbins Scientific). DNA samples were then added to wells containing the master mix. Cycleplates were subjected to a brief centrifugation in a model GR-422 refrigerated floor centrifuge (Jouan) fitted with microtiter plate carriers. Cycleplates were sealed with Scotch Brand heavy duty aluminum foil tape and floated in a 37°C waterbath for a minimum of 1 hr. cDNAs contained in pBluescript (Stratagene) were treated similarly except that *Hin*dIII was replaced with *Xho*I. After incubation, samples were centrifuged briefly to collect the reactions in the bottoms of the wells and 2 μl of 6× loading dye II (Sambrook et al. 1989) were added. Samples were stored at 4°C prior to electrophoresis.

Agarose Gel Electrophoresis and cDNA Size Determination

1.8% agarose gels (SeaKem LE, FMC Bioproducts, Rockland ME) in 1× TAE (Sambrook et al. 1989) containing 48 μg/l ethidium bromide were cast in gel trays constructed of UV transparent plastic following a design of B. Brownstein (unpubl.) modified by D. Panussis (unpubl.). Gel dimensions were 22.4 cm long × 13.2 cm wide. Slots for 4 combs were machined into the gel casting tray at 5.2-cm intervals. Combs were designed to form 24 sample wells with two flanking wells for size standard markers. Sample well-to-well spacing was 0.475 cm (measured center to center), which permitted the use of 12-channel pipettors in sample loading. Each gel could thus accommodate 96 samples divided among four separate rows, with each row containing two marker lanes. Electrophoresis was conducted in 1× TAE in a buffer tank constructed to stack two gel trays at once. In this way high-throughput electrophoresis was achieved, with each electrophoresis apparatus separating 2 × 96 samples per run. Gels were run at 2.1 V/cm (distance measured between the electrodes) for approximately 3 hr, or until the bromophenol blue marker band had nearly reached the next set of wells in the gel. Gels were imaged using a Fluorimager SI Vistra (Molecular Dynamics) with scanner control settings as follows: pixel size, 200 μm; digital resolution, 16 bits; detection sensitivity, normal; PMT voltage, 750. Restriction fragment sizes were obtained semi-interactively using FragmeNT Analysis version 1.1 (Molecular Dynamics) and comparison to DNA size standard VI (Boehringer Mannheim) for all samples that showed complete digestion and were free of apparent contaminating restriction fragments.

An alternative method used to determine the sizes of a small number of cDNA inserts early in the project was the polymerase chain reaction (PCR; Saiki et al. 1988) using vector specific primers flanking the cDNA insert. Electrophoresis and sample analysis were as described above.

DNA Sequencing

Di-deoxy terminator sequencing reactions (Sanger et al. 1977) were performed as described (Fulton and Wilson 1994) or with the following modifications. In the second half of the project we made extensive use of ThermoSequenase DNA polymerase and DYEnamic ET dye primers (Amersham). Sequencing reactions contained 1.5 units of ThermoSequenase and were assembled essentially as described by the manufacturer (Amersham). Dye primer cycle sequencing reactions were conducted in a 96-well format using a MJ Research PTC 200 thermal cycler. Temperature profiles were as follows. For M13 Universal and T3 dye-primer sequencing reactions, the profile was: 95°C, 4 sec; 55°C, 10 sec; 70°C, 60 sec; for a total of 15 cycles, followed by 15 cycles of: 95°C, 4 sec; 70°C, 60 sec. For M13 Reverse dye-primer sequencing reactions, the profile was the same except the annealing temperature was 50°C. The dye primers used on each of the libraries and the primer sequences are given in Table 8. After thermal cycling, sequencing reactions were ethanol precipitated, resuspended in loading buffer containing formamide, denatured, and electrophoresed on ABI 373 or 377 sequencing machines.

Initial Data Processing and Submission of Data to dbEST

Following gel image analysis and DNA sequence extraction, ABI sequence data were automatically processed to: (1) assess EST quality; (2) trim flanking vector sequences; (3) mask repetitive elements; (4) remove contaminated ESTs; (5) identify similarities to ribosomal RNA, human mRNA, and a nonredundant protein data base; (6) identify other cloning artifacts; and (7) determine which portion of the EST to submit. The resulting sequences were annotated with similarity information, sequence quality information (i.e., position in the sequence at which high-quality data ends), and other pertinent information (e.g., library information) and submitted to dbEST/GenBank typically within 48 hr after reaction products had completed gel electrophesis. The details of these procedures follow.

EST Quality

An assessment of EST quality was achieved using peak to highest noncalled peak signal-to-noise ratios and peak to shoulder-width ratios to determine the base position of the start and stop of the high-quality portion of the EST.

Table 8. Sequencing Primers

Vector	Libraries[a]	Vector source	Primers tested[b]	bp from cDNA[c]	Signal intensity[d]	Preferred primers	Primer source	Primer sequences 5′–3′
pBluescript	Stratagene (5)	Stratagene	M13RP (5′)	102	**		ABI	CAGGAAACAGCTATGACC
	Olfactory epithelium library		T3 (5′)	67	*	T3 (5′)	ABI	ATTAACCCTCACTAAAGGGA
	Fetal cochlea library		M13UP (3′)	53	**	M13UP (3′)	ABI	TGTAAAACGACGGCCAGT
			T7 (3′)	26	*		ABI	TAATACGACTCACTATAGGG
Lafmid BA	Soares 1NIB	Soares	M13RP (5′)	12	**	M13RP (5′)	ABI	CAGGAAACAGCTATGACC
			M13UP (3′)	17	–		ABI	TGTAAAACGACGGCCAGT
			M13-40 UP (3′)	37	**	M13-40 UP (3′)	Promega	GTTTTCCCAGTCACGAC
pT7T3-Pac	Soares (14)	Pharmacia	M13RP (5′)	46	–		ABI	CAGGAAACAGCTATGACC
			M13RP Mod. (5′)	46	**		Promega	CAGGAAACAGCTATGAC
			M13REV2 ET (5′)	43	****	M13REV2 ET (5′)	Amersham	AGGAAACAGCTATGACATG
			M13UP (3′)	38	**		ABI/Promega	TGTAAAACGACGGCCAGT
			M13-40UP ET (3′)	38	****	M13UP-40 ET (3′)	Amersham	GTTTTCCCAGTCACGACG
			T7 (5′)	12	*		ABI	TAATACGACTCACTATAGGG
			T3 (3′)	7	*		ABI	ATTAACCCTCACTAAAGGGA

[a]The number in parentheses indicates the number of libraries sequenced.
[b]Dye primers specific for the 5′ and 3′ ends of cDNA clones are indicated.
[c]The number of vector bases from the 3′ end of the primer to the start of the cDNA insert.
[d]A measure of the strength of the signal obtained with each dye primer. (*) Lowest signal; (****) highest signal.

ESTs with 25% or more ambiguous base calls ("Ns") were removed as were ESTs with 50 or fewer bases of high or intermediate quality sequence. We elected to submit shorter high-quality sequences not less than 50 bases for two reasons. First, this length is sufficient to provide an unambiguous gene tag and permit development of STSs. Second, sequences this length were relatively rare. Sequences shorter than 50 bases were submitted only if they showed similarity to a data-base entry. We also elected to submit some lower quality data extending beyond regions of high quality (see below) because this too can be of use. Individual evaluation of the quality of each EST is made possible by the availability of all trace data over the internet from our server.

Removal of Vector Sequences and Masking of Repeats

Vector sequences were trimmed using the programs VEP (Dear and Staden 1991), WEP (W. Gish, unpubl.) and BLASTN (S = 133, S2 = 133, M = 5, N = −8). WEP also served to identify incorrect adaptor sequences and 3′ ESTs lacking poly(A) tails. Repetitive elements were then iden-

tified and masked as follows. HMMFS (S. Eddy and G. Miklem, unpubl.) and REP (Dear and Staden 1991) were used to mask Alu sequences. blastx_and_mask (G. Miklem, unpubl.), which uses BLASTX (S = 50) against a data base of human repetitive elements translated in all six frames, was used to mask other human repetitive elements. The programs TANDEM and INVERTED (R. Durbin, unpubl.) were used to mask local tandem and inverted repeats, and BLASTN (S = 100) of ESTs against a data base of homopolymeric runs was used to identify and mask homopolymer runs.

Contaminating Sequences

Sequences determined to be vector (BLASTN S = 133, S2 = 133, M = 5, N = −11, W = 8 against a vector subset of GenBank), bacterial (BLASTN S = 133, S2 = 133, M = 5, N = −11 against the bacterial division of GenBank) or mitochondrial (BLASTN S = 133, S2 = 133, M = 5, N = −11 against Genbank: HUMMTCG, the human mitochondrion complete genome sequence) were not submitted to the public data bases or included in further analysis. Thresholds for EST matches to "contaminating" sequence is a critical issue. For any threshold, one or more examples of possible contaminants could be found in the large data sets we process. In general, we have tried to set the threshold criteria to avoid identifying a human sequence with a similarity to a bacterial gene as bacterial and thereby withhold a useful sequence. This policy undoubtedly results in the rare submission of bacterial sequences.

Identification of Similarities to Existing Data Sets

EST similarities to ribosomal RNA were identified using BLASTN (S = 133, S2 = 133, M = 5, N = −11) to compare the entire masked EST extending through the lower quality data with the RNA division of GenBank. EST similarities to human mRNAs were identified using BLASTN (S = 133, S2 = 133, M = 5, N = −11) in searches against a nonredundant human mRNA data base. Similarities to proteins were identified using BLASTX (S = 100) searches against SWIR, which is a nonredundant protein data base containing sequences culled from PIR, SWISS-PROT, and a data base of predicted *C. elegans* proteins called WORMPEP (E. Sonnhammer, unpubl.).

EST Data-base Submission

Generally ESTs are submitted from the first high-quality base to the position where two ambiguous base calls (Ns) in five bases are found after the last high-quality base. For high-quality ESTs exhibiting strong BLAST scores this rule was followed if the similarity segment ended before the end of the high-quality region; alternatively, if the BLAST alignment extended beyond the high-quality region, the sequences up until the end of the alignment were submitted. ESTs with no high-quality sequence were nonetheless processed through all the above searches; those that had significant similarity to the GenBank RNA subsection, nonredundant protein data base, or nonredundant human mRNA data base were submitted with the expectation that such sequences and their traces could be of use. In these cases only the portion of the EST that is identified in the BLAST alignment is submitted (in rare cases this has resulted in a submission of fewer than 50 bases).

Comparisons with Other Publicly Available Data

To perform comparative analyses, we considered all our submitted data as of April 1, 1996 (280,223 ESTs). These were masked for known human repetitive elements and low entropy sequences as described above to provide a data base against which to conduct the following searches. Each human protein in SWISS-PROT release 32 (3281 proteins) was compared with our ESTs using TBLASTN (version 1.4) with the parameters M = BLOSUM62, S = 100. The program MSPcrunch (Sonnhammer and Durbin 1994) was used to remove low-entropy similarities. Using the BLOSUM62 amino acid scoring matrix as opposed to the PAM120 or PAM180 scoring matrices biased the results against short stretches of identity. The proportion of human proteins in this data base hit by at least one EST was then determined for 15 P values between 10^{-9} and 10^{-149} in 10^{-10} increments.

Comparisons with the Human mRNA Data Base

A nonredundant human mRNA data base (Boguski and Schuler 1995) of 4169 human sequences was compiled from GenBank release 94. These sequences were compared with our data set using BLASTN2 (W. Gish, in prep.) with the following parameters: S (score corresponding to minimum significance for a single alignment) = 170, gapS2 (minimum reported score) = 150, M (match) = 5, N (mismatch) = −11, Q (gap initiation penalty) = 11, R (gap extension penalty) = 11, B = 5000, filter = seg (removes low-entropy sequences such as homopolymeric tracts from consideration). Unlike BLAST (Altschul et al. 1990), BLASTN2 permits gaps, thereby allowing alignments that included insertions and deletions to be included in the comparison. The proportion of sequences in this data base hit by at least one EST was then determined for 15 P values between 10^{-9} and 10^{-209} in 10^{-10} increments.

Measuring the Frequency of Discrepancies

We obtained an estimate for the average frequency of discrepancies between ESTs and sequences in the nonredundant human mRNA data base as follows. BLAST searches were performed to identify 3000 mRNAs that had hits to both 5′ and 3′ ESTs generated from the same cDNA clone. The 5′ and 3′ EST alignments to the human mRNA sequence were then examined in detail using cross_match (P. Green, unpubl.), a modified Smith-Waterman alignment program. The discrepancies thus identified were classified as substitutions (which included ambiguous bases, or Ns), deletions (bases not present in the EST sequence but present in the mRNA sequence), or insertions (bases present in the EST but not present in the mRNA sequence). For each of these categories the total number of discrep-

ancies per number of bases in the alignment was computed and then averaged across all 5′ EST alignments and across all 3′ EST alignments. Two such calculations were performed, one for the regions of sequences annotated as high-quality in the dbEST entry, and one for the entire length of the submitted EST, which included regions considered to be of lower quality. Rates of substitution, deletion, and insertion errors as they are distributed along the length of the sequence were also calculated for the high-quality portions of the 6000 5′ and 3′ EST sequences.

3′ Clustering to Assay Library Complexity

We applied a clustering algorithm that demanded that 3′ ESTs derived from the same poly(A) tract should align at the ends, thus providing an estimate of the number of anchored 3′ ESTs. The 136,474 3′ ESTs obtained from all libraries were processed as follows: All ESTs containing known repeats were removed and any 5′ polyT regions were trimmed. Any EST containing more than 30 bases of low-entropy sequence [as defined by "dust" (R. Tatusov and D. Lipman, in prep.)] was removed. All ESTs were trimmed back to their high-quality stop position as indicated in each dbEST entry. Any EST less than 150 bases was removed from consideration. This left 111,189 high quality 3′ ESTs for the remaining analyses. Bases 40–140 of each EST [40 bases upstream of the poly(T) tract (approximately bases 90–190, the highest quality portion of the original trace)] were then used in comparisons with the full length of all other ESTs in the set using BLASTN2 and the following parameters: S = 170, gapS2 = 150, M = 5, N = −11, Q = 11, R = 11. Each EST that showed greater than 95% identity on the correct strand, covered at least bases 50–130 of the 101 base query sequence, and the similarity of which ended before base 200 of the subject sequence was considered for pairwise comparisons. Pairwise comparisons of the full length of these 3′ ESTs were performed using cross_match (P. Green, unpubl.) with the following parameters: -minmatch 12, -minscore 80, -penalty -2, -gap_init -4, -gap_ext -4.

After the full-length Smith-Waterman pairwise comparisons, ESTs showing less than 8% mismatch across the full length of the EST were considered identical. The criterion of 8% mismatch was chosen empirically to reflect the increased error associated with the full length of the ESTs included in the alignments and because the first step in the analysis (the 3′ EST anchoring by the initial BLAST comparisons) was considered sufficiently stringent to exclude matches between related but nonidentical ESTs. ESTs were then grouped into clusters based on their pairwise similarities. The ESTs that could not be grouped into clusters were used to calculate the proportion of singletons contributed by each library to the entire data set. To examine the depletion of the singletons within each library, only intralibrary similarities were considered. The number of singletons as a fraction of the number of ESTs sampled in that library was calculated at intervals of 500 sequences as described above.

Internal Priming

To obtain an estimate of the frequency of internal priming per library, the results of the comparison between our masked EST data set and the nonredundant human mRNA data base were used. All ESTs showing similarity to an mRNA sequence with a P value of 10^{-100} or better were examined. The 3′ ESTs showing similarities at these levels on the "correct strand" were classified as follows: (1) 3′ EST whose similarity fell within 50 bases of the end of the human mRNA sequence; (2) 3′ EST whose similarity ended before 50 bases from the end of the human mRNA, but another EST showed similarity in that region; and (3) 3′ EST whose similarity fell more than 50 bases from the end of the human mRNA, and no other EST showed similarity in that region. Only those meeting the latter criterion were judged to be a candidate nonspecifically primed cDNAs.

Estimating the Frequency of Chimeric Clones and the Frequency of Inverted Clones

From the results of the comparison between our masked EST data set and the nonredundant human mRNA data base, an estimate of the frequency of chimeric clones was obtained. All 3′ ESTs showing similarity to a human mRNA sequence at a probability of 10^{-100} or better were considered for this analysis. If the 3′ EST matched and the corresponding 5′ EST had not been successful (failed during initial sequencing), the match was not considered. If both the 5′ and 3′ ESTs matched the same human mRNA sequence at the 10^{-100} cutoff the cDNA clone was considered unlikely to be a chimera. Conversely, if the 3′ EST matched a human mRNA and the 5′ EST was submitted to dbEST but failed to match the same human mRNA, the cDNA from which the ESTs were derived was considered potentially chimeric. In a similar fashion, using the results of the comparison between our masked EST data set and the nonredundant human mRNA data base, we generated an estimate of the frequency of reversed (or flipped) clones. The BLASTN2 alignments of all 3′ ESTs matching a human mRNA sequence with a P value of 10^{-100} or better were examined. Matches categorized as being on the minus strand were considered to identify cDNAs in the correct orientation; conversely, matches on the plus (or coding) strand were considered to identify cDNAs that were in inverted orientation.

Comparisons with Human Genomic Sequence

The 176 human genomic sequences larger than 25 kb as of April 7, 1996 (a total of 8,244,729 bases) were retrieved from GenBank release 94. These were compared with our masked ESTs using BLASTN2 and the following parameters: S = 170, gapS2 = 150, M = 5, N = −11, Q = 11, R = 11. All ESTs showing similarity to the genomic sequence with a P value of 10^{-100} or better were categorized as described in Results.

Comparisons with Protein Families

The 171 available seed alignments for protein families were obtained (Eddy 1995; Sonnhammer et al. 1996) and HMMs were built using the hmmer-1.9j package with the

following building parameters: hmmb -d -R -PBLOSUM62. The -d option allows for maximum discrimination; the -R option allows that some family members may be fragmentary (useful for EST data); and BLOSUM62 (Henikoff and Henikoff 1992) is the scoring matrix used. The program HMMFS was then used to search the resulting HMMs against the entire set of 5′ ESTs, translated in all six frames using the program "orfer" (S. Eddy, unpubl.). A cutoff score of 20.0 bits was used to determine whether a given clone belonged to a given family. This is a log-odds score; a score of 20 indicates that a sequence is 2^{20}-fold more likely to match the HMM than not. The use of HMMs in general and the hmmer-1.9j package specifically is detailed in a users' guide available upon request from Sean Eddy (eddy@genetics.wustl.edu).

Comparison of ESTs with a Yeast Protein Data Base

A nonredundant yeast protein data base (yeast_nrpep. fasta) containing 8269 yeast proteins was obtained from www-genome.stanford.edu on June 11, 1996. This data base contains yeast proteins culled from SWISS-PROT, GenPept, and PIR, with redundant entries (only 100% identical amino acid sequences) removed. Each of the yeast proteins was compared with the entire set of masked 5′ and 3′ ESTs using TBLASTN and the following parameters: M = PAM120, T = 17, W = 4, V = 10000, filter = seg. The number of yeast proteins with similarities to ESTs was then calculated for decreasing P values commencing at 10^{-9}.

ACKNOWLEDGMENTS

The authors wish to thank Drs. A. Williamson, K. Elliston, and J. Aaronson for support in all aspects of the project, Drs. D. Lipman, M. Boguski, C. Tolstoshev, G. Schuler, E. Koonin, and D. Bassett at the NCBI for assistance in various aspects of data analysis and submission, Drs. F. Collins, D. Smith, and M. Vaudin for encouragement and support, and Dr. Sean Eddy for advice on HMMs and their uses. We are grateful to Melissa Allen, Donald Blair, Louise Bowles, Michael Holman, Michele Steptoe, Brenda Theising, and Todd Wylie for technical assistance and valuable discussion. We gratefully acknowledge the support of all staff at the Washington University Genome Sequencing Center who have contributed to this effort. The work of G.L. and C.P. was supported by the U.S. Department of Energy under contract W-7405-Eng-43. All cDNA clones are available on a cost-per-clone basis from Genome Systems (St. Louis, MO), Research Genetics (Huntsville, AL), American Type Culture Collection (Rockville, MD), United Kingdom Human Genome Mapping Project Resource Center (Hinxton, UK), and Reference Library Data Base (Berlin). Queries concerning the IMAGE consortium can be directed to info image.llnl.gov. Sequencing traces and other EST data are available at the WashU-Merck WWW site (http:// genome.wustl.edu/est/esthmpg.html). The sequence data described in this paper have been submitted to the GenBank data library.

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

REFERENCES

Aaronson, J.S., B. Eckman, R.A. Blevins, J.A. Borkowski, J. Myerson, S. Imran, and K.O. Elliston. 1996. Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST seqence data. *Genome Res.* (this issue).

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno et al. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

Adams, M.D., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White, et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–17.

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

Bairoch, A. and B. Boeckmann. 1994. The SWISS-PROT protein sequence data bank: Current status. *Nucleic Acids Res.* **22:** 3578–3580.

Bassett, D.E., M. Boguski, and P. Hieter. 1996. Yeast genes and human disease. *Nature* **379:** 589–590.

Boguski, M. and G. Schuler. 1995. Establishing a human transcript map. *Nature Genet.* **10:** 369–371.

Bonaldo, M.F., G. Lennon, and M.R. Soares. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* (this issue).

Dear S. and R. Staden. 1991. A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.* **19:** 3907–3911.

Dujon, B. 1996. The yeast genome project: What did we learn? *Trends Genet.* **12:** 263–169.

Eddy, S. 1995. Multiple alignment using hidden Markov models. ISMB-95: Proceedings Third International Conference on Intelligent Systems for Molecular Biology. pp. 114–120. AAAI Press, Menlo Park, CA.

Eddy, S., G. Mitchison, and R. Durbin. 1995. Maximum discrimination hidden Markov models of sequence consensus. *J. Computat. Biol.* **2:** 9–23.

Fulton, L.L. and R.K. Wilson. 16994. Variations on cycle sequencing. *Bio-Techniques* **17:** 298–301.

Hawthorn L.A. and J.K. Cowell. 1996. Regional assignment of EST sequences on human chromosome 13. *Cytogenet. Cell Genet.* **72:** 72–77.

Henikoff S. and J.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Houlgatte, R., R. Mariage-Samson, S. Duprat, A. Tessier, S. Bentolila, B. Lamy, and C. Auffray. 1995. The Genexpress Index: A resource for gene discovery and the genic map of the human genome. *Genome Res.* **5:** 272–304.

Johnston, M. 1996. The complete code for a eukaryotic cell. *Curr. Biol.* **6:** 500–503.

Khan, A.S., A.S. Wilcox, and M.H. Polymeropoulos. 1992. Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genet.* **2:** 180–185.

Krogh, A., M. Brown, I.S. Mian, K. Sjoelander, and D. Haussler. 1994. Hidden Markov model in computational biology. Applications to protein modelling. *J. Mol. Biol.* **235:** 1501–1531.

Lennon, G.G., C. Auffray, M. Polymeropoulos, and M.B. Soares. 1996. The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. *Genomics* **33:** 151–152.

McCombie, W.R., M.D. Adams, J.M. Kelley, M.G. FitzGerald, T.R. Utterback, M. Khan, M. Dubnick, A.R. Kerlavage, J.C. Venter, and C. Fields. 1992. Caenorhabditis elegans expressed sequence tags identify gene families and potential disease gene homologues. *Nature Genet.* **1:** 124–131.

McGuire, R.E., S.A. Jordon, V.V. Braden, G.G. Bouffard, P. Humphries, E.D. Green, and S.P. Daiger. 1996. Mapping the RP10 locus for autosomal dominant retinitis pigmentosa on 7q: Refined genetic positioning and localization within a well-defined YAC contig. *Genome Res.* **6:** 255–266.

Matsubara, K. and K. Okubo. 1993. cDNA analysis in the human genome project. *Gene* **135:** 265–274.

Milner, R.J. and J.G. Sutcliffe. 1983. Gene expression in rat brain. *Nucleic Acids Res.* **11:** 5497–54520.

Newman, T. 1994. Genes galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. *Plant Physiol.* **106:** 1241–1255.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Putney, S.D., W.C. Herlihy, and P. Schimmel. 1983. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302:** 718–721.

Robertson, N.G., U. Khetarpal, G.A. Gutierrez-Espeleta, F.R. Bieber, and C.C. Morton. 1994. Isolation of novel and known genes from a human fetal cochlear cDNA library using subtractive hybridization and differential screening. *Genomics* **23:** 42–50.

Saiki, R.K., D.H. Gelfand, S. Stoffel, S.J. Scharf, R. Higuchi, G.T. Horn, K.B. Mullis, and H.A. Erlich. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239:** 487–491.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: A laboratory manual*, 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanger, F., A.R. Coulson, B.G. Barrell, A.J.H. Smith, and B.A. Roe. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74:** 5463–5467.

Sasaki, T., J. Song, Y. Koga-Ban, E. Matsui, F. Fang, H. Higo, H. Nagasaki, M. Hori, M. Miya, E. Murayama-Kayano, et al. 1994. Toward cataloguing all rice genes: Large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. *Plant J.* **6:** 615–624.

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

Sonnhammer, E.L.L. and R. Durbin. 1994. A workbench for large scale sequence homology analysis. *Computer Applic. Biosci.* **10:** 301–307.

Sonnhammer, E.L.L., S. Eddy, and R. Durbin. 1996. Pfam: A comprehensive data base of protein domain families based on seed alignments. *Proteins* (in press).

Sulston J., Z. Du, K. Thomas, R. Wilson, L. Hillier, R. Staden, N. Halloran, P. Green, J. Thierry-Mieg, L. Qiu, S. Dear, A. Coulson, M. Craxton, R. Durbin, M. Berks, M. Metzstein, T. Hawkins, R. Ainscough, and R. Waterston. 1992. The C. elegans sequencing project: A Beginning. *Nature* **356:** 37–41.

Tugendreich, S., M.S. Boguski, M.S. Seldin, and P. Hieter. 1993. Linking yeast genetics to mammalian genomes: Identification and mapping of the human homolog of CDC27 via the expressed sequence tag (EST) data base. *Proc. Natl. Acad. Sci.* **90:** 10031–10035.

Waterston, R., C. Martin, M. Craxton, C. Huynh, A. Coulson, L. Hillier, R.K. Durbin, P. Green, R. Shownkeen, N. Halloran, T. Hawkins, R. Wilson, M. Berks, Z. Du, K. Thomas, J. Thierry-Mieg, and J. Sulston.

1992. A survey of expressed genes in Caenorhabditis elegans. *Nature Genet.* **1:** 114–123.

Wilcox, A.S., A.S. Khan, J.A. Hopkins, and J.M. Sikela. 1991. Use of 3′ untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: Implications for an expression map of the genome. *Nucleic Acids Res.* **19:** 1837–1843.

Wilson, R., R. Ainscough, K. Anderson, C. Baynes, M. Berks, J. Bonfield, J. Burton, M. Connell, T. Copsey, J. Cooper, et al. 1994. The C. elegans genome project: Contiguous nucleotide sequence of over two megabases from chromosome III. *Nature* **368:** 32–38.

Received June 25, 1996; accepted in revised form July 29, 1996.