RESEARCH

Toward the Development of a Gene Index to the Human Genome: An Assessment of the Nature of High-throughput EST Sequence Data

Jeffrey S. Aaronson,[1] Barbara Eckman, Richard A. Blevins, Joseph A. Borkowski, Joseph Myerson, Shahid Imran, and Keith O. Elliston[2]

Merck Research Laboratories, Department of Bioinformatics, Rahway, New Jersey 07065, and West Point, Pennsylvania 19486

A rigorous analysis of the Merck-sponsored EST data with respect to known gene sequences increases the utility of the data set and helps refine methods for building a gene index. A highly curated human transcript data base was used as a reference data set of known genes. A detailed analysis of EST sequences derived from known genes was performed to assess the accuracy of EST sequence annotation. The EST data was screened to remove low-quality and low-complexity sequences. A set of high-quality ESTs similar to the transcript data base was identified using BLAST; this subset of ESTs was compared with the set of known genes using the Smith–Waterman algorithm. Error rates of several types were assessed based on a flexible match criterion defining sequence identity. The rate of lane-tracking errors is very low, ~0.5%. Insert size data is accurate within ~20%. Reversed clone and internal priming error rates are ~5% and 2.5%, respectively, contributing to the incorrect identification of reads as 3′ ends of genes. Follow-up investigation reveals that a significant number of clones, miscategorized as reversed, represent overlapping genes on the opposite strand of entries in the transcript data base. Relevance of these results to the creation of a high-quality index to the human genome capable of supporting diverse genomic investigations is discussed.

The random sequencing of cDNA clones has been used for more than a decade as a method for gene discovery (Costanzo et al. 1983). This technique more recently has been termed expressed sequence tag (EST) analysis, and has resulted in the partial characterization of a large variety of human genes (Adams et al. 1991, 1995; Wilcox et al. 1991; Khan et al. 1992; Okubo et al. 1992; Houlgatte et al. 1995), as well as those of other species (Waterson et al. 1992; Cooke et al. 1996). Significant effort has gone into EST sequencing; the majority of sequence entries found in the public data base GenBank (Benson et al. 1994) are the result of EST sequencing projects. However, aside from generalized estimates of numbers of genes represented and a quality-control analysis measuring levels of vector contamination and nonnuclear RNA (Adams et al. 1995; Boguski and Schuler 1995; Aaronson and Elliston 1996), little is known about the nature of the data resulting from the automated, single-pass sequencing of cDNA clones.

In 1994, a public effort to generate both 5′ and 3′ EST sequences from the majority of human genes was developed (Williamson et al. 1995). This collaboration, consisting of Merck & Co., Inc., the Integrated Molecular Analysis of Genomes and Their Expression (IMAGE) Consortium (Lennon et al. 1996), the Genome Sequencing Center (GSC) at Washington University of St. Louis School of Medicine, the National Center for Biotechnology Information (NCBI), and the Computational Biology and Informatics Laboratory (CBIL) at the University of Pennsylvania, has concentrated on the characterization of normalized libraries, and has produced more than 300,000 EST sequences from more than 180,000 IMAGE cDNA clones. These EST sequences com-

[1]Corresponding author.
E-MAIL aaronson@merck.com; FAX (908) 594-2929.
[2]Current address: Bayer Corporation, Genomics Section, West Haven, Connecticut 06516.

prise >75% of publicly available human EST sequences. These sequences, derived from 27 distinct cDNA libraries, have been placed into the public EST data base dbEST (Boguski et al. 1993), as well as GenBank. The processing of these EST sequences has included quality assessment, contamination removal, and initial automated annotation, prior to their submission to the data base (Hillier et al., this issue). This annotation, including clone_id, insert size, orientation, and endedness, provides important constraints governing the relationships between EST sequences and the clones from which they are derived, thereby greatly increasing the tractability of large-scale analyses of the data. A sufficient quantity of annotated EST data has been produced from the source libraries to assess both the accuracy of this annotation and the general nature of high-throughput EST data, with detailed statistics for each characterized library.

The characterization of sequence data that is generated from predominantly unknown transcription units is a difficult problem. However, the ESTs that are derived from known genes represented in public data bases can be readily studied. Analysis of human genes represented in GenBank has resulted in about 4000 characterized genes (UniGene, EGAD). Surveys of these genes have shown that >60% are represented currently by one or more EST sequences (Adams et al. 1995; Boguski and Schuler 1995; White and Kerlavage 1996). These results are consistent with this study, which finds coverage of known genes to be over 62% (data not shown). A thorough comparison and analysis of the set of annotated EST sequences that are derived from these known genes can begin to elucidate the general nature of ESTs representing all human genes. As known genes have been subject to closer study, they provide a standard with which less reliable EST data can be compared. Such a comparison is the subject of this study, with the immediate goal of enhancing the utility of the Merck-sponsored data for the genomics community. The ultimate goal is the development of an index to the human genome (Merck Gene Index) that classifies each sequenced clone by both gene and transcript. A high-quality Gene Index can facilitate studies of gene expression, gene function, and the assessment of their relevance to human disease.

EXPERIMENTAL DESIGN

The detailed sequence analysis of the EST data was accomplished through a multistep process. First, a high-quality human transcript data base was identified for use in comparisons to partial transcript sequences. This provides a reference data set that can be used to study the nature of the ESTs that match these known genes. The choice of a transcript data base was limited to: (1) a set of transcript sequences culled from GenBank using annotated CDS, mRNA, and prim_transcript features; (2) the UniGene (Boguski and Schuler 1995) data set developed at NCBI; and (3) EGAD (White and Kerlavage 1996), the Expressed Gene Anatomy Database developed at The Institute for Genomic Research. After careful consideration, EGAD was selected as the reference data set, because of its highly curated nature and the methods in which multiple transcripts for individual human genes are represented hierarchically. All known alternative splice forms are associated with each EGAD gene, and a sequence is given for each splice form. Next, the transcript data base was processed to remove artifacts, including vector sequences and redundant clone sequences. Searches were performed over the remaining transcripts in order to annotate coding region boundaries and locations of A-rich regions. In parallel, the EST data retrieved from GenBank was processed to remove low-quality sequence, using an "index quality" analysis. This analysis restricted the level of ambiguity allowed in sequences by truncating all sequences before they reach a level of 5% Ns, and removed all sequences less than 100 bp in length. In addition, all sequences containing repeats were removed, in order to eliminate spurious matches as a result of alignments of repeat sequences during sequence comparison. This resulted in a data set of 175,666 sequences (as of December 15, 1995). Searches were performed over this index-quality data set to annotate positions of polyA signals. An initial comparison of the processed EGAD data set against these sequences was performed using BLAST (Altschul et al. 1990). This reduced the data to a set of sequences with sequence similarity to the known genes described in EGAD. Both the EGAD data and index-quality EST data were uploaded into a relational data base. At this point, a full Smith–Waterman (Smith and Waterman 1981) comparison of the transcript data base and the BLAST-selected ESTs was performed, and the results were stored in a relational data base for further analysis. The results presented are the summaries of this analysis.

RESULTS

Preprocessing of Data Sets

The EGAD data base was analyzed for vector contamination and other artifacts prior to the full EGAD:EST analysis. One transcript, HT1138, was identified as containing vector and was removed from analysis. A screen for sequence redundancy yielded seven transcripts represented in both orientations. Each pair of transcripts was analyzed, and the appropriate entry removed (HT0464, HT1512, HT2417, HT2998, HT3667, HT4390, and HT4565). The final EGAD subset used for analysis represented 4412 transcripts (HTs) from 3970 genes (HGs).

In order to restrict the EST sequence analysis to high-quality portions of these single-pass sequences, several criteria were adopted to trim remaining poly(A) tails and screen low-quality sequence from the data set. To confine the analysis to the more reliable portions of each sequence read, each sequence was trimmed to contain only the high-quality portions of sequence, as annotated by GSC. Sequences of less than 100 bp were removed, with a reduction of 20,165 entries. To further refine the set of sequences for analysis to high-quality reads, the ESTs were trimmed to reduce ambiguous base calls on the trailing end of reads. This was accomplished by truncating each read before the rate of Ns reached 5%. A grace period of 99 bp was allowed before the criterion was enforced, with 5253 sequences of less than 100 bp eliminated. Finally, to remove repetitive elements from the analysis, a set of 23,162 ESTs annotated in GenBank as containing repetitive sequence elements were eliminated from the remaining ESTs. This entire process reduced a set of 224,246 ESTs to a set of 175,666 of index-quality ESTs for further characterization. The results of the index-quality assessment are summarized by library in Table 1. Each library is characterized for the total number of EST sequences culled from dbEST, with the number of sequences passing the index-quality assessment indicated. The wide variation in library quality reported here may be caused by a number of factors, including library construction, sequence specific effects (e.g., high or low expression of GC-rich transcripts, high or low incidence of repeat sequence) or variations in sequencing protocol and reagents. Quality values range from 66.05% index-quality ESTs for Soares pineal gland N3HPG to 86.65% for Soares infant brain 1NIB.

Selection of EGAD-specific ESTs

The detailed analysis of sequence similarities be-

Table 1. Results of Index Quality Screening

Library	ESTs	Index quality	Index quality (%)
Soares pineal gland N3HPG	1,461	965	66.05
Stratagene ovary	3,146	2,135	67.86
Stratagene lung (#937210)	8,172	5,741	70.25
Weizmann Olfactory Epith.	1,645	1,157	70.33
Soares adult brain N2b5HB	8,047	5,798	72.05
Soares breast 3NbHBst	9,530	6,903	72.43
Soares adult brain N2b4HB	4,406	3,260	73.99
Stratagene liver	8,417	6,245	74.20
Soares fetal liver spleen	70,463	52,649	74.72
Soares retina N2b5HR	2,303	1,734	75.29
Soares retina N2b4HR	4,043	3,045	75.32
Soares breast 2NbHBst	7,745	5,896	76.13
Stratagene placenta	3,640	2,790	76.65
Stratagene fetal spleen	6,263	4,979	79.50
Soares ovary tumor NbHOT	795	633	79.62
Soares melanocyte 2NbHM	2,850	2,281	80.04
Soares placenta 8 to 9 weeks	1,345	1,090	81.04
Soares placenta Nb2HP	33,754	28,320	83.90
Morton Fetal Cochlea	1,609	1,389	86.33
Soares infant brain 1NIB	44,611	38,656	86.65

tween known gene sequences and single-pass EST sequences requires the use of a rigorous pairwise alignment method, such as that of Smith and Waterman. However, the computational requirements of a full Smith–Waterman analysis of 4412 full-length transcript sequences against 175,666 ESTs were prohibitive. To reduce the matrix of comparison, a set of sequences similar to the transcript data base was selected by comparing each transcript with the refined EST data set using the rapid, but less sensitive, BLASTN program. In order to minimize false negatives, sequences meeting the modest criterion of $P \leq 10^{-30}$ were binned and subjected to the full Smith–Waterman comparison. A set of 42,668 ESTs was selected (5′, 58%; 3′, 42%), with an average processed size of 280 bp (5′, 281 bp; 3′, 279 bp), thus reducing the matrix of comparison by a factor of four, and reducing the calculation by an order of magnitude. The Smith–Waterman output was parsed into structured fields organized around individual searches and alignments, and the resulting tables were loaded into a relational data base.

Identity Criteria and Reliability

The detailed analysis of EST data that corresponds to known genes relies upon the definition of a sequence match. EST data by its nature is error prone, with the average sequence fidelity hovering around 97% (Nishikawa and Nagai 1996). Members of highly conserved multigene families have regions that surpass 95% identity, and often have regions of 99% or more identity (Efstratiadis et al. 1980). Thus, the choice of criteria for the determination of a match is important. Rather than arbitrarily choosing a single match criterion, these experiments were performed using three different criteria based upon the full-length of a Smith–Waterman optimal alignment: 95% identity over at least 100 bp, 97% identity over at least 100 bp, and 90% identity over at least 200 bp. A rigorous definition of sequence match, while deserving of further study, is beyond the scope of this paper. The chosen criteria accommodate the observed sequence error in single-pass data, while providing stringent criteria for match determination. Evaluating the Smith–Waterman comparisons between EGAD and the 42,668 BLAST-selected ESTs with similarity to EGAD yielded sets of 36,593, 32,413, and 33,285 ESTs with identity to EGAD, according to each of the described criteria (100/95, 100/97, and 200/90).

Sequence Analysis

Reversed Clones/Mislabeled Ends

The orientation of the inserts in cDNA cloning is determined by the biochemistry of the reverse-transcription and cloning reactions. The libraries involved in this project were derived from oligo(dT) primed cDNA that was cloned in a unidirectional fashion (Soares et al. 1994). The sequencing was performed using distinct forward and reverse primers to generate 5′ and 3′ sequence, respectively. The association between primer and sequence must be maintained strictly in order to record the correct orientation of the insert. A failure in this process will result in the mislabeling of 5′ sequence as 3′, and 3′ sequence as 5′. Similarly, the cloning of the insert in the opposite orientation also will lead to the incorrect association of orientation and sequence, and therefore incorrect annotation of sequence. Since raw sequence is trimmed prior to GenBank submission, vector and poly(T) tracks that might distinguish 3′ from 5′ ESTs are not present in the public data base. Our methodology does not discriminate between these two instances, hereafter referred to as reversed clones.

The orientation of the ESTs is encoded by the GSC in several ways in the public data bases. The primary method uses the s/r primer notation in the suffix of the read_id to encode 3′/5′. The secondary method encodes orientation using unstructured "3′" or "5′" labels in dbEST and GenBank flat-file entries. These two criteria should be in agreement for all EST sequences. However, 0.39% of entries studied (186 of the 46,668 BLAST-selected ESTs) have conflicting labels, indicating that errors occurred during annotation. Sequences are represented in the orientation that they are read from the sequencing gel; therefore, true 3′-end sequences can be detected by identifying sequence matches to the noncoding, or − strand. Conversely, true 5′- end sequences can be detected by identifying sequence matches to the coding, or + strand. Thus, true 3′ reads should match transcript sequences on the − strand, and true 5′ reads should match transcript sequences on the + strand. A reversed clone is characterized by a true 5′ sequence that is labeled as 3′, and a true 3′ sequence that is labeled as 5′ (Fig. 1A).

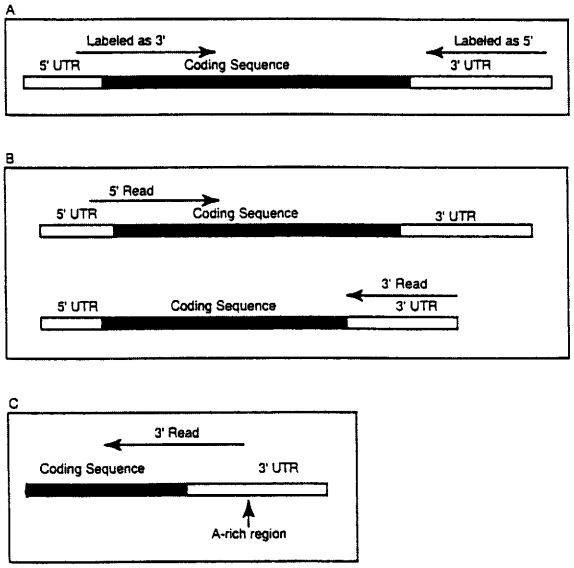

Figure 1 Idealized models for the analyses of EST sequences vs. EGAD transcripts are presented. (*A*) An example of a reversed clone. The labeled 3′ EST matches the transcript on the (+) strand, and the labeled 5′ EST matches the transcript on the (−) strand. (*B*) An example of a lane-tracking error. The 5′ read and 3′ read of a single clone do not match the same gene. (*C*) An example of an internal priming event. An A-rich region 20 bp upstream of the 3′ end of the EST is the source of internal priming during reverse transcription stage of library construction, resulting in a 3′ EST that is not anchored to the poly(A) tail.

The reversed clone analysis partitioned the set of EST clones that match known genes into three disjoint classes: (1) clone orientation agrees with EGAD (i.e., all ESTs from a clone that match EGAD agree in orientation); (2) clone orientation disagrees with EGAD (i.e., all ESTs from a clone that match EGAD disagree in orientation); and (3) clone orientation is mixed with respect to EGAD (i.e., the set of ESTs from a clone represent both orientations). The read_id annotation was found to be more reliable and was used to specify clone orientation for this analysis. Summary totals are given in Table 2A for the 3 match criteria. Over 5% of the EST clones are reversed in orientation with respect to EGAD, and only a few clones are in mixed orientation. Reversed clone rates for the 100/97 match criterion, organized by library, are given in Table 2B. The rates of reversed clones vary widely across libraries, from roughly 0.5% for a non-normalized fetal spleen library to 18% for normalized retina libraries. In general, the normalized (Soares) libraries have a higher rate of reversed clones than the non-normalized libraries.

Lane Tracking/Chimeric Clones

A main concern in large-scale EST sequencing projects is maintaining the correct association of sequence to clone. To maintain this association, it is essential to correctly assign and track the individual lanes on the sequencing gels. Lane-tracking errors introduce incorrect associations between sequence and clone. These can be identified more readily if there are multiple sequences associated with the same clone. Since both 5′ and 3′ sequences have been determined for the majority of clones in this project, we can readily identify lane-tracking errors by first selecting a set of clones for which both 5′ and 3′ sequences are represented, and then identifying the subset where the 5′ sequence matches different genes than the 3′ sequence. This method cannot distinguish between a lane tracking error and a chimeric clone; thus, they are not differentiated in this analysis. The resequencing, or full-length sequencing, of the affected clones could differentiate the rate of chimerism versus lane-tracking errors.

A set of clones was identified where both the 5′ and 3′ reads matched a known transcript. Occasionally, a clone matched more than one transcript, as a result of either representation of multiple transcripts for the same gene or the presence of members of multigene families whose sequences are very highly conserved. This complication was resolved by comparing the sets of genes matching a clone's 5′ read and 3′ read, respectively, according to the three distinct match criteria, and then identifying those that had no gene in common (Fig. 1B). Summary statistics for each of the three match criteria are given in Table 3A, and a listing by library is given in Table 3B. The frequency of clones that exhibit lane-tracking errors in the sample set is ~1%. Assuming that only one of these reads is incorrectly assigned, the rate of errors per sequence is approximately one-half the rate per clone, or ~0.5%.

Insert Size

An important component of the Merck-sponsored EST project is the estimation of insert sizes for each clone sequenced. Several tech-

Table 2. Comparison of ESTs

	Clones	Multiple EST clones	Same orientation	% Same	Reverse orientation	% Reversed	Mixed orientation	% Mixed
Criteria								
100/95[a]	25,919	10,647	24,546	94.70	1366	5.27	7	0.03
100/97[b]	23,924	8,483	22,666	94.74	1253	5.24	5	0.02
200/90[c]	24,309	8,969	23,051	94.82	1252	5.15	6	0.02
Library[d]								
Stratagene fetal spleen	925	323	920	99.46	5	0.54	0	0.00
Stratagene liver	2,092	697	2,075	99.19	16	0.76	1	0.05
Stratagene placenta	729	286	719	98.63	10	1.37	0	0.00
Soares infant brain 1NIB	3,340	1,302	3,279	98.17	57	1.71	4	0.12
Morton Fetal Cochlea	219	117	214	97.72	5	2.28	0	0.00
Soares fetal liver spleen	6,441	2,276	6,140	95.33	301	4.67	0	0.00
Weizmann Olfactory Epith.	193	96	183	94.82	10	5.18	0	0.00
Stratagene lung	971	374	918	94.54	53	5.46	0	0.00
Soares ovary tumor NbHOT	128	33	121	94.53	7	5.47	0	0.00
Soares melanocyte 2NbHM	222	77	207	93.24	15	6.76	0	0.00
Soares placenta 8 to 9 weeks	158	40	147	93.04	11	6.96	0	0.00
Soares breast 3NbHBst	1,317	303	1,223	92.86	94	7.14	0	0.00
Soares placenta Nb2Hp	3,956	1,637	3,643	92.09	313	7.91	0	0.00
Stratagene ovary	555	233	511	92.07	44	7.93	0	0.00
Soares adult brain N2b4HB	415	79	373	89.88	42	10.12	0	0.00
Soares adult brain N2b5HB	820	241	736	89.76	84	10.24	0	0.00
Soares breast 2NbHBst	873	235	783	89.69	90	10.31	0	0.00
Soares pineal gland N3HPG	92	24	82	89.13	10	10.87	0	0.00
Soares retina N2b5HR	167	43	137	82.04	30	17.96	0	0.00
Soares retina N2b4HR	311	67	255	81.99	56	18.01	0	0.00

Analysis of ESTs for reversed or mislabeled clones was done by comparing EST clones with the EGAD transcript data base. The number of EST clones matching EGAD and the number of EST clones with more than one EST matching EGAD are listed. The number and percentage of EST clones that are in the same orientation, reverse orientation, and mixed orientation with respect to EGAD are shown.

[a]At least 95% identity over 100 bp.

[b]At least 97% identity over 100 bp.

[c]At least 90% identity over 200 bp.

[d]Results are shown for the 100/97 match criterion.

Table 3. Analysis of Lane-tracking Errors

	Clones	Clones with error	% Clones with error
Criteria[a]			
100/95	10,635	101	0.95
100/97	8,474	78	0.92
200/90	8,959	92	1.03
Library[b]			
Soares adult brain N2b5HB	241	0	0.00
Morton Fetal Cochlea	117	0	0.00
Weizmann Olfactory Epith.	96	0	0.00
Soares melanocyte 2NbHM	77	0	0.00
Soares retina N2b5HR	43	0	0.00
Soares ovary tumor NbHOT	33	0	0.00
Soares pineal gland N3HPG	24	0	0.00
Stratagene placenta	286	1	0.35
Soares fetal liver spleen	2,276	12	0.53
Soares placenta Nb2HP	1,636	12	0.73
Soares breast 3NbHBst	303	3	0.99
Stratagene lung	374	4	1.07
Soares breast 2NbHBst	235	3	1.28
Soares adult brain N2b4HB	78	1	1.28
Stratagene ovary	233	3	1.29
Soares infant brain 1NIB	1,296	17	1.31
Soares retina N2b4HR	67	1	1.49
Stratagene fetal spleen	323	5	1.55
Stratagene liver	696	15	2.16
Soares placenta 8 to 9 weeks	40	1	2.50

Number of clones having both a 3′ and 5′ read matching EGAD are shown, as are the number and percentage of those clones showing a lane-tracking error.
[a]Criteria as in Table 2.
[b]Results are shown for the 100/97 match criterion.

niques have been used to estimate insert size, including PCR and restriction digestion (Hillier et al., this issue). The insert length for a clone can be calculated if both a 3′ read and a 5′ read are available and match a known transcript sequence. The alignments determine the extent of the transcript corresponding to the clone, and from this, the true insert size of the clone can be calculated. This method is complicated by the presence of alternative splice forms of genes, as well as by closely related members of gene families. Associating a clone with a related but distinct transcript will invalidate the result. Thus, clones whose sequences matched more than one transcript were excluded. Clones whose sequences appeared to be inappropriately trimmed were also removed. The possibility of underestimating the actual insert size was investigated by limiting the analysis to ESTs whose leading (5′) end was included in the EGAD alignment. Introducing this constraint did not affect the results significantly, and, therefore, it was not enforced in the final analysis. Summary statistics of insert sizes for each of the three match criteria are given in Table 4A. The average insert size of sequenced clones is ~0.9 kb; the average error in insert size determination is about 21.5%, where error refers to the difference between calculated and reported insert size, expressed as a percentage of the calculated size. Results organized by library are shown in Table 4B, according to the 100/97 match criterion.

Internal Priming/Alternative Termination

The development of a nonredundant set of clones representing all of the transcripts characterized in the EST project relies upon analysis of comparable regions of 3′ untranslated region

Table 4. Results of Insert-size Analysis

	Clones	Average reported size	Average calculated size	Minimum difference (%)	Maximum difference (%)	Average difference (%)
Criteria[a]						
100/95	7664	979	890	0.00	523.25	21.78
100/97	6197	988	901	0.00	500.54	21.41
200/90	6296	971	886	0.00	523.25	21.37
Library[b]						
Soares retina N2b4HR	11	1666	1476	1.82	22.81	15.22
Soares adult brain N2b4HB	62	1469	1321	2.49	107.71	18.18
Soares infant brain 1NIB	1028	1605	1439	0.07	307.91	18.71
Soares breast 3NbHBst	266	876	772	0.36	128.97	19.01
Soares breast 2NbHBst	184	868	760	0.28	177.86	19.11
Stratagene fetal spleen	245	803	733	0.09	135.61	19.42
Soares fetal liver spleen	1844	959	853	0.00	500.54	19.92
Soares placenta Nb2HP	1404	843	753	0.00	444.82	19.98
Soares adult brain N2b5HB	210	780	675	0.77	172.29	20.36
Soares retina N2b5HR	7	867	738	4.84	75.67	22.94
Soares pineal gland N3HPG	14	940	900	2.85	57.28	23.02
Weizmann Olfactory Epith.	43	834	669	1.34	138.40	24.09
Soares ovary tumor NbHOT	21	894	722	2.77	99.20	25.20
Stratagene placenta	188	791	748	0.09	305.96	27.77
Stratagene lung	224	648	689	0.43	113.32	30.06
Stratagene ovary	129	512	458	0.78	379.53	30.74
Stratagene liver	317	703	931	0.22	139.68	37.24

Reported and calculated insert sizes of EST clones are shown as measured by alignment to EGAD.
[a]Criteria as in Table 2.
[b]Results are shown for the 100/97 match criterion.

(UTR) sequence. This analysis is fundamental to the initial development of the Merck Gene Index, as it provides a classification of expressed sequence into transcript equivalence-classes. The 3′ UTR region is the most diverse region of transcripts (Ko et al. 1994), and each 3′ UTR fragment can serve as a unique identifier for its source transcript. All 3′ UTR fragments resulting from 3′ reads are anchored at the poly(A) site, as they are primed by an oligo(dT) primer at the poly(A) tail and therefore are directly comparable. A potential problem with this approach is the occurrence of false 3′ ends because of internal priming. This can occur during library construction as a result of priming from A-rich regions upstream of the poly(A) tail during the reverse transcription process, which results in 3′ ends that are not anchored to the poly(A) site (Fig. 1C).

To assess the level of internal priming, the set of 3′ ESTs matching known transcripts were analyzed to determine whether or not they fall at the 3′ end of transcripts. ESTs that fall close to the end of a transcript are considered to have been primed correctly; ESTs that only fall upstream of the poly(A) site may result from either an alternative 3′ end or an internal priming event. Since internal priming should occur from a region similar to the oligo(dT) primer, an attempt was made to distinguish these events by looking for canonical signals characteristic of 3′ ends, and A-rich regions that can serve as a template for oligo(dT)-primed cDNA synthesis. ESTs that do not fall at the 3′ end are classified as: (1) internal ESTs with canonical poly(A) signals, suggestive of an alternative 3′ end; (2) internal ESTs that lack canonical poly(A) signals and are proximal to an A-rich region, considered likely to be the result of internal priming; or (3) internal ESTs that lack canonical poly(A) signals that are not proximal to an A-rich region. These ESTs may be derived

from true 3′ ends that lack a canonical poly(A) signal or may be the result of an internal priming event not captured by the A-rich criteria defined in this analysis. Therefore, no judgment was made as to the import of this latter class. A threshold of 20 bp was used for determining whether an EST was proximal to the end of a transcript. Two values, 20 bp and 50 bp, in turn, were used for determining whether an EST was positioned appropriately with respect to an A-rich region. Summary statistics for the three match criteria are given in Table 5A. The overall level of internal priming as measured by these criteria is relatively low, within the range of 2–3%. The rate of internal priming varies considerably by library: Several libraries have rates greater than 3% (Table 5B).

DISCUSSION

A gene index is ideally a collection of information about genes in which all the information pertaining to a particular gene is organized into a single gene class, and each gene class is distinct from all other gene classes. The partial and error-prone nature of EST data complicates the definition and formulation of the set of gene classes that form an index. Moreover, there are many reasonable characterizations of a gene index, driven by different intended uses of the information. The Merck Gene Index is intended to be an informatics resource for the community, designed to facilitate diverse and comprehensive investigations into the nature of the human genome, including studies of gene mapping, gene expression, and gene function. Such disparate endeavors require a flexible index that can generate a varied set of reports designed to address the specific needs of each investigation, reflecting the particular scientific assumptions and goals of individual researchers. This can be accomplished by creating a data base that one can query that represents explicitly the relationships between EST sequences and other relevant data, and a sophisticated indexing methodology capable of producing targeted reports satisfying the criteria specified in the query.

This study has made significant steps toward the development of a dynamic gene index in several ways. First, levels of accuracy have been established for important annotation to the majority of public human EST sequences, thereby enabling individuals to make informed decisions on the extent to which they will rely upon the annotations. A background rate of biological and informatics errors is unavoidable in large-scale sequencing projects. Moreover, the error rates will naturally change in response to applications of different technology, improvements in methodology, and an evolving understanding of the underlying biology. To perform effectively in this noisy environment, it is imperative that confidence levels be established and analytical methodology be developed that can withstand violating the constraints imposed by the annotations.

Second, the development of a flexible data repository that will be the foundation of the Merck Gene Index has begun. A relational representation of annotated EST sequence data, annotated transcript data, and sequence similarity search results provides the basis for these analyses. Parameterized queries over this data are employed, granting flexibility in choosing criteria governing aspects of these analyses. A flexible methodology is crucial in order to (1) provide robustness and maintain relevance in the emerging field of genomic science, and (2) enable basic research into the nature of the genome, by providing a framework in which hypotheses can be tested and poorly understood or unknown biological phenomena can be explored.

Third, a multilevel sequence comparison strategy has been adopted, utilizing the BLAST algorithm to rapidly discriminate sequences that are clearly too dissimilar to be identical, and employing the rigorous Smith–Waterman algorithm to distinguish similarity from identity. The initial development of the Gene Index depends upon binning EST clones into transcript equivalence classes, based upon sequence comparisons of error-prone single-pass EST data. By using BLAST to quickly screen out dissimilar sequences, a reduced set of comparisons can be performed via the computationally intensive Smith–Waterman algorithm, focusing greater sensitivity on the difficult problem of discerning identity from similarity using EST data.

The detailed analysis of EST data begins to provide feedback about the accuracy of annotations attached to EST sequence. Specifically, this analysis has evaluated clone_id, insert size, orientation, and endedness. The correct association of 5′ reads with corresponding 3′ reads using the GSC clone_id has been assessed through lane-tracking. The rate of sequences exhibiting lane-tracking errors is ~0.5%, indicating that the mapping from sequence to clone is of high fidelity. This has two important implications. First, physi-

Table 5. Internal Priming Analysis Results

		(−) Strand ESTs[b]	Internal to transcript[c]	% Internal	Internal with poly(A)[d]	% Internal with poly(A)	Internal without poly(A)	Internal priming[e]	% Internal priming
Criteria[a]									
100/95		15,558	1815	11.67	723	39.83	1092		
	20							339	2.18
	50							406	2.61
100/97		13,324	1682	12.62	665	39.54	1017		
	20							333	2.50
	50							399	2.99
200/90		13,792	1785	12.94	718	40.22	1067		
	20							337	2.44
	50							404	2.93
Library[f]									
Soares melanocyte 2NbHM		114	16	14.04	3	18.75	13	0	0.00
Soares placenta 8 to 9 weeks		78	9	11.54	2	22.22	7	0	0.00
Stratagene ovary		367	22	5.99	10	45.45	12	2	0.54
Stratagene lung		579	61	10.54	31	50.82	30	4	0.69
Soares adult brain N2b5HB		412	67	16.26	24	35.82	43	5	1.21
Soares breast 3NbHBst		574	55	9.58	14	25.45	41	7	1.22
Soares retina N2b4HR		159	39	24.53	6	15.38	33	2	1.26
Weizmann Olfactory Epith.		132	17	12.88	5	29.41	12	2	1.52
Soares placenta Nb2HP		2,178	236	10.84	90	38.14	146	37	1.70
Soares breast 2NbHBst		406	48	11.82	9	18.75	39	8	1.97
Soares adult brain N2b4HB		184	18	9.78	4	22.22	14	4	2.17
Stratagene placenta		435	44	10.11	21	47.73	23	10	2.30
Stratagene fetal spleen		624	87	13.94	65	74.71	22	15	2.40
Soares infant brain 1NIB		1,995	244	12.23	128	52.46	116	50	2.51
Soares retina N2b5HR		73	16	21.92	4	25.00	12	2	2.74
Soares ovary tumor NbHOT		71	9	12.68	3	33.33	6	2	2.82
Soares fetal liver spleen		3,607	444	12.31	166	37.39	278	128	3.55
Morton Fetal Cochlea		128	24	18.75	10	41.67	14	5	3.91
Stratagene liver		1,165	213	18.28	70	32.86	143	47	4.03
Soares pineal gland N3HPG		43	13	30.23	0	0.00	13	3	6.98

[a]The summary lists the number of ESTs that are within 20 or 50 bp of an appropriate A-rich region, with results for each of the three match criteria implemented. Criteria as in Table 2.
[b]Number of true 3' ESTs (sequences that match the − strand of transcripts).
[c]3′ ESTs that do not align within 20 bp upstream of the reported 3′ end of transcripts.
[d]Internal ESTs that contain a canonical poly(A) signal.
[e]Internal ESTs that do not contain a poly(A) signal and are proximal to an A-rich region that could act as a template for priming by an oligo(dT) primer.
[f] Results are shown for the 20 bp and 100/97 match criteria.

cal clones identified through sequence similarity can be reliably retrieved. Second, all sequences annotated with the same clone_id are derived from the same transcript. This is of great utility for the construction of an index to the genome, which exploits sequence identity to categorize all clones representing a single transcript (or gene) into a single index class.

Accurate insert size data for EST clones enables the selection of the longest clone from each index class, and is a valuable aid to performing sequence assemblies within a class. However, the insert size analysis has shown this data to be accurate only within ~20%—the average error rate of reported insert sizes is about 21% overall, and exceeds 15% for each library. This complicates the use of insert size data in building contigs and the development of index classes. However, because the average error over most of the clones falls in the 15–20% range, a systematic, and therefore correctable, error during insert size determination may be responsible. Differing techniques used to determine insert size may account for the differences observed among libraries, as may the differential presence of alternatively spliced transcripts in various tissues.

Reversed clones and internal priming events each contribute to the incorrect identification of sequences as 3′ ends of genes, and to the reduction in fidelity of an index that relies upon annotated 3′ ends to categorize clones. This study estimates the level of reversed clones in the EST data set to be ~5% overall, and the rate of internal priming to be 2–3%. An initial investigation into a small sample of identified reversed clones indicates that a match on the wrong strand is not sufficient to categorize a clone as reversed, and suggests that real transcripts identified as reversed clones may have been removed while preprocessing EGAD. A significant number of these opposite strand matches result from genes overlapping in the opposite orientation. This phenomenon also has been reported elsewhere (Houlgatte et al. 1995), suggesting that overlapping genes is not a rare occurrence. A rigorous analysis of the extent of this phenomenon is imperative, and methodology must be developed in order to avoid merging distinct, overlapping genes into an index class.

METHODS

EGAD–EST Comparisons

Bioccelerator (Compugen 1995) caches were built from FastA (Pearson 1991)-formatted versions of EGAD transcripts and EST data, using the bic_makecache command. Each of the 4420 EGAD transcripts was searched against the set of 42,668 ESTs, using a (two-stranded) multiquery Smith–Waterman search on a four-board Biocellerator-2, with gap creation and extension penalties of 4.50 and 0.05, respectively. Parameters used were listsize = 200, gapweight = 4.50, length = 0.05, minlist = 2.50, minseq = 2, noaverage, and nonormalize. The scoring matrix used sets all matches to 1, all mismatches to −0.6 and all Ns (four-way ambiguities) to 0. Alignments were done as a postprocessing step with the identical parameters, using the program bic_align. Alignments for the top 100 hits were performed. If the 100th alignment was judged significant according to any of the three match criteria, an additional 100 alignments were performed.

Uploading Search Results to Sybase

A Perl (Wall and Schwartz 1992) script was run over search and alignment output files, building bulk copy files suitable for direct upload of the search and alignment results into Sybase (Sybase 1994) tables. Tables were uploaded through the Sybase bcp utility.

Preprocessing of Transcript Data Base

The EGAD data base was screened for the cmvsport, arbl2skm, lt7t3d-pac, and lafmid vectors (WebMaster 1996), using the Smith–Waterman algorithm. Searches were run on the Biocellerator using the parameters gapweight = 5.00, length = 0.30, minlist = 2.50, minseq = 2, noaverage, and nonormalize. HT1138 was identified as contaminated by cmvsport (ZScore = 16.08) and was eliminated from the data set.

The EGAD data base was screened for redundant clones by examining pairs of transcripts that strongly matched the same EST on different strands. Seven pairs were identified in this manner: HT0464-HT4215, HT1512-HT4343, HT2417-HT3615, HT2998-HT4551, HT3391-HT3667, HT2896-HT4390, and HT3183-HT4565. The overlap region of the pair was computed using the GCG (Genetics Computer Group 1994) bestfit program, and used to query the nonredundant nucleotide data base Merck.DNA (Blevins et al. 1995) using BLAST. The BLAST reports were analyzed to ascertain the correct orientation of each clone. The member of each pair corresponding to the likely incorrect orientation was eliminated from the data set (HT0464, HT1512, HT2417, HT2998, HT3667, HT4390, and HT4565).

Relational Representation of Data

Genes, Transcripts, ESTs, and Alignments

A relational data-base schema was created to represent EGAD, ESTs, and the results of the Smith–Waterman alignments. This enabled queries to be performed against both the EGAD data and the alignment results. An Extended Entity Relationship diagram (Chen 1976) for the schema is given in Figure 2. The central concepts are EGAD genes and transcripts, ESTs, and alignments between them, rep-

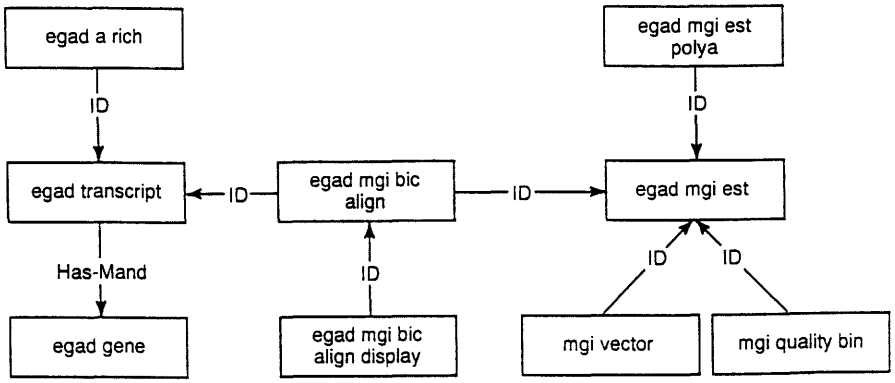

Figure 2 Schema for tables used in analysis. The diagram was drawn using ERDRAW (Szeto and Markowitz 1993). In these schemas Has arcs are many-to-one in the direction of the arrows; the suffix -Mand on an arc label designates that the relationship is mandatory. An ID arc, also many-to-one in the direction of the arrow, denotes that the primary key of the "one" table forms part of the primary key of the "many" table.

resented by the **egad_gene, egad_transcript, egad_mgi_est,** and **egad_mgi_bic_align** tables, respectively. Detailed descriptions of these tables are given in Table 6. There can be many transcripts for a single gene, representing alternative splicing. Most of the information in the gene and transcript tables came from EGAD, with the exception of three columns which were added to the **egad_transcript** table: **cds_start, cds_end,** and **cap_to_polya.** Values for the **cds_length** column were recalculated from **cds_start** and **cds_end** values. The alignment displays were stored in a separate table, **egad_mgi_bic_align_display,** to maximize efficiency of querying the alignment results. All calculated alignments were stored.

Identifying Index Quality EST Sequence

The results of the quality screening were stored in the **mgi_quality_bin** table. Each accession was assigned a quality code. Possible values are: 1 = too short by GSC's standards; 2 = too short by Merck's standards; 3 = marked by NCBI as containing repeat sequence; 4 = index quality. Counts of ESTs in each category can be obtained by executing a simple query in the Structured Query Language (SQL) relational data-base language.

Poly(A) Signals in ESTs

To annotate ESTs for the presence or absence of poly(A) signals, a simple string search was used to identify the presence of three canonical poly(A) signals (AATAAA, ATTAAA, and AATAAT; Birnstiel et al. 1985) within 35 bp of the leading end of all EST sequences. Based on the sequencing methodology, the signals are represented on the opposite strand; therefore, the reverse complement of each signal was utilized in the search (TTTATT, TTTAAT, and ATTATT). The signal type and the offset within the nucleotide sequence were stored in the **egad_mgi_est_polya** table. A sequence could have many poly(A) signals. A null **offset** was used to designate that no poly(A) signal was found within the 35-bp window. This table was used in queries as follows: An EST with accession number **a** has a poly(A) signal if there is a row in the **egad_mgi_est_polya** table with **accession = a** and non-null **offset.**

A-rich Regions of Transcripts

Two different indicators of A-rich regions were used: (1) 6 consecutive A's present, and (2) 14 A's present within 18 consecutive bases. A-rich regions within EGAD transcripts were identified via a Perl script utilizing these criteria. The starting and ending positions of the A-rich regions within the transcript sequence were recorded in the **startpos** and **endpos** columns of the **egad_a_rich** table, along with an integer code representing the relevant indicator. A transcript could have many A-rich regions.

Coding Region Boundaries

Coding region boundaries were derived for EGAD transcripts by identifying the protein coding region of the transcript. This was done through a Perl script. The starting and ending boundaries were recorded in the **cds_start** and **cds_end** columns of the **egad_transcript** table; the length of the coding sequence was recalculated and recorded in the column **cds_length.** In some instances, EGAD did not associate a protein sequence with a transcript, or the nucleotide sequence included no region that could code for the given protein. In these cases, **null** values were recorded for the **cds_start, cds_end,** and **cds_length** columns. In all cases where a coding region could be determined, however, it was unique.

cap-to-poly(A) length

The poly(A) site was determined for each EGAD transcript as the last base prior to the poly(A) tail. The length of the transcript from the 5′ end to this site was recorded in the **cap_to_polya** column in the **egad_transcript** table. If a transcript lacked a poly(A) tail, **cap_to_polya** was set to the EGAD transcript length. For increased accuracy in analyses involving the 3′ end of the EGAD transcript, the **cap_to_polya** length was used instead of the actual EGAD transcript length.

Vector Contamination

EGAD transcripts were screened for vector as described. For all hits, the Zscore and an integer code representing the vector type were recorded in the **mgi_vector** table, which is in a many-to-one relationship to **egad_transcript.** If

Table 6. Relational Schema for EGAD, ESTs, and Smith–Waterman Results

	Column	Type	Description
egad_gene			
	HG_id	varchar(10)	EGAD internal identifier for a gene
	gene_name	varchar(255)	Description of gene, reminiscent of GenBank description line
	KEYS: PRIMARY KEY (HG_id)		
egad_transcript	HG_id	varchar(10)	EGAD internal identifier for a gene
	HT_id	varchar(10)	EGAD internal identifier for a transcript
	seq_name	varchar(255)	Sequence name, reminiscent of GenBank description line
	cds_length	int	Length of coding sequence in nt. Calculated by TIGR. Constraint: (cds_end − cds_start) + 1 = cds_length.
	cds_start	int	Starting position of the coding region in the nt sequence (1 − indexed). Constraint: cds_length = (cds_end − cds_start) + 1.
	cds_end	int	End position of the coding region in the nt sequence (1 − indexed). Constraint: cds_length = (cds_end − cds_start) + 1.
	tx_length	int	Length of the transcript sequence in nucleotides.
	nt_sequence	text	Actual nucleotide sequence
	prot_sequence	text	Actual protein sequence
	cap_to_polya	int	Length of transcript from cap site to polya site. = (polya site − 1). PolyA site is defined as a string of at least 1 A at the 3′ end of the transcript.
	KEYS: PRIMARY KEY (HT_id) FOREIGN KEY (HG_id) REFERENCES egad_gene		
egad_mgi_est			
	accession	varchar(15)	GenBank accession number
	wu_clone_id	varchar(20)	GSC identifier for clone
	insert_size	int	Size of clone insert in bp. Constraint: should be the same for 2 ESTs derived from same clone.
	wu_read_id	varchar(20)	GSC identifier for sequence read. .r = 3′ end of clone, .s = 5′ end of clone.
	gb_length	int	Length of GenBank sequence
	wu_length	int	Length of high-quality sequence according to GSC (high-quality sequence always starts at position 1)
	index_start	int	Start position of index-quality sequence
	index_stop	int	End position of index-quality sequence

Table 6. *(Continued)*

	Column	Type	Description
	index_length	int	Length of index-quality sequence (derivable from index_start, index_stop)
	index_ambig	int	Number of ambiguous bases in index-quality sequence
	library	varchar(30)	Library from which clone originated
	p_end	char(1)	Which end of clone was sequenced? 5 = 5′, 3 = 3′, 0 = unknown.
	KEYS: PRIMARY KEY (accession)		
egad_mgi_bic_align			
	HT_id	varchar(10)	EGAD internal identifier for a transcript
	accession	varchar(15)	GenBank accession number for an EST
	rank	int	Relative strength of the target sequence (EST) hit for the specified query sequence (transcript). Minimum value = 1, maximum value = 200.
	tgt_strand	char(1)	Strand of target sequence: +/− ?
	zscore	float	Z-score of alignment
	sw_score	float	Smith–Waterman score
	align_qual	float	Alignment quality
	align_length	int	Length of alignment
	align_ratio	float	Normalized alignment quality
	align_gaps	int	Number of gaps in alignment
	align_sim	float	Percent similarity of alignment
	align_ident	float	Percent identity of alignment
	align_query_start	int	Start position of alignment in query sequence
	align_query_end	int	End position of alignment in query sequence
	align_tgt_start	int	Start position of alignment in target sequence
	align_tgt_end	int	End position of alignment in target sequence
	align_query_ambig	int	Number of ambiguous bases in query sequence in alignment
	align_tgt_ambig	int	Number of ambiguous bases in target sequence in alignment
	align_id	int	Unique internal identifier for alignment
	KEYS: PRIMARY KEY (HT_id, accession, rank)		
	ALTERNATE KEY (align_id)		
	FOREIGN KEY (HT_id) REFERENCES egad_transcript		
	FOREIGN KEY (accession) REFERENCES egad_mgi_est		
	FOREIGN KEY (align_id) REFERENCES egad_mgi_bic_align_display		

no significant vector contamination was found, the vector type was set to 0.

General Definitions and Techniques

Identifiers

ESTs were identified uniquely by their GenBank **accession** number. Their clones were identified by **wu_clone_id,** the clone id assigned by the GSC. EGAD transcripts and genes were identified by their **HT_id** and **HG_id,** respectively.

Definition of Match between EST and Transcript

All queries were written as parameterized Sybase stored procedures. Every query involving alignments had input parameters **@align_length** (length of the alignment) and **@align_ident** (percent identity of the alignment). A transcript **t** was considered to match or hit an EST **e** if there was a row in the **egad_mgi_bic_align** table with **HT_id = t, accession = e, align_length** ≥ **@align_length** and **align_ident** ≥ **@align_ident.** An EGAD gene was counted as hit by an EST if at least one transcript associated with it was hit by that EST.

Identification of ESTs as 3′ or 5′ Sequence Reads

The EGAD transcript sequence is reported 5′ to 3′; by definition, the + strand of the transcript is its coding strand. Therefore, a + strand hit by an EST on the EGAD transcript identifies the coding strand of the EST, and a − strand hit corresponds to the noncoding strand. Due to the antiparallel nature of DNA, and the fact that sequencing reactions proceed 5′ to 3′, a 5′ read from a cDNA clone represents the + strand of the gene, and a 3′ read represents the − strand. Regardless of their label in the data base, 3′ ESTs are in the same orientation as the − strand of a transcript, and 5′ ESTs are in the same orientation as the + strand.

Summary of Results by Library

Summaries and subtotals grouped by library were accomplished using the SQL **GROUP BY** operator, combined with aggregate functions such as **sum, max, min, avg,** and **count.**

Detailed Descriptions of Selected Queries

The relational data-base schema was designed specifically to facilitate the queries over sequence alignments comprising this study. Consequently, many queries consisted of relatively simple SQL statements using the general definitions and techniques outlined. Detailed descriptions of the more complex queries are provided below.

Reversed Clones/Mislabeled Ends

A 3′ EST is one that hits the − strand of an EGAD transcript, and a 5′ EST is one that hits the + strand of an EGAD transcript. This independent identification of 3′ and 5′ reads was compared with an ESTs **wu_read_id,** the identifier given to the sequence read by the GSC. A true 3′ read should correspond to a **wu_read_id** containing .s, and a true 5′ read should correspond to a **wu_read_id** containing .r. Where discrepancies arise, either the sequence reads have been mislabeled or the clone's orientation was reversed during library construction. Since putative reversed clones were eliminated from the EGAD data set by identifying pairs of transcripts that strongly hit the same EST on different strands, each EST hits EGAD either on the + or the − strand.

The query strategy is as follows. Identify ESTs whose **wu_read_id** corresponds to the strand on which the EST hits, and ESTs whose **wu_read_id** disagrees with the strand on which the EST hits. For each clone, use the SQL **GROUP BY** operator and **count** function to compute **numaccurate** (the number of ESTs with accurate **wu_read_ids**) and **numests** (the total number of ESTs associated with the clone). A clone's orientation completely agrees with EGAD if **numaccurate** = **numests.** A clone's orientation completely disagrees with EGAD if **numaccurate** = 0. A clone is in mixed agreement with EGAD if **numaccurate < numests** and **numaccurate** ≠ 0.

Lane Tracking/Chimeric Clones

Using the **egad_mgi_bic_align, egad_transcript,** and **egad_mgi_est** tables and the definition of 3′ and 5′ EST, two sets of gene-clone pairs were identified: gene-clone pairs where the gene hits the clone via a true 3′ EST from that clone, and pairs where the gene hits the clone via a true 5′ EST from that clone. From these pairs, clones were identified that are in both subsets, i.e., they have at least one EGAD hit on each of their ends. Of these, a lane tracking error/chimeric clone was identified as a clone for which the set of genes hit by its 3′ EST(s) and the set hit by its 5′ EST(s) are disjoint—the 3′ and 5′ EST hits have no genes in common.

Verifying Estimated Insert Sizes

The query strategy was as follows. Identify ESTs that hit only one EGAD transcript. Of these, retrieve clones with both 3′ and 5′ ESTs hitting the same transcript. Of these, require that a clone's 5′ and 3′ ESTs hit in the expected manner, i.e., the 5′ hit begins upstream of the 3′ hit, and the 3′ hit ends downstream of the 5′ hit: (**align_query_end** of 3′ EST alignment ≥ **align_query_end** of 5′ EST alignment) and (**align_query_start** of 5′ EST alignment ≤ **align_query_start** of 3′ EST alignment). Calculate the putative insert size of the clone from the alignment: insert size = (**align_query_end** of 3′ EST alignment − **align_query_start** of 5′ EST alignment). This calculated size **p** was compared with the reported insert size **i,** and the difference **d** expressed as a percentage of the calculated size: **d = 100*p − i/p.**

Internal Priming/Alternative 3′ Ends

This query required two input parameters in addition to

the two parameters specifying the match criterion. **@close_enough** was the maximum permitted distance of the alignment from the 3′ end of the transcript in order for the alignment to be considered at the 3′ end of the transcript. **@close_enough_arich** was the maximum permitted distance of the alignment from the 5′ end of an A-rich region on the transcript in order for an internal priming event to be inferred.

For this query, only 3′ hits (hits on the − strand of the transcript) were considered. A hit was considered to be internal if the alignment was more than **@close_enough** bp from the end of the transcript: (**cap_to_polya − align_query_end**) > **@close_enough.** The query strategy was as follows. Identify 3′ ESTs hitting **only** internal sites on transcripts. Of these, identify the ESTs with poly(A) signals within 35 bp of their 3′ ends by looking up the EST in the **egad_mgi_est_polya** table. These EST hits are interpreted as representing alternative 3′ ends of the EGAD gene. Of the ESTs hitting internal sites but without poly(A) signals, identify those that hit some transcript within **@close_enough_arich** and upstream of an A-rich site on the transcript: (**egad_a_rich.startpos − align_query_end**) ≤ **@close_enough_arich** and (**egad_a_rich.startpos − align_query_end**) ≥ 0. These hits are interpreted as instances of internal priming.

ACKNOWLEDGMENTS

We thank B. Waterston, R. Wilson, M. Marra, L. Hillier and the GSC team, G. Lennon, B. Soares, and the other members of IMAGE, C. Overton and M. Gibson of CBIL, M. Boguski and C. Tolstoshev of NCBI, and A. Williamson and C.T. Caskey of Merck Research Laboratories. We also thank the reviewers for their critical and insightful comments, which have strengthened the paper considerably.

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. and K.O. Elliston. 1996. ftp://avery.merck.com/mgi/IndexReport.

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropolous, H. Xiao, C.R. Merril, Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McCombie, and J.C. Venter. 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, G. Sutton, J.A. Blake, R.G. Brandon, M. Chiu, R.A. Clayton, R.T. Cline, M.D. Cotton, J. Earle-Hughes, L. Fine, L.M. FitzGerald, W.M. FitzHugh, J.L. Fritchman, N.S. Geoghagen, A. Giodek, C.L. Gnehm, M.C. Hanna, E. Hedblom, P.S. Hinkle Jr., J.M. Kelley, K.M. Klimek, J.C. Kelley, L. Liu, S.M. Marmaros, J.M. Merrick, R.F. Moreno-Palanques, L.A. McDonald, D.T. Nguyen, S.M. Pelligrino, C.A. Phillips, S.E. Ryder, J.L. Scott, D.M. Saudek, R. Shirley, K.V. Small, T.A. Spriggs, T.R. Utterback, J.F. Weldman, Y. Li, R. Barthlow, D.P. Bednarik, L. Cao, M.A. Cepeda, T.A. Coleman, E. Collins, D. Dimke, P. Feng, A. Ferrie, C. Fischer, G.A. Hastings, W. He, J. Hu, K.A. Huddleston, J.M. Greene, J. Gruber, P. Hudson, A. Kim, D.L. Kozak, C. Kunsch, H. Ji, H. Li, P.S. Meissner, H. Olson, L. Raymond, Y. Wei, J. Wing, C. Xu, G. Yu, S.M. Ruben, P.J. Dillon, M.R. Fannon, C.A. Rosen, W.A. Haseltine, C. Fields, C.M. Fraser, and J.C. Venter. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–174.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–10.

Benson, D., M. Boguski, D. Lipman, and J. Ostell. 1994. GenBank. *Nucleic Acids Res.* **22:** 3441–3444.

Birnstiel, M.L., M. Busslinger, and K. Strub. 1985. Transcription termination and 3′ processing: The end is in site *Cell* **41:** 349–359.

Blevins, R., J. Aaronson, J. Myerson, G. Hamm, and K. Elliston. 1995. PROFILER: A tool for automatic searching of internally maintained data bases. *Comput. Appl. BioSci.* **11:** 667–673.

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

Boguski, M.S., T.M.J. Lowe, and C.M. Tolstoshev. 1993. dbEST—Database for "expressed sequence tags." *Nature Genet.* **4:** 332–333.

Chen, P.P. 1976. The entity-relationship model—Towards a unified view of data. *ACM Trans. Database Sys.* **1:** 9–36.

Compugen. 1995. BIOCCELERATOR Users Manual. http://www.compugen- us.com/.

Cooke, R., M. Raynal, M. Laudie, F. Grellet, M. Delseny, P.C. Morris, D. Guerrier, J. Giraudat, F. Quigley, G. Clabault, Y.F. Li, R. Mache, M. Krivitzky, I.J. Gy, M. Kreis, A. Lecharny, Y. Parmentier, J. Marbach, J. Fleck, B. Clement, G. Philipps, C. Herve, C. Bardet, D. Tremousaygue, and J. Hofte. 1996. Further progress towards a catalogue of all Arabidopsis genes: Analysis of a set of 5000 non-redundant ESTs. *Plant J.* **9:** 101–124.

Costanzo, F., L. Castagnoli, L. Dente, P. Arcari, M. Smith, P. Costanzo, G. Raugel, P. Izzo, T.C. Pietronaolo, L. Bougueleret, F. Cimino, F. Salvatore, and R. Cortese. 1983. Cloning of several cDNA segments coding for human liver proteins. *EMBO J.* **2:** 57–61.

Efstratiadis, A., J.W. Posakony, T. Maniatis, R.M. Lawn, C. O'Connell, R.A. Spritz, J.K. DeRiol, B.G. Forget, S.M. Weissman, J.L. Slightom, A.E. Blechl, O. Smithies, F.E. Baralle, C.C. Shoulders, and N.J. Proudfoot. 1980. The

structure and evolution of the human beta-globin gene family. *Cell* **21:** 653–658.

Genetics Computer Group. 1994. *Program manual for the Wisconsin package.* Genetics Computer Group, Madison, WI.

Hillier, L., G. Lennon, M. Becker, M. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, M. Soares, F. Tan, E. Trevaskis, K. Underwood, P. Wohldman, R. Waterston, R. Wilson, and M. Marra. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* (this issue).

Houlgatte, R., R. Mairage-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.

Khan, A.S., A.S. Wilcox, M.H. Polymeropoulos, J.A. Hopkins, T.J. Stevens, M. Robinson, A.K. Orpana, and J.M. Sikela. 1992. Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genet.* **2:** 180–185.

Ko, M.S., X. Wang, J.H. Horton, M.D. Hagen, N. Takahashi, Y. Maezaki, and J.H. Nadeau. 1994. Genetic mapping of 40 cDNA clones on the mouse genome by PCR. *Mamm. Genome* **5:** 349–355.

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.

Nishikawa, T. and K. Nagai. 1996. EST error analysis in a large-scale GenBank search of ESTs using rapid-identity-searching program for DNA sequences. In *Genome mapping and sequencing,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Okubo, K., N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2:** 173–179.

Pearson, W.R. 1991. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA Algorithms. *Genomics* **11:** 635–650.

Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

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

Sybase. 1994. *Sybase SQL server reference manual* (2 vols). Server Publication Group, Sybase, Inc., Emeryville, CA.

Szeto, E. and V.M. Markowitz. 1993. *Erdraw 5.3: A graphical editor for extended entity-relationship schemas. Reference manual.* Lawrence Berkeley National Laboratory, Berkeley, CA.

Wall, L. and R.L. Schwartz. 1992. *Programming Perl.* O'Reilly and Associates. Sebastopol, CA.

Waterson, R., C. Martin, M. Craxton, C. Huynh, A. Coulson, L. Hillier, R. Durbin, P. Green, R. Shownkeen, N. Halloran, M. Metzstein, 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.

WebMaster. 1996. http://www-bio.llnl.gov/bbrp/image/humlib_info.html.

White, O. and R. Kerlavage. 1996. TDB: New databases for biological discovery. *Methods Enzymol.* **266:** 24–40.

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.

Williamson, A.R., K.O. Elliston, and J.L. Sturchio. 1995. The Merck Gene Index, a public resource for genomics research. *J. NIH Res.* **7:** 61–63.

Received June 13, 1996; accepted in revised form August 5, 1996.

Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data.

J S Aaronson, B Eckman, R A Blevins, et al.

Genome Res. 1996 6: 829-845
Access the most recent version at doi:10.1101/gr.6.9.829