Assembly, Annotation, and Integration of UNIGENE Clusters into the Human Genome Draft

Degen Zhuo, Wei D. Zhao, Fred A. Wright, et al.

Genome Res. 2001 11: 904-918
Access the most recent version at doi:10.1101/gr.164501

Resource

Assembly, Annotation, and Integration of UNIGENE Clusters into the Human Genome Draft

Degen Zhuo,[1,2,5] Wei D. Zhao,[1,2,5] Fred A. Wright,[2,5] Hee-Yung Yang,[4] Jian-Ping Wang,[1,2] Russell Sears,[1,2] Troy Baer,[3] Do-Hun Kwon,[1,2] David Gordon,[1,2] Solomon Gibbs,[1,2] Dean Dai,[4] Qing Yang,[1,2] Joe Spitzner,[4] Ralf Krahe,[2] Don Stredney,[3] Al Stutz,[3] and Bo Yuan[1,2,6]

[1]Bioinformatics Group, [2]Division of Human Cancer Genetics, James Cancer Hospital and Solove Research Institute, The Ohio State University, Columbus, Ohio 43210, USA; [3]Ohio Supercomputer Center (OSC), Columbus, Ohio 43212, USA; [4]Labbook.Com, Columbus, Ohio 43229, USA

The recent release of the first draft of the human genome provides an unprecedented opportunity to integrate human genes and their functions in a complete positional context. However, at least three significant technical hurdles remain: first, to assemble a complete and nonredundant human transcript index; second, to accurately place the individual transcript indices on the human genome; and third, to functionally annotate all human genes. Here, we report the extension of the UNIGENE database through the assembly of its sequence clusters into nonredundant sequence contigs. Each resulting consensus was aligned to the human genome draft. A unique location for each transcript within the human genome was determined by the integration of the restriction fingerprint, assembled genomic contig, and radiation hybrid (RH) maps. A total of 59,500 UNIGENE clusters were mapped on the basis of at least three independent criteria as compared with the 30,000 human genes/ESTs currently mapped in Genemap'99. Finally, the extension of the human transcript consensus in this study enabled a greater number of putative functional assignments than the 11,000 annotated entries in UNIGENE. This study reports a draft physical map with annotations for a majority of the human transcripts, called the Human Index of Nonredundant Transcripts (HINT). Such information can be immediately applied to the discovery of new genes and the identification of candidate genes for positional cloning.

There are at least three principal products of the Human Genome Project: the sequence itself, the genes, and the map integrating the genes and the sequence. This has been partially accomplished by the release of the human genome draft, which represents 85%–90% of the entire human genome (http://www.ncbi.nlm.nih.gov/genome/seq/HsHome.shtml). The Ensembl consortium has annotated the draft with known transcripts, protein sequences, and hypothetical genes (http://www.ensembl.org). However, most of the transcripts used in the Ensembl database were individual ESTs, which were single pass, potentially highly redundant, and partial cDNA fragments derived from either the 5′- or 3′ -untranslated regions of human genes (Boguski et al. 1994; Adams et al. 1995; Schuler et al. 1996). We proposed that assembly of individual ESTs into a consensus sequence could provide the basis for discovery of additional relationships with the genomic sequence, orthologous genes, and/or conserved functional domains.

Several groups have compiled sets of human transcripts, generating gene-oriented clusters, indices, or consensus sequences (Boguski and Schuler 1995; Miller et al. 1999; Ewing and Green 2000; Quackenbush et al. 2000). We chose UNIGENE as a starting point to annotate the human genome, because, in contrast to other gene indices, it provides links to the IMAGE, Genemap'99, Locuslink, SAGE, and protein databases, as well as cytogenetic and other mapping information (Boguski and Schuler 1995). The current version of UNIGENE has ~87,000 nonredundant sequence clusters assembled from >2.1 million individual ESTs and other cDNA sequences (http://www.ncbi.nlm.nih/UniGene/). However, the resulting UNIGENE database is still composed of individual transcripts, each the longest in its progenitor cluster. We thus set out to assemble the UNIGENE database to extend the human transcript consensus.

Transcript mapping is used to facilitate gene discovery. Because >370,000 ESTs and other mRNA se-

[5]These authors contributed equally to this work.
[6]Corresponding author.
E-MAIL yuan.33@osu.edu; FAX (614) 688-4761.
Article published on-line before print; *Genome Res.,* 10.1101/gr.164501. Article and publication are at www.genome.org/cgi/doi/10.1101/gr.164501.

11:904–918 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

quences had already been mapped by the international radiation hybrid (RH) consortium (Deloukas et al. 1998), we initially sought to integrate UNIGENE and Genemap'99 to map all the consensus transcripts. However, only ~19,750 UNIGENE clusters were found to contain at least one STS or EST that could be localized on either the GB4 or G3 map—significantly fewer than the 30,000 distinct genes described in Genemap'99 (Deloukas et al. 1998). Thus we were motivated to use the draft of the human genome to map all human genes with sufficient transcript representation. Because >95% of the finished and nonfinished clones sequenced by the Human Genome Project have been physically mapped, and the human genome draft represents >85%–90% of the nonredundant human genomic sequences, global sequence alignments between the consensus transcripts and the working draft were possible. The overall goal of our gene-indexing project is to use the genome draft to develop a transcript map that integrates both positional and functional information for most human genes.

We report a computational approach for producing high-quality transcript assemblies. To ensure the integrity of the assembly process, we assessed the Smith–Waterman scores for each individual transcript contributing to the consensus. Unalignable variants and other chimeric contaminants were first identified and removed. When more than one contig was generated from each cluster, consensus sequences were matched to exclude any internal associations. Problematic contigs were reassembled by use of more stringent conditions. Our resulting database, the Human Index of Nonredundant Transcript (HINT), contains a total of 97,687 consensus transcripts, including 61,066 contigs and 36,621 singlets assembled from 87,000 UNIGENE clusters. In addition, 8100 potential splicing variants were identified during the sequence assembly, of which only 5% had been reported previously in the literature. Interestingly, many of the potential variants had unique tissue library profiles in their progenitor clusters, implicating alternative splicing as possibly a highly regulated event.

We demonstrate the use of the human genome draft to map all human transcripts. We emphasize the importance of quality control and map integration to resolve ambiguous placements for both the genomic clones and consensus transcripts. Nonoverlapping, high-scoring segment pairs (HSPs) from the same transcripts spliced on the same genomic contig were retained. Remaining overlapping alignments were resolved by scoring both the length and identity of all of the transcripts involved, resulting in the assignment of a unique location for a given cDNA fragment on the human genome. Finished and unfinished genomic clones, along with their aligned transcripts, were physically ordered by integration of the restriction fingerprint map for the genomic clones (Marra et al. 1997), the genomic contigs assembled by Haussler and colleagues (http://genome.cse.ucsc.edu/goldenPath), Genemap'99 (Deloukas et al. 1998), and the e-PCR map developed in this study.

The final transcript map was substantiated by comparison of the clone, contig, genetic, and cytogenetic mapping information, and evaluating supporting and conflicting evidence. Together, a total of 59,500 UNIGENE clusters have been mapped, providing an early glimpse of a complete transcript map for the human genome.

RESULTS AND DISCUSSION

Assembly of UNIGENE Clusters

First, we assembled the UNIGENE clusters. Rather than creating a new gene index, the goal of our assembly was to extend the human transcript consensus to facilitate gene annotation and mapping. Assembly was carried out strictly within each UNIGENE cluster. Chimeric sequences were first identified and removed (see the section concerning splicing variants). A total of 46,202 UNIGENE clusters were each assembled into single contigs (53%), with an additional 7570 UNIGENE clusters resulting in more than one contig (9%). A total of 6240 ESTs were not assembled along with other contigs in the same clusters. Because they represented unique and nonconflicting sequences, these singlets were incorporated into the consensus of their progenitor clusters. A total of 2200 UNIGENE clusters had to be assembled under more stringent conditions (2.5%), during which an additional 3800 chimeric ESTs were identified and discarded as conflicting singlets. A total of 770 UNIGENE clusters, most of which contained either many ESTs and/or very long transcripts, could not be assembled on the basis of our criteria (0.7%). In this case, the longest transcript was used to represent the consensus. The remaining 30,381 UNIGENE singletons (35%) were not assembled.

Second, we compared each resulting consensus to its contributing transcripts. The accuracy of sequence assembly can be significantly improved by use of the quality scores derived from the original sequencing traces (Ewing and Green 1998; Ewing et al. 1998). As original chromatograms were not used in our assembly, we applied position-dependent quality scores based on a published assessment of six public sequencing projects (Richterich 1998). In this analysis, it was shown that the first 30 bp had high error rates, in the range of 5%–10%. Toward the end of sequence reads, the error rates generally exceeded 10% between bases 300–700, depending on the template, sequencing chemistries, sequencing machines, or gel length used in each of the six projects (Richterich 1998). We thus applied a conservative approach, by assigning higher

quality scores only to the 31–300 base positions for ESTs that were longer than 500 bp. A nearly perfect quality score was given to all bases of known genes (Methods).

Many of the UNIGENE clusters (15%) contained at least one known gene. The overall Smith–Waterman scores for the known genes in their corresponding contigs were 1.8% for mismatch, 0.7% for insertion, and 0.5% for deletion, indicating that the resulting consensus was reasonably accurate at the sequence level. Similar overall Smith–Waterman scores were obtained for the remaining transcripts associated with the known genes (mismatch 1.5%, insertion 0.9%, and deletion 0.8%). This accuracy might be partially attributed to the high `Phred` quality score given to the known genes, which would reduce the contribution from ESTs of potentially poorer qualities. Alternatively, the redundancy of current EST information was high enough that discrepancies due to random sequencing errors had been largely compensated. This appeared to be true as the overall Smith–Waterman scores for rest of the contigs were in the same range (mismatch 1.9%, insertion 1.7%, and deletion 0.9%). Interestingly, no substantial discrepancies in Smith–Waterman scores were observed for the ESTs of different lengths. We believe that the observed accuracy for the resulting contigs could also be attributed to the stringent threshold we set during the assembly process, that any transcripts presenting higher than the accumulative >5% error rate would be discarded as singlets. We speculate that some of the sequence mismatches are due to single-nucleotide polymorphisms (SNPs) in human ESTs (Buetow et al. 1999; Irizarry et al. 2000).

Extension of UNIGENE Consensus Sequences

Third, we assessed the benefits of our assembly by comparing each UNIGENE index to its corresponding HINT consensus (Table 1 and Fig. 1). We did not include any of the chimeric ESTs to assure that only unique consensus sequences were compared. We divided the UNIGENE data set into anonymous ESTs and known gene categories using the UNIGENE descriptions (see Methods). Surprisingly, almost 65% (7234) of the 11,191 known genes could be extended further by the assembly process, by an average of 20% (Table 1). In contrast, 30% (22,796) of the 75,925 anonymous EST clusters were extended, and the resulting HINT clusters were an average of 50% longer than their UNIGENE counterparts. This result was somewhat expected, as known genes may have been more easily characterized due to high expression (as evidenced by their greater number of transcripts per cluster, resulting in a greater chance of assembly extension). However, the full-length mRNA forms an upper limit to proper extension. The majority (72.5%) of the known gene UNIGENE clusters were >1000 bp, while almost all (98.5%) of the anonymous EST clusters were <1000 bp (corresponding to the limit of single-pass sequencing). Furthermore, known gene UNIGENE consensus of <1000 bp tended to be directly sequenced ESTs rather than deposited mRNA sequences. Thus, we indicated 1000 bp as a rough division between the categories (Fig. 1). All of these observations were consistent with the nature of current EST information, which consisted mainly of redundant and partial cDNA sequences (Aaronson et al. 1996).

Table 1. Benefit Assessment of the Assembly by Comparing the Original UNIGENE Transcripts to Their Corresponding Consensus

	Known genes	Anonymous ESTs
UNIGENE clusters	11191	75,925
Singletons	692	29,689
UNIGENE clusters extended	7237	22,795
Average number of transcripts	97	18
Average length of UNIGENE (bp)	2280	560
Average length of HINT (bp)	2749	843
UNIGENE clusters unchanged (Non-singletons)	3262	23,441
Average number of transcripts	86	4
Average length of UNIGENE (bp)	2833	506

UNIGENE clusters containing known genes or only anonymous ESTs were divided (see Methods). The sizes and lengths for each UNIGENE cluster were based on the original UNIGENE dataset. bp, Base pair.

Overall, we assembled the 1.51 million ESTs (260 Mb) comprising UNIGENE into 42.5 Mb of consensus sequence. Thus, we estimated the redundancy of the EST information as about sixfold.

Integrity Assessment of the HINT Database

We assessed the integrity of our HINT database through three independent approaches. First, we integrated Genemap'99 into the UNIGENE database for the individual ESTs with RH mapping information. Inconsistent chromosomal assignments of multiple RH markers in the same UNIGENE clusters suggested misassembly. A total of 19,464 UNIGENE clusters could be placed on either the GB4 and/or G3 map on the basis of ~375,000 markers. Only an additional 286 UNIGENE clusters were not consistently mapped, containing a total of 2695 misplaced ESTs. About one-third of these inconsistent clusters (93) were later resolved on the basis of the majority of mapped ESTs in the same clusters. The remaining 193 clusters were clearly misassembled, containing more than one distinct mapping group, each having at least two mapped ESTs (see Supplementary Table 1 at http://pandora.med.ohio-state.edu/HINT). Interestingly, almost all of the misassemblies represented known genes belonging to either large gene families or evolutionarily conserved func-

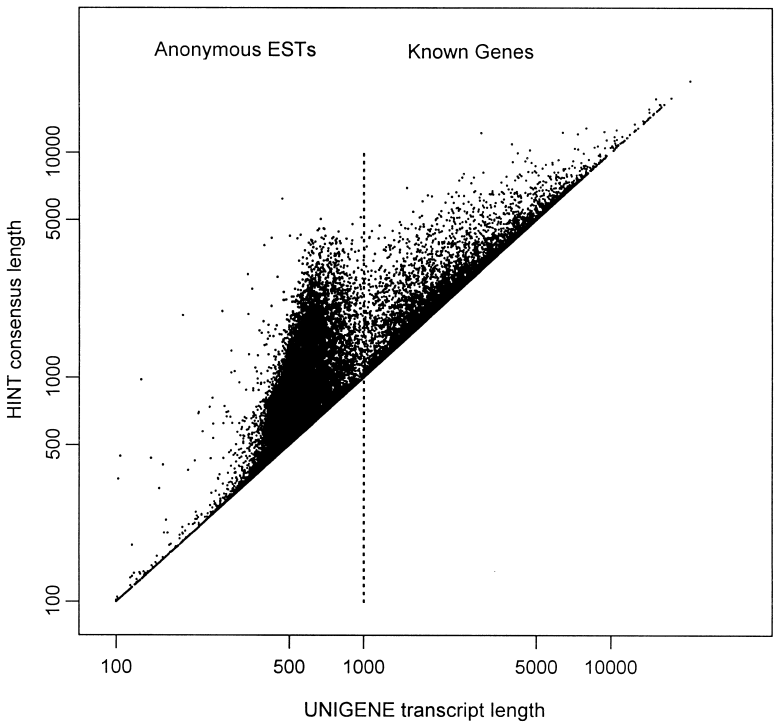

Figure 1 The extension of human transcript consensus sequences after the assembly of UNIGENE clusters. The lengths of transcript consensus were compared for each UNIGENE cluster before and after the sequence assembly (log–log scale). The (*x*-axis) Sequence length for the original UNIGENE transcripts; (*y*-axis) sequence length for the assembled HINT consensus. Almost none of the anonymous UNIGENE EST entries is longer than 1000 bp (above the limit at regular single sequencing reads), while most of the known genes in the original UNIGENE consensus are longer than 1000 bp.

tions, including ATP binding cassette-containing transporters, membrane receptors, CDCs, protein kinases and phosphatases, ubiquitin-specific proteases, cytokines, G-proteins, histones, and electron transporters. The top 10 UNIGENE clusters with the most inconsistency (>40 inconsistently mapped ESTs in each of the clusters) were *PEDF* (protease inhibitor), *TUBB2* (tubulin beta 2), *SDHA* (flavorin-containing dehydrogenase), *IDH2* (NADP+-associated dehydrogenase), *PRDX1* (peroxidase), *EPB41L2* (erythrocyte membrane band 4.1-like), *UBE2V1* (ubiquitin-conjugating enzyme E2), *ATP5J2* (ATP synthase, H+ transporting), *HDAC3* (histone deacetylase 3), and *PPP1R2* (protein phosphatase 1, subunit 2) (see Supplementary Table 2 at http://pandora.med.ohio-state.edu/HINT). These observations imply that paralogous genes (sequence similarity due to gene duplication) or homologous genes (sequences containing evolutionarily conserved functional domains) might be assembled into the same UNIGENE clusters, contributing to >95% of the misplaced ESTs (2601/2695).

We reasoned that sequencing errors, mapping errors, and other artifacts would otherwise result in such misassembly on a random basis. Sequence similarity thus appeared to be a major pitfall when having to assemble a large amount of unique, paralogous, and homologous human genes. This was very true in the UNIGENE building procedure, during which anonymous ESTs had to be clustered largely on the basis of sequence similarities.

The international RH consortium has indicated that at least 30,000 genes were mapped in Genemap'99 (Deloukas et al. 1998), from which almost 95% of the markers (~375,000) were contained in UNIGENE. Though the TXMAP was not used for this integration (Deloukas et al. 1998), which could place an additional 5000 genetically mapped UNIGENE clusters, we speculated that significantly more redundant ESTs had been typed than the estimate indicated in Genemap'99.

As a second assessment of the integrity of HINT, we compared the orientations for the individual ESTs in their contigs to those in their original cDNA clones. Inconsistent orientations of multiple ESTs in the same sequence contig suggested misassembly. About 15% of the ESTs assembled (290,000) showed inconsistent sequence orientations in their contigs (see Supplementary Table 3 at http://pandora.med.ohio-state.edu/HINT). To determine whether this inconsistency was due to poor assembly, we compared the overall Smith–Waterman scores for the apparently inverted ESTs to the rest of the database. No significant differences in the Smith–Waterman scores were observed, strongly suggesting that the clone direction annotations might be mislabeled in the original IMAGE database, or the error rate in the directional cloning process was significantly high. An alternative explanation for the inversion in the EST assemblies, suggested by Burke and colleagues, could be attributed to two different genes overlapping on two opposite strands (Burke et al. 1998). Evidence of overlapping genes from two different strands has been noted in the human genome (Ashworth 1993; Tsai et al. 1994; Houlgatte et al. 1995; Hillier et al. 1996; Burke et al. 1998). We thought that the inversion issue could be at least partially resolved if multiple known mRNAs assembled in the same UNIGENE clusters had different functional annotations. However, on the basis of our preliminary database integration study using both the UNIGENE and GenBank descriptions, no obvious discrepancies were observed for the clusters or individual genes. Thus the cause for the apparent transcript inversion remains to be determined.

As a third assessment of HINT integrity, we used the human genome draft to integrate and compare our consensus to Ewing and Green's (EG) EST assemblies using high-quality traces (Ewing and Green 2000). We attempted to address the following questions: (1) Because the HINT consensus was built without the use of original chromatograms, what impact would this have on the fidelity of the resulting consensus? (2) How many exonic sequences were in common between the two indices, and how much unique transcribed information was contributed from each? The availability of the draft genome enables a direct comparison of transcript indices. Using BLAST, assembled transcripts from HINT and the EG indices were assigned to the genomic sequences. UNIGENE singletons were not used for these comparisons, as the EG indices were built from multiple ESTs only.

The comparisons revealed HINT (BLAST score 472, sequence identity 93.77%) was about 3% lower in average quality score than the EG assemblies (490, 95.43%). This appeared consistent with the overall Smith–Waterman error rates for HINT (~5%) and the threshold we set (5%) during the assembly process. The average BLAST scores occupied a limited range (471–475) for UNIGENE clusters of different sizes, lengths, or known genes versus anonymous ESTs. Similarly, sequence identity for all of these groups was near 93%.

The overlap between the two indices was substantial but not complete (e.g., 62% of the EG indices appear in HINT). The unique transcripts column in Table 2 enumerates transcripts that appeared in one index and not the other. Note that the grouping of ESTs into clusters differed somewhat for the two indices, so that comparisons of HINT clusters to those of EG were not always one to one. Because the two indices represented two different transcript populations, the comparison was narrowed further to the overlapping HSPs within each transcript. For simplicity, we referred to the (spliced) segments (HSPs) mapped to the genome as exons, although the computational evidence awaited further biological confirmation. A total of 5363 HINT transcripts had at least one exon in complete overlap (allowing up to a 2-bp difference) with an EG exon, while 5675 EG consensus sequences had an exon in common with HINT. Again, the two values differed because the indices did not match one to one. The genomic sequence identity of HINT versus EG overlapping exons was about 3% lower than EG versus HINT overlapping exons (Table 2). Similar comparisons were given for consensus transcripts containing exons that were longer (extended) in HINT versus EG, truncated in HINT versus EG, and transcripts that overhang, that is, contain overlap that represented only a portion of both the HINT and EG exon.

Table 2. The Integration of HINT Consensus and Ewing and Green's Assemblies into the Human Genome Draft

Comparison of Mapped Transcripts		
	HINT	**Ewing and Green (EG)**
Transcripts mapped	43,484	29,582
HSP length (kb)	34,226	12,723
Average BLAST score to genome	472	490
Overall sequence identity with genome	93.77%	95.43%
Unique transcripts	28,026	11,002
HSP length (kb)	24,568	3,452
Average BLAST score to genome	471	490
Overall sequence identity with genome	93.53%	95.18%
Comparison of Exons in Transcripts		
	HINT vs. EG	**EG vs. HINT**
Transcripts with an exact exon overlap	5,363	5,675
HSP length (kb)	1,731	1,731
Average BLAST score to genome	475	490
Overall sequence identity with genome	93.86%	95.25%
Transcripts with an exon extended	7,580	4,759
HSP length (kb)	4,682	2,129
Average BLAST score to genome	472	490
Overall sequence identity with genome	93.88%	95.69%
Transcripts with an exon truncated	5,142	2,012
HSP length (kb)	1,724	2,989
Average BLAST score to genome	471	490
Overall sequence identity with genome	93.62%	96.02%
Transcripts with an exon overhanging	229	850
HSP length (kb)	253	656
Average BLAST score to genome	463	483
Overall sequence identity with genome	91.41%	93.32%

(HSP) High-scoring segment pair, representing exons. Average BLAST score was obtained by dividing the total BLAST score by the total length of the HSPs involved, represented as per 100 base pair HSP. Overall sequence identity was similarly averaged by the total length of HSP involved.
(Overlap) HSPs overlap by ± 2 base pairs on either side between the HSPs of HINT and EG.
(Extended) HSPs from one index were longer than the other one.
(Truncated) HSPs from one index were shorter than the other one. In both cases, the positions of two HSPs agreed on either 5′ or 3′ side of their HSPs within ± 2 base pairs.
(Overhanging) both the 5′ and 3′ positions of the two HSPs disagreed (> ± 2 base pairs).
(kb) Kilobase pairs.

A majority of the extended and truncated HSPs (>75%) represented either the 3′ or 5′ end of a gene index, strongly indicating that the difference was largely due to the length of the consensus rather than misassembly. Though it was expected that longer consensus would be generated from the UNIGENE database consisting of a significantly larger population of ESTs, still a significant number of EG consensus was extended further for the corresponding transcripts. We attributed the remaining internal variations (25%) largely to alternative splicing. After the HSPs contributed from the splicing variants in the HINT database were removed (see the section concerning splicing variants), the internal difference dropped to about 15%. To determine whether the remaining internal discrepancies could represent misassembled consensus, we compared EG consensus to known GenBank sequences as well as to TIGR's HGI. We observed a similar degree of internal variations among all of the gene indices tested, strongly suggesting that alternative splicing might occur as a rather frequent event (B. Yuan, unpubl.).

Finally, a minor amount of imperfect (overhanging) overlaps was observed, implicating misassembly, misplacement, or genes on two different strands (Table 2). To distinguish these possibilities, we checked the orientations and evidence of splicing. Interestingly, most of them (~80%) represented HSPs on two different strands with evidence of splicing, suggesting overlapping genes. Though the gene-overlapping issue remains largely unresolved, it appeared that such a phenomenon might occur more frequently in the human genome than previously thought (Burke et al. 1998). An additional 15% appeared to be caused by confusion during the resolution of the BLAST placements (see Methods), suggested by the lack of evidence for splicing and the overall drop of sequence identity (~2%). Similar observations were also made when comparing Ewing and Green's with other gene indices (B. Yuan, unpubl.). Even so, we still could not rule out completely the possibility of misassembly for at least some of the imperfect or truncated overlaps.

The integration of transcript consensus into the human genome draft also allowed us to find overlapping HINT consensus (~5450 HSPs), suggesting agglomerate clusters. A similar phenomenon was also observed (~1570) within Ewing and Green's index (see Supplementary Tables 4 and 5 at http://pandora.med.ohio-state.edu/HINT;). We were not able to determine whether they were distinct genes, which were either contained in and/or overlapped other genes, or components of the same genes. This issue was complicated further due to the lack of original orientation information for at least some of the consensus. We speculated a majority of them represented fragmented indices, as many of the sandwiched indices were singletons or consensus sequences derived from small clusters (see Supplementary Tables 4 and 5 at http://pandora.med.ohio-state.edu/HINT).

We were surprised by the lack of common consensus between HINT and EG indices (Table 2). Here, it was clearly indicated that the two data sets represented two very different populations of human transcript information, with only ~30% of the UNIGENE index being represented in Ewing and Green's assemblies. Even though UNIGENE was thought to consist of a majority of human transcript information, still ~15% of Ewing and Green's assemblies were not represented. Similar observations were made between our HINT and TIGR's HGI (B. Yuan, unpubl.). This raises the issue that all existing human transcripts including known genes are required to create a complete and nonredundant gene index for the human genome.

An estimate of 35,000 human genes was obtained on the basis of Ewing and Green's assembly (Ewing and Green 2000). At least one major difference between this approach and UNIGENE (Boguski and Schuler 1995) or TIGR's HGI (Liang et al. 2000) has been that consensus sequences from single unconfirmed ESTs were not included. To address whether singletons represented real genes, we integrated the ~30,000 UNIGENE singletons into the human genome draft, resulting in 16,305 unique placements (see Methods). A total of 3340 singletons had evidence of splicing (see Supplementary Table 6 at http://pandora.med.ohio-state.edu/HINT). Given that 85% of the UNIGENE consensus was placed in the human genome (see the section concerning transcript mapping), still only a small portion of the singletons (15%) could be real genes. Thus, the issue whether or not singletons represent human genes or genomic contaminants and other artifacts remains largely unresolved.

Thus, gene counts cannot be based directly on transcript clustering and assembly. Lack of overlapping evidence among individual consensus sequence could significantly inflate the current estimates. This can be particularly true as human transcript information was largely derived from the 3′ or 5′ termini of human genes. Second, genomic contaminants, artificial or inaccurate ESTs could have inevitably been included in the UNIGENE data sets, which would contribute a significant number of additional singletons as genes. Third, new genes and their representations remain to be discovered, which can be particularly true for genes that are either very low in abundance or expressed only in specific tissues or limited times. Wheelan and Boguski proposed algorithms to identify and annotate all human genes (Wheelan and Boguski 1998). We believed that the integration of all human transcribed information into the genome, together with predicted exons, protein homology, and new algorithms could significantly facilitate the identification and annotation of at least a majority of human transcriptional units.

Identification of Splicing Variants in the UNIGENE Database

Splicing variants and other chimeric contaminants had to be first identified and removed from each cluster before sequence assembly. In fact, chimerism was indicated as one of the main reasons why no attempt had been made at NCBI to assemble the UNIGENE clusters (http://www.ncbi.nlm.nig.gov/UniGene). For our efforts, only the transcripts with an internal deletion or internal nonoverlapping alignment compared with their consensus were considered as potential splicing variants. A total of 8100 such candidates were identified, consisting of 6500 deletional plus about 1600 insertional variants (see Supplementary Tables 7 and 8 at http://pandora.med.ohio-state.edu/HINT). An additional 14,500 chimeric sequences were identified and discarded on the basis of the fact that the nonoverlapping alignments were only found at their 5′ or 3′ termini compared with their corresponding consensus. We speculated that these ESTs could largely represent partially processed RNAs or genomic contaminants. The 8100 potential variants were derived from a total of 6713 individual UNIGENE clusters, of which only 287 had more than one variant. This observation indicated that a majority of the alternative splicing events (6426, 80%) was each supported by only one transcript. Our preliminary validations revealed no significant differences in the size distributions for the variants (~505 bp) compared with the rest of the ESTs (~525 bp). Because we placed strict conditions on the formation of the consensus and included the deletion variations that were only internal to a transcript, the possibility of intron contamination was ruled out in the vast majority of the variants we considered. We later also ruled out the possibility of artifacts (genomic contamination or incomplete RNA processing) for a majority of the insertional variants by identifying their alternatively spliced exons in the human genome draft (see Supplementary Table 9 at http://pandora.med.ohio-state.edu/HINT).

Because many of the ESTs in the UNIGENE database are single-pass reads of low sequence quality, gross sequence variations found near the 5′ or 3′ ends of the ESTs were ignored and trimmed before the subsequent round of assembly. Some transcript variations could also be attributed to the hypervariable microsatellites contained in exons. A recent study suggested that such sequence variations could exist in as many as 20% of all human genes (Wren et al. 2000). By use of the program `RepeatMasker` on the sequences identified as deleted or inserted in this study, no significant simple repetitive sequences were observed.

Because splicing variations often resulted in gain or loss of a function, we compared the patterns of protein motifs on the variants to their counterparts. Interestingly, unique exons contributed from 75% of the 1300 insertional variants tested contained at least one conserved protein motif, detected by use of the program `HMMER` and the Pfam database (Bateman et al. 2000). The top three frequently found motifs were kinase (62), tyrosine phosphatase (18), and SH2/SH3 (15) domains, all involved in signal transduction (see Supplementary Table 10 at http://pandora.med.ohio-state.edu/HINT). More extensive molecular analysis on all potential splicing variants is underway.

To determine whether any of the variants were tissue, pathology, or development specific, we standardized the vocabularies used to describe each individual cDNA library. Terms were then divided into organ, tissue, cell, pathology, developmental, or experimentally treated categories (see Supplementary Table 11 at http://pandora.med.ohio-state.edu/HINT). The cDNA source for each variant was compared with the profile of all cDNA libraries of its progenitor cluster (see Supplementary Table 12 at http://pandora.med.ohio-state.edu/HINT). By doing so, 40% of the variants were found to be tissue, development, or neoplastic specific, implicating alternative splicing as a potentially highly regulated event. However, because the relative expression level of individual genes and the cDNA libraries used in the EST sequencing project can both affect the relative abundance of a particular tissue source represented in a given UNIGENE cluster, cDNA library profiles could only be used as presumptive evidence when inferring gene expression profiles. This was particularly true in cases where early cDNA cloning efforts were more successful with highly expressed genes.

To estimate how many of the splicing variants were potentially novel, splicing-related annotations such as alternative splicing were identified in the SWISSPROT, PIR, and TrEMBL protein databases, as well as GenBank; only ~5% of the variants had previously been identified. However, this estimation was based mainly on the information already annotated in the available sequence databases rather than the entire public literature database. A number of genomewide searches for splicing variants have been similarly conducted by others as well (Burke et al. 1998; Mironov et al. 1999).

In the end, the HINT consensus was composed of 61,066 contigs, 36,621 singletons, and 8100 splicing variants, representing the original 87,000 UNIGENE indices.

Functional Annotation of the HINT Database

Functional annotation was simplified by the use of consensus to represent each gene. To assess the benefits of using our extended consensus on sequence annotation, the original UNIGENE indices and their corresponding consensus sequences were compared for their alignments to protein databases (Fig. 2). More than 13,000 UNIGENE clusters, most of which were

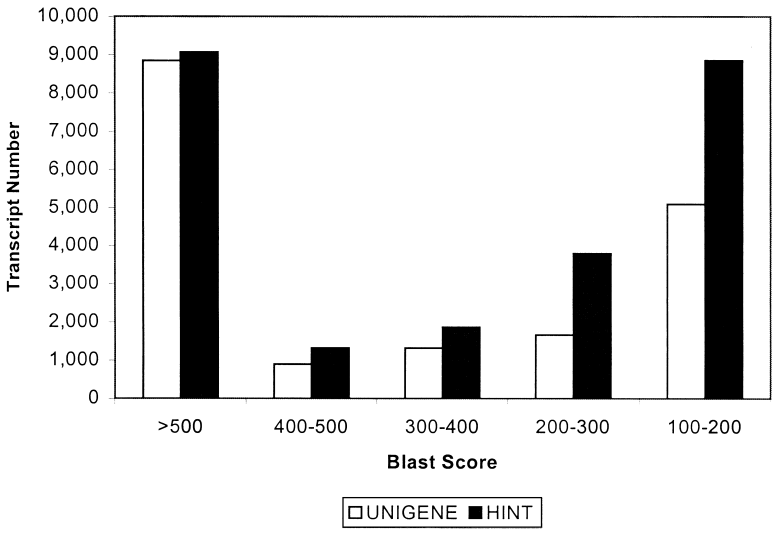

Figure 2 The assembly of human transcripts enabled a greater number of UNIGENE entries with a protein homology. The UNIGENE and HINT consensus sequences were aligned to the SWISSPROT, TrEMBL, and PIR protein databases by use of the program BLASTX. Protein alignments with the highest scores were selected from each UNIGENE and HINT transcript consensus and broken into five groups according to their BLAST scores, ranging from the highest >500 to 100–200 (*x*-axis). (*y*-axis) Numbers of transcript consensus. (Open bars) UNIGENE transcripts; (closed bars) HINT consensus.

previously labeled as ESTs, had at least one new putative annotation based on their homology to known protein sequences (see Supplementary Table 13 at http://pandora.med.ohio-state.edu/HINT). A greater number of highly significant hits (BLAST score >400) was obtained as the result of the extension (~1,500). However, a substantial amount of new protein-related information was obtained with marginal BLAST scores (<200), suggesting that current EST databases were still largely composed of 3′ and 5′ ends of human transcripts. Further extension of current transcript consensus sequence is necessary for more functional annotations to be incorporated.

We chose the SWISSPROT, TrEMBL, and PIR protein databases to annotate the HINT database, because functional annotations for each protein entry in these databases had already been organized and largely standardized (Junker et al. 1999). Additional function-related annotations were obtained through the links of UNIGENE to the Locuslink database (see Locuslink TBL in Fig. 3).

UNIGENE has been built as a nonredundant and largely gene-oriented index, with each UNIGENE cluster having at least one known gene, or two ESTs representing the 3′ terminus of a gene, or at least one EST containing a poly(A) signal for the singletons (http://www.ncbi.nlm.nih.gov/UniGene/build.html). It is our belief that the integration of this gene index into the genome draft together with the identification of more exonic sequences could significantly facilitate the current gene-oriented annotation efforts.

The Draft for Human Transcript Map

Our transcript mapping effort faced three major problems: sequence fragmentation and redundancy in the current human genome draft, and paralogous/homologous sequences in the assembled gene index. Genomic sequence fragmentation results in ambiguities for the placement and assembly of genomic sequences. Genomic sequence redundancy results in multiple alignments for the same transcript onto different clones of the human genome. Paralogous and homologous genes, including pseudogenes, result in multiple alignments of presumably different transcripts to the same genomic region.

We used the program BLAST to integrate HINT indices into the human genome draft. Multiple alignments were frequently found between one transcript and several genomic clones, presumably resulting from the splicing of the transcript, or the sequence redundancy in the genome draft. Alternatively, multiple alignments could be caused by the sequence homology of paralogous and homologous genes or pseudogenes. Thus, the determination of exact associations among genomic clones at the sequence level was essential to resolving these three possibilities.

We integrated the BLAST results into the genomic contigs assembled on the basis of the restriction fingerprints (Marra et al. 1997). Evidence of splicing was first scored by retaining only the nonoverlapping HSPs from the same transcripts spanning on the same genomic contig. The same process was applied to the HSPs on the genomic sequences (see Methods). The remaining overlapping HSPs were removed, which presumably represented paralogous sequences due to gene duplication or homologous sequences representing evolutionally conserved functions. This appeared more true for the overlaps to be paralogs when more than one pair of the overlapping HSPs were from the same transcripts and their underlying genomic contigs were physically nearby.

Nearly identical tandem genomic duplications occur with appreciable frequency (Eichler 1998). A paralog was suspected if more than one pair of the overlapping segments was from the same transcript and their underlying genomic clones were physically nearby. This was made possible by integration of the columns containing the contigs, clones, and transcripts along with their positions and orientations in the same

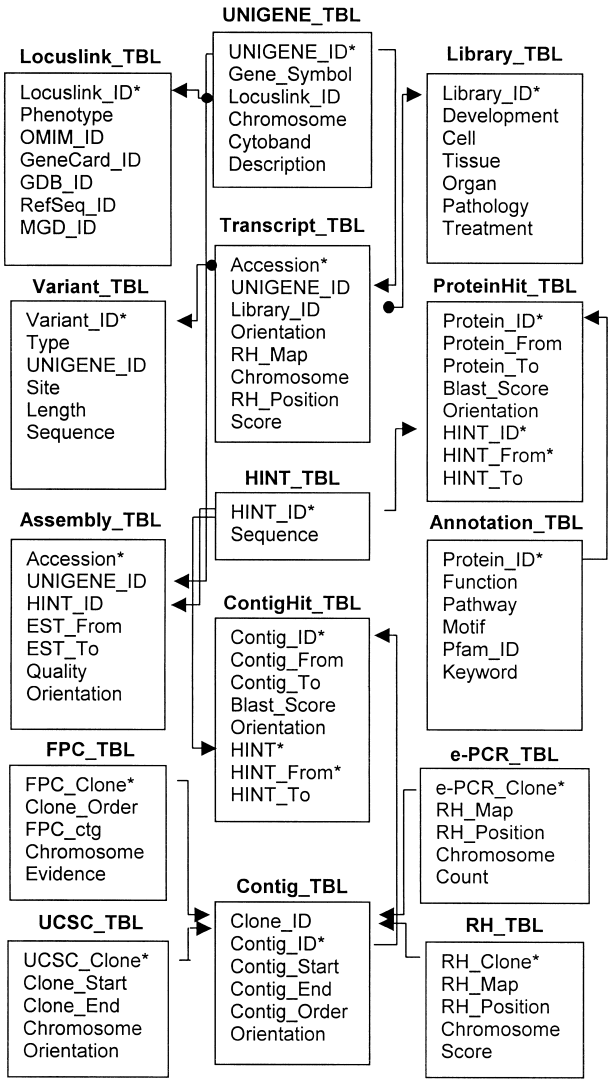

Figure 3 The database schema for the integration and mining of functional and mapping information of human transcript consensus. The UNIGENE_TBL and Locuslink_TBL were derived from the original UNIGENE and Locuslink data sets, respectively. The Transcript_TBL was created by the integration of UNIGENE and Genemap'99 databases. The cDNA library information in the Library_TBL was standardized further into appropriate categories: Development (e.g., embryonic stages), Cell, Tissue, Organ, Pathology (e.g., tumor), and Treatment. The Assembly_TBL defined the start and end positions and the orientation for each transcript in its contig, and its Smith–Waterman score compared with the consensus. All consensus sequences were stored in the HINT_TBL. The potential splicing variants were represented in the Variant _TBL by the site and length of the insertion or deletion for a given variant. The ProteinHit_TBL and ContigHit_TBL were generated by BLASTing the transcript consensus with the protein (SWISSPROT, PIR, and TrEMBL), and genomic (Ensembl) sequence databases, respectively. A majority of the functional annotations in the Annotation_TBL were derived from the SWISSPROT data set, supplemented with the key words from the PIR and TrEMBL protein entries. The Contig_TBL, based on the Ensembl database, was used to order and position individual sequencing contigs (see Methods). Mapping information for individual clones was integrated by use of four different maps: the assembled genomic contigs (UCSC_TBL); the fingerprint map (FPC_TBL); and the radiation hybrid maps (RH_TBL and e-PCR_TBL). (*) Primary or joint keys for each table; (arrows) one to many relationships; (arrows with closedcircles) one to one relationships.

spreadsheet (generated by the relational database). We depended largely on the following four criteria to specifically place a potential paralog: first, the continuity of the splicing pattern for all the HSPs involved; second, the total length and sequence identity for each of the HSP involved; third, the genomic contigs ordered in the fingerprint map; and finally, the HSPs had to be at least partially different (sequence diverging). In fact, we considered identically spliced and overlapped HSPs as the results of overlapping genomic clones, resolved by retaining only the HSPs on the longest genomic clone. We did not have evidence that human paralogous genes were ~100% identical at least in many of the cases.

To compensate in part for any of the potential errors (<5%) existing in our transcript consensus, we used the default penalty score (−4 for N) in the program BLASTN for sequence mismatch. Although ambiguities in transcript placements could be largely resolved on the basis of the HSPs spliced on the same genomic contig, more stringent search strategies might be used in the future (more penalties to the sequence mismatch as well as decreasing the sensitivity). This might be especially beneficial when paralogs or homologs had to be distinguished at the sequence level.

Because ~50% of the genomic clones in the first draft of human genome had not been mapped on either the GB4 or G3 radiation hybrids in the Genemap'99, we performed a genomewide e-PCR experiment using the available PCR primer information from the current RHdb. A total of 104,026 sequence contigs, representing 19,991 of the genomic clones, were mapped with at least one RH marker. We scored the likelihood of a putative positional placement for a genomic clone by counting the number of consistent STSs mapped on the clone. We integrated this new mapping information with Genemap'99. When an inconsistent mapping result was observed, the location scores in Genemap'99, the number of consistent e-PCR results in this study, and the fingerprint map were integrated to provide the most likely placement. The remaining 2332 clones which were not mapped by either the Genemap'99 or the e-PCR results of this study had at least fingerprint mapping information, thus limiting the complexity that must be overcome to establish possible sequence overlaps.

We have observed at least seven types of errors during our transcript mapping process: (1) Genemap'99 RH assignments for ESTs; (2) cytogenetic or chromosomal assignments in UNIGENE; (3) Gene-

map'99 RH assignments for genomic clones; (4) e-PCR map; (5) fingerprint map (FPC); (6) GoldenPath sequence contigs; and (7) misplacement due to BLAST.

We tested the mapping consistency by the integration of multiple and largely independent maps available for each HINT consensus as illustrated in Figure 4. We scored the errors on the basis of the inconsistency of a mapping assignment compared to a majority of other maps for a given consensus. The determination was also facilitated by incorporation of the likelihood scores for the following four maps: Genemap'99 (the number of consistently mapped ESTs within a UNIGENE cluster), e-PCR (the number of consistently mapped STSs within a genomic clone), FPC (the original supporting evidence), and BLAST (evidence of splicing). Thus, transcripts containing more than one mapped EST would be weighted more than other maps when having to discern more than one mapping inconsistency for a given transcript consensus. The same rule also applied to the genomic clones on which more than one STS could be consistently mapped. Errors due to BLAST placement were scored when the transcripts and the genomic clones were consistently (>2 evidence each) mapped to two different chromosomes.

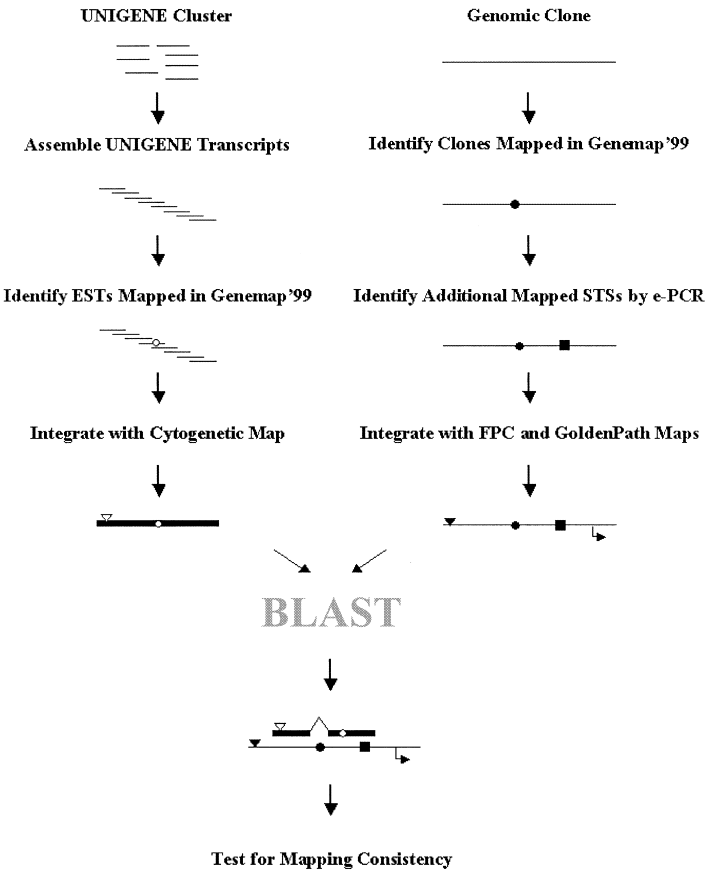

Figure 4 The data flow in mapping human transcript consensus sequences. UNIGENE clusters were assembled into transcript consensus sequences. The RH mapping information for the individual ESTs and other cDNAs available from Genemap'99 was incorporated into the transcript consensus sequences, along with the cytogenetic and/or chromosomal assignments available in the original UNIGENE data set. Mapping information for each genomic clone was compiled on the basis of the integration of four largely independent maps: Genemap'99, e-PCR and fingerprint maps, and GoldenPath sequence contigs. The resulting transcript consensus was BLASTed against the mapped genomic clones. The final transcript map was substantiated by the integration of multiple maps available for both the transcripts and the genomic clones. (Open circles) Genemap'99 EST RH assignments; (open inverted triangles) UNIGENE cytogenetic or chromosomal assignments; (closed circles) Genemap'99 RH assignments for genomic clones; (closed boxes) additional mapped STSs; (closed inverted triangles) fingerprint-mapping assignments; (bent arrows) GoldenPath positional assignments.

We systematically estimated the error rates for the currently available mapping procedures using a set of 14,469 UNIGENE consensus sequences (Fig. 5 and Table 3, Category I). It was indicated that the current mapping information was at least 95% correct, with the physical maps including the fingerprints and the genome assemblies being slightly higher in errors. Interestingly, the error rate decreased in the recent release of the genome assembly compared with its first version a few months ago (Fig. 5). Because GoldenPath contigs were built largely on the basis of the fingerprint map of genomic clones, we speculated that further FPC improvement would significantly increase the accuracy of the resulting sequence assemblies. All of the misplacements due to BLAST were singleton HSPs containing more than one mapped EST. Thus, evidence of splicing appeared crucial to bring a transcript and a genomic clone together. We speculated that the observed 1% error rate due to BLAST placement might be reduced further with more stringent BLAST parameters.

Mapping errors can be significantly reduced when multiple maps are to be integrated (<0.2%). It is important to note that because only the inconsistencies in the chromosomal assignments were tested, an additional number of finer mapping errors might still exist in these data sets. Furthermore, other hidden errors, which were not revealed due to incomplete mapping information, could also be found for the transcripts of Categories II–IV (Table 2). Finally, at least 6350 of consensus transcripts (8%) were still not mapped. It is thus conceivable that a substantial amount of genomic sequence information remains to be generated by the Human Genome Project.

Therefore, this transcript map was not a perfect and final draft for all human genes; rather it integrates multiple largely independent mapping information to substantiate an assignment, providing both supporting and conflicting evidence

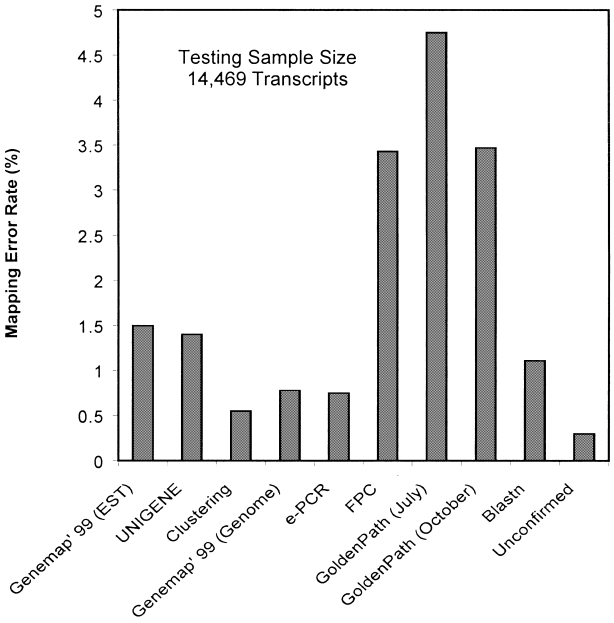

Figure 5 Estimated mapping error rates by the integration of multiple mapping data sets. A total of 14,469 transcript consensus sequences all containing the following mapping information were tested: the Genemap'99 for the ESTs contained in each of the clusters; the original UNIGENE cytogenetic or chromosomal assignments; the Genemap'99, fingerprint, and GoldenPath contigs (July and October versions), and the e-PCR maps for the genomic clones contained within the transcript consensus. Mapping errors were scored by an inconsistent chromosomal assignment of one map compared with five other maps. Nonspecific BLAST alignments were scored when the transcripts and the genomic clones putatively harboring the transcripts could be consistently mapped to different chromosomes. Errors occurring during the original UNIGENE clustering were detected when two distinct and consistent RH mapping assignments were found within the same UNIGENE clusters. Unconfirmed mapping errors were scored when more than one inconsistent mapping assignment was found for a given transcript.

(Table 4). Our final map was created based on 59,500 HINT consensus sequences, each substantiated by at least three supporting criteria (Table 4 and Supplementary Table 14 at http://pandora.med.ohio-state.edu/HINT).

A significant overall difference for transcript density was observed among all the human chromosomes (Fig. 6). The average transcript number was estimated to be about 1 per 50 kb. Human chromosomes 19 and 17 were highest in transcript density, whereas the two human sex chromosomes and chromosomes 21 and 13 were lowest. These observations were consistent with the previous gene density estimations (Schuler et al. 1996).

We were able to use the human genome draft to integrate a majority of the human transcript consensus. Such information, placing human genes and their functions in a complete positional context, can be very useful in finding candidate genes and genes in close vicinity. As the human genome draft is quickly evolving, more placements of genomic clones and finished contigs will become available, allowing more genes to be accurately mapped and the human genome to be systematically studied.

METHODS

Assembly of UNIGENE Clusters

The HINT database was built upon the most recent release of UNIGENE by NCBI, in which untrimmed vector, linker, ribosomal, mitochondrial, low-complexity sequences, repeats, and other external contaminants have already been removed. NCBI has also attempted to validate the sequence associations among nonoverlapping 5′ and 3′ ESTs, requiring at least two sets of sequencing reads from a single cDNA clone to be present to qualify as belonging to the same cluster. Our initial assembly was performed with the UNIGENE build 111 using Phrap, a sequence assembly program kindly provided by Phil Green (http://www.phrap.org). We first set out to identify any chimeric sequences within each UNIGENE cluster by using stringent alignment parameters of Phrap, with gap initiation and extension penalties both set at −10, and other parameters set at default. Under such stringent alignment criteria, deletion or insertion of sequences as short as 20 bp can be detected. The output of Phrap was parsed to identify the ESTs with deletions or multiple sites of nonoverlapping alignments. For an EST to qualify as a potential splicing variant, the deletional or insertional event had to be internal to the sequence, occurring at least >50 bp away from both its termini and with the insertion or deletion being at least 40 bp long. All chimeric ESTs were removed from each cluster before sequence assembly.

The Phrap system takes sequence quality information into account for assembly. Based on the error estimation for six sequencing projects (Richterich 1998), we automatically assigned Phred quality score 10 (corresponding to an error probability of 10%) to the first 30 bp for all ESTs that were longer than 500 bp, with positions 31–300 at 25 and the remainder at 15. A Phred quality value of 45 was assigned at all positions for the 37,400 known mRNAs and GenBank CDSs included in the UNIGENE database. This was not applied for the ESTs that were shorter than 500 bp. In this case, a Phred quality score of 15 was assigned to all positions downstream of the first 30 bases. The program cross_match, kindly provide by Phil Green (http://www.phrap.org), was used to assure that no internal association existed among multiple contigs generated from the same cluster. Default parameters of cross_match were used for this purpose. The following Phrap parameters were used for assembly: mismatch penalty −2, gap_init penalty −4, gap_ext penalty −4, vector_ bound 0, repeat_stringency 0.98, and force_high. Once a low-quality contig was detected, or multiple contigs generated from the same clusters were internally related, the entire cluster was reassembled under more stringent conditions (mismatch penalty −6, gap_init penalty −12, and gap_ext penalty −12). Each new cluster was assembled in the same fashion until a set of high-quality contigs was generated. The organization, assembly, quality control, and database integration were automated by use of a combination of shell and Perl scripts developed in-house. Computing was carried out on 5 DEC Alpha workstations (500 MHz, 1 GB of RAM each) running Digital UNIX 4.0D. Data were parsed and stored in a relational database (see Assembly_ TBL in Fig. 3) powered by the Sybase

Table 3. Assessment of Our Transcript Mapping Strategy by the Integration of Multiple Mapping Datasets

Mapping Categories	Genemap'99 (EST)	UNIGENE	Clustering	Genemap'99 (Genome)	e-PCR	Fingerprint	GoldenPath	Transcripts Mapped	Mapping Consistent for All Available Maps	Genemap'99 (EST)	UNIGENE	Clustering	Genemap'99 (Genome)	e-PCR	Fingerprint	GoldenPath	BLAST	Unconfirmed
										Mapping Errors								
I	■	■	■	■	■	■	■	14,469	12,330	216	202	79	112	108	532	687	160	43
II	■		■	■	■	■	■	3,145	2,911	24		20	68	53	17		46	6
III		■	■	■	■	■	■	4,578	3,619		103	35	8	7	44	155		607
IV			■	■	■	■	■	38,003	37,215			44	139	115	188	267		35
IV				■	■		■	1,368	866			12	41	32		67		350
VI	■	■	■					1,801	1,750		23	3						25

Mapping criteria are divided into six categories based on the types of mapping information available for each of the transcript consensus and genomic clones, as shown in columns 2–8. The sample size for each category is listed as Transcript Mapped. Mapping errors were scored by an inconsistent chromosomal assignment for a given transcript compared to at least three other independent maps. At least three independent and consistent mapping criteria were required for a transcript to be qualified as consistently mapped. Transcripts with more than one inconsistent mapping assignment were disqualified, and labeled as unconfirmed.

Adaptive Server Enterprise (Version 11.9.2) on a Pentium-based PC running RedHat Linux 6.2. The data inheritance was checked upon each new release of UNIGENE, with only updated clusters being reassembled as follows: The sequence composition of previously built UNIGENE clusters was compared with the same cluster in the current release. The original description in the UNIGENE database was used to divide the UNIGENE clusters into known and anonymous categories. Test descriptions containing EST, ESTs, Alu, or sequences highly, moderately, and weakly similar to protein sequences were considered as anonymous genes. The RepeatMasker program was kindly provided by Arian Smith and Phil Green (http://ftp.genome.washington.edu/RM/RepeatMasker.html). The HMMER program was kindly provided by Eddy Sean (http://hmmer.wustl.edu).

Integration of Transcript Consensus into the Human Genome Draft

The June 26, 2000 version of the repeat-masked draft sequences was downloaded from http://www.ensembl.org and BLASTed against the HINT and Ewing and Green's indices by use of the BLASTN program (Altschul et al. 1990) compiled from the NCBI toolkit (6.1) on a 32-node SGI Linux/Intel Cluster, with four 550 MHz Pentium III Xeons processors and 2 GB of RAM on each node. Ewing and Green's consensus was from http://www.phrap.org/est_assembly/fasta_contigs.human.000317.tar.Z. The genomewide hit expectation value was set at $1 - e^{15}$ to filter out nonspecific high-scoring segment pairs (HSPs). Default parameters of BLASTN were used. The BLAST report was parsed into field-specific tables by use of the program MSPcrunch (Version 2.3, ftp://ftp.cgr.ki.se/pub/prog). The resulting table was processed by use of a set of Perl scripts (http://pandora.med.ohio-state.edu/HINT Scripts) by first retaining only the HSPs that were spliced from the same consensus on the same genomic contig. The same process was then applied to the HSPs on the genomic sequences, that spliced HSPs from the same transcripts were first retained followed by the singleton HSPs that were both longer and higher in sequence identity over their overlapping counterparts, resulting in a unique placement for each cDNA segment on the genomic backbone. The resulting table was saved in the relational database, establishing the relationship between the transcript consensus and the genomic contigs (see ContigHit_TBL in Fig. 3).

The BLAST results for HINT and Ewing and Green's were independently processed by use of the same Perl scripts until no overlapping transcripts and genomic segments were observed. The two processed BLAST reports were later merged and sorted according to the same genomic clones and positions. The unique, overlapped, extended, truncated, and staging HSPs were scored by comparison of the start and end positions of each HSP, allowing plus or minus 2 bp for two HSPs to be qualified as aligned. All algorithms were implemented in Perl, available from our Web site (http://pandora.med.ohio-state.edu/HINT Scripts).

Map Integration

The relative position and order for a given genomic clone were often different in different maps. Different maps also had different resolutions. For instance, a genomic clone could be precisely mapped to the base position on a particular chromosome by use of the GoldenPath contig map. Less specific information would be available from the FPC map, where only the relative order of different clones was obtained. Often this resolution could not be achieved based only on the GB4 or G3 radiation hybrid map. Thus, a relational database was used to integrate all mapping information.

The order and the orientation of individual sequencing contigs within each unfinished clone were available from ftp://ftp.sanger.ac.uk/pub/ensembl/data/mysql/contig.txt.table.gz. This table was used to both integrate the sequencing contigs into the genomic clones (see Contig_TBL in Fig. 3) and allow the transcript consensus to be associated with the genome draft (see ContigHit_TBL in Fig. 3). The remaining

Table 4. The Human Transcript Map (Example)

Chromo	Clone	From (bp)	To (bp)	GB4 (cR)	Evidence	HINT ID	Description
16	AC002040	31778068	31778186	192.52	(+0+0)	AI128170_2	ESTs
16	AC002040	31831164	31831640	192.52	(+0+0)	AI285970_4	ESTs
16p12.2	AC025273	31960092	32037384	194.27	(++0+)	X87159_30	Sodium channel, nonvoltage-gated 1, beta (Liddle syndrome)
16	AC025273	32055445	32055838	195.29	(+0+0)	AI635217_4	RETINOPATHY PROTEIN (Score: 140)
16p12	AC025273	32087870	32088023	196.52	(++0+)	AB029003_296	KIAA1080 protein
16	AC025273	32096920	32097148	195.29	(+0+0)	AI022922_3	NEURONAL THREAD PROTEIN AD7C-NTP (Score: 243)
16	AC025273	32109671	32109902	195.29	(+0+0)	AW207413_1	ESTs
16p12	AC002400	32159709	32203727	196.52	(++0+)	AB029003_296	KIAA1080 protein
16	AC002400	32194539	32194738	195.26	(+0+0)	AI609753_1	CG14939 PROTEIN (Score: 581)
16	AC002400	32215385	32215802	193.96	(++0+)	AA833716_22	ESTs
16	AC002400	32216451	32217395	193.96	(++0+)	AW361820_26	ESTs
16	AC002400	32217644	32237865	195.26	(+0+0)	AC002400_5	GLUTAMYL-TRNA SYNTHETASE (Score: 329)
16	AC002400	32251204	32264728	195.26	(+0+0)	AC002400_34	Ubiquitin-binding protein homolog human (Score: 2041)
16	AC002400	32265227	32267320	196.77	(++0+)	W22792_117	ESTs
16	AC002400	32275528	32275628	195.26	(+00+)	AC002400_120	NADH dehydrogenase (ubiquinone) 1, alpha/beta subcomplex
16	AC002400	32277889	32278513	196.17	(++0+)	H66286_10	ESTs
16	AC009043	32364421	32368017	196.77	(++0+)	W22792_117	ESTs
16	AC008870	32425556	32540868	196.77	(++0+)	AA811478_5	DYNACTIN SUBUNIT P25 (Score: 167)
16	AC009043	32434020	32437557	196.77	(++0+)	AI697625_20	ESTs
16	AC008870	32437193	32447003	196.77	(++0+)	AI697625_20	ESTs
16	AC012185	32441029	32441121	193.96	(+0+0)	AW445082_3	RETINOBLASTOMA BINDING PROTEIN 2 HOMOLOG 1 (Score: 111)
16	AC012185	32474457	32474527	193.96	(+0+0)	AA714835_9	ESTs
16	AC012185	32477696	32507184	193.96	(+0+0)	U01038_83	Polo (*Drosophila*)-like kinase
16	AC012185	32493020	32493133	193.96	(+0+0)	AA436947_7	ESTs
16p11.2	AC012185	32496241	32497643	198.28	(++0+)	X07109_116	Protein kinase C, beta 1
16	AC012185	32497057	32498921	193.96	(+0+0)	AC002302_4	SID470P (Score: 160)
16	AC012185	32499777	32500136	193.96	(++00)	AA527435_16	ESTs
16	AC012185	32503878	32505108	193.96	(+0+0)	AW294825_1	ESTs
16	AC012185	32507584	32508921	193.96	(+00+)	AI732416_20	IRE1, *Saccharomyces cerevisiae,* homolog of
16	AC012185	32517480	32519058	193.96	(+0+0)	AI676060_3	IRE1 (Score: 222)

Only a portion of the human 16p is presented. A more complete human transcript map is available as supplemental information from our Web site at http://pandora.med.ohio-state.edu/HINT Supplementary Table 14. Column 1, chromosome or cytoband information; Column 2, GenBank accession numbers for the genomic clones assembled in the GoldenPath map (July version); Columns 3 and 4, the start and end positions in base pair for each transcript consensus based on the GoldenPath chromosomal contigs (July version); Column 4, the GB4 positions for the genomic clones in cR, available from the Genemap'99 or the e-PCR map; Column 5, supporting or conflicting evidence (+ = supportive, − = conflicting, 0 = no data available) ordered by genomic and transcript radiation hybrid, fingerprint, and UNIGENE maps; Column 6, the consensus ID (HINT ID), defined by the GenBank accession number of the longest transcript followed by the number of the transcripts in the same UNIGENE clusters; Column 7, gene description derived from the original UNIGENE dataset and supplemented with additional protein homology information for the previously-anonymous ESTs. The newly annotated information can be found in the rows ending with BLAST scores.

clone–clone relationships among FPC, e-PCR, Genemap'99, and GoldenPath provided an integrated view for a clone to be mapped on different scales (see FPC_TBL, UCSC_TBL, RH_TBL and e-PCR_TBL in Fig. 3). The fingerprint (Marra et al. 1997) and GoldenPath contig maps were downloaded from http://genome.wustl.edu/gsc/human/Mapping (version June 15) and http://genome.cse.ucsc.edu/goldenPath (Versions July 17 and October 7 2000), respectively. Genemap'99 (Deloukas et

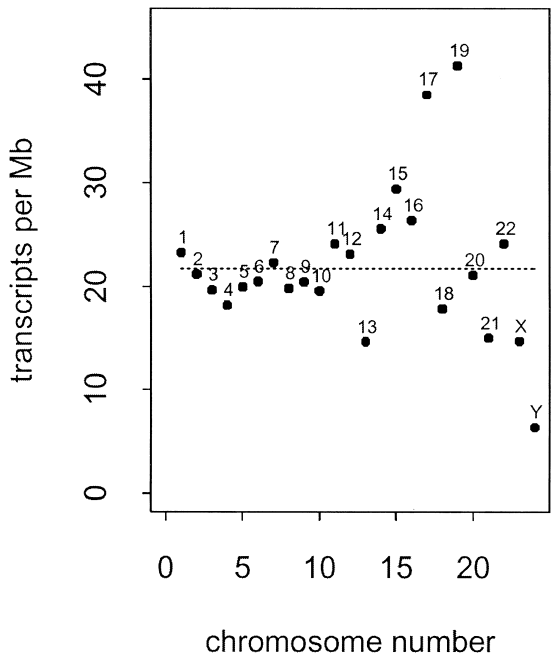

Figure 6 Estimated transcript densities on individual human chromosomes. (*y*-axis) Putative average number of consensus transcripts per 1 Mb; (*x*-axis) 24 human chromosomes. The broken line indicates the average genomewide transcript density (~22 genes per 1 Mb).

al. 1998) was downloaded from ftp://ncbi.nlm.nih.gov/repository/genemap/Mar1999. Because a substantial number of the clones in the working draft had not been typed physically with RH markers, the program e-PCR (Schuler 1997) and primers collected in the RHdb (Release 17.0, http://corba.ebi.ac.uk/RHdb) were used to map the genomic clones in the Ensembl data set. An SQL join statement was executed to integrate all related mapping information according to the identities of the genomic clones. Because our BLAST analysis was performed on individual sequence contigs, updated mapping information could be integrated readily by use of the same database schema.

Genemap'99 was integrated into UNIGENE clusters. For the UNIGENE clusters with more than one mapped EST, an averaged position was obtained and stored in the database (see Transcript_TBL in Fig. 3). Cytogenetic bands and chromosomal assignments inherited from the original UNIGENE database were integrated into the same relational database (see UNIGENE_TBL in Fig. 3). In addition, original supporting and conflicting information was also included to aid mapping assignments, such as the number of consistently mapped STSs within a genomic clone in the e-PCR map (see the ePCR_TBL in Fig. 3), the original supporting information in the FPC map (see the FPC_TBL in Fig. 3), and the score in the original Genemap'99 (see the Transcript_TBL in Fig. 3).

The relative position and order for each cDNA consensus were thus based on the integration of all the available mapping information for a given clone as illustrated in the schema (Fig. 3). To obtain a sequence level map for the transcripts, the contigs and the coordinates of the BLAST results (see ContigHit_TBL in Fig. 3) were integrated into the GoldenPath sequence map (see UCSC_TBL in Fig. 3). On the basis of such an integrated database schema, mapping information from sequence, clone, contigs, RH, and cytogenetic positions for a given transcript could be obtained through a SQL joint statement (Table 4).

Functional Annotation of Consensus Transcripts

HINT and UNIGENE consensus sequences were aligned to SWISSPROT (Release 39; Bairoch and Apweiler 2000), TrEMBL (Release 63; Bairoch and Apweiler 2000), and PIR (Release 65.01; Barker et al. 2000) by use of the program BLASTX (Altschul et al. 1990), compiled from the NCBI toolkit (6.1) on the same SGI Linux/Intel Cluster. BLAST was performed at expectation value of $1 - e^3$ to filter out nonspecific high-scoring segment pairs. All BLAST reports were compiled into field-specific tables by use of the program MSPcrunch (Version 2.3) and loaded into the underlying relational database (see ProteinHit_TBL in Fig. 3). The database interface for the Sybase Adaptive Server Enterprise was developed by use of the DBI, Sybperl, DBD:Sybase, and CGI.pm perl modules (http://www.cpan.org). Additional JavaScript and CGI scripts in Perl were developed in-house to customize the web interface. Assembled consensus sequences and singletons, along with their mapping and functional annotations, were available through BLAST analysis or query from our web site (http://pandora.med.ohio-state.edu/HINT). The UNIGENE cluster IDs, the GenBank accession numbers, STS and other markers are used to query the database. The results of different queries were linked. Embedded links to the UNIGENE, Locuslink, Genemap'99, GenBank, and SAGE databases via the Internet were made available. A simple description for the cluster, protein homology, and a profile of all the cDNA libraries contributing to the cluster were used for functional annotations. The consensus sequences were BLASTed via the Internet and retrieved as FASTA-formatted texts. Graphical representations were used to highlight the relative positions of each individual EST relative to its consensus. By use of our Web browser, the quality of assembly could be visually evaluated. The mathematical sum of percent mismatch, percent insertion, and percent deletion was represented by use of three different colors labeled on individual transcripts. Alignments with <2% of the discrepancies to the consensus were defined as excellent, whereas 2%–4% were fair, and >4% were considered questionable. In addition, potential deletion or insertion variants were also included, which could be queried along with rest of the cluster from which the variants were originally derived.

ACKNOWLEDGMENTS

We thank Albert de la Chapelle for his support and encouragement for the development of bioinformatics at The Ohio State University. We also thank Fred J. Hendler, Robert B. Chadwick, Cheryl Johnson, Paivi Peltomaki, Gail Herman, Christoph Plass, Gustavo Leone, Charis Eng, and Brad Harris for their support. We acknowledge Pete Wyckoff, Steve Gordon, Douglass Johnson, and James Giuliani at the Ohio Supercomputer Center (OSC), Adel Mikhail, Shawn Green, Jeffery Spitzner, Bob Rumpf, Eric Rentschler, and Seth Kraut at the Labbook. Com, and Mario Lauria at the Department of Computer and Information Science of The Ohio State University for their invaluable suggestions, assistance, and technical expertise, and William J. Lemon, Laura Rush, and Ming You for critical reading of the manuscript and useful suggestions.

This work was supported in part by the James Cancer Hospital and the Solove Research Foundation.

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

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 173–174.

Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S., Elliston, K.O. 1996. Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST sequence data. *Genome Res.* **6:** 829–845.

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

Ashworth, A. 1993. Two acetyl-CoA acetyltransferase genes located in the t-complex region of mouse chromosome 17 partially overlap the Tcp-1 and Tcp-1x genes. *Genomics* **18:** 195–198.

Bairoch, A., and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.

Barker, W.C., Garavelli, J.S., Huang, H., McGarvey, P.B., Orcutt, B.C., Srinivasarao, G.Y., Xiao, C., Yeh, L.S., Ledley, R.S., Janda, J.F., et al.. 2000. The protein information resource (PIR). *Nucleic Acids Res.* **28:** 41–44.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L.L. 2000. The Pfam protein families database. *Nucleic Acids Res.* **28:** 263–266.

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

Boguski, M.S., Tolstoshev, C.M., Bassett, Jr., D.E. 1994. Gene discovery in dbEST. *Science* **265:** 1993–1994.

Buetow, K.H., Edmonson, M.N., and Cassidy, A.B. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet.* **21:** 323–325.

Burke, J., Wang, H., Hide, W., and Davison, D.B. 1998. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* **8:** 276–290.

Deloukas, P., Schuler, G.D., Gyapay, G., Beasley, E.M., Soderlund, C., Rodriguez-Tomé, P., Hui, L., Matise, T.C., McKusick, K.B., Beckmann, J.S., Bentolila, S., et al. 1998. A physical map of 30,000 human genes. *Science* **282:** 744–746.

Eichler, E.E. 1998. Masquerading repeats: Paralogous pitfalls of the human genome. *Genome Res.* **8:** 758—762.

Ewing, B., and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186—194.

———. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25:** 232–-234.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175—185.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

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

Irizarry, K., Kustanovich, V., Li, C., Brown, N., Nelson, S., Wong, W., and Lee, C.J. 2000. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat. Genet.* **26:** 233–236.

Junker, V.L., Apweiler, R., and Bairoch, A. 1999. Representation of functional information in the SWISS-PROT data bank. *Bioinformatics* **15:** 1066–1067.

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

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7:** 1072–1084.

Miller, R.T., Christoffels, A.G., Gopalakrishnan, C., Burke, J., Ptitsyn, A.A., Broveak, T.R., and Hide, W.A. 1999. A comprehensive approach to clustering of expressed human gene sequence: The sequence tag alignment and consensus knowledge base. *Genome Res.* **9:** 1143–1155.

Mironov, A.A., Fickett, J.W., and Gelfand, M.S. 1999. Frequent alternative splicing of human genes. *Genome Res.* **9:** 1288–1293.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. 2000. The TIGR gene indices: Reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* **28:** 141–145.

Richterich, P. 1998. Estimation of errors in "raw" DNA sequences: A validation study. *Genome Res.* **8:** 251–259.

Schuler, G.D. 1997. Sequence mapping by electronic PCR. *Genome Res.* **7:** 541–-750.

Schuler, G.D., Boguski, M.S., Hudson, T.J., Hui, L., Ma, J., Castle, A.B., Wu, X., Silva, J., Nusbaum, H.C., Birren, B.B., et al. 1996. The human transcript map. *Science* **274:** 547–562.

Tsai, J.Y., Namin-Gonzalez, M.L., and Silver, L.M. 1994. False association of human ESTs. *Nat. Genet.* **8:** 321–322.

Wheelan, S.J., and Boguski, M.S. 1998. Late-night thoughts on the sequence annotation problem. *Genome Res.* **8:** 168–169.

Wren, J.D., Forgacs, E., Fondon, III, J.W., Pertsemlidis, A., Cheng, S.Y., Gallardo, T., Williams, R.S., Shohet, R.V., Minna, J.D., and Garner, H.R. 2000. Repeat polymorphisms within gene regions: Phenotypic and evolutionary implications. *Am. J. Hum. Genet.* **67:** 345–356.

Received September 11, 2000; accepted in revised form February 5, 2001.