# Gene Index analysis of the human genome estimates approximately 120,000 genes

Feng Liang, Ingeborg Holt, Geo Pertea, Svetlana Karamycheva, Steven L. Salzberg & John Quackenbush

**Although sequencing of the human genome will soon be completed, gene identification and annotation remains a challenge. Early estimates suggested that there might be 60,000–100,000 (ref. 1) human genes, but recent analyses of the available data from EST sequencing projects have estimated as few as 45,000 (ref. 2) or as many as 140,000 (ref. 3) distinct genes. The Chromosome 22 Sequencing Consortium estimated a minimum of 45,000 genes based on their annotation of the complete chromosome, although their data suggests there may be additional genes[4]. The nearly 2,000,000 human ESTs in dbEST provide an important resource for gene identification and genome annotation, but these single-pass sequences must be carefully analysed to remove contaminating sequences, including those from genomic DNA, spurious transcription, and vector and bacterial sequences. We have developed a highly refined and rigorously tested protocol for cleaning, clustering and assembling EST sequences to produce high-fidelity consensus sequences for the represented genes (F.L. *et al.*, manuscript submitted) and used this to create the TIGR Gene Indices[5]—databases of expressed genes for human, mouse, rat and other species (http://www.tigr.org/tdb/tgi.html). Using highly refined and tested algorithms for EST analysis, we have arrived at two independent estimates indicating the human genome contains approximately 120,000 genes.**

To assemble the Human Gene Index (HGI 5.0), we began with 1,610,947 EST sequences downloaded from the dbEST division of GenBank. These were 'cleaned' to remove contaminating sequences (see Methods, http://genetics.nature.com/supplementary_info), which eliminated 82,228 ESTs (5.1%) containing mitochondrial, ribosomal, vector, adaptor and bacterial sequences. We included 54,506 human gene sequences: 47,283 human transcripts (NP sequences) parsed through Entrez from CDS and CDS-join features in GenBank records; and 7,223 curated expressed transcript (ET) sequences from the TIGR expressed gene anatomy database (EGAD; http://www.tigr.org/tdb/egad/egad.html).

These sequences were compared using FLAST, a rapid sequence comparison program based on DDS (ref. 6), and those with greater than or equal to 95% identity over at least 40 bp with unmatched overhangs less than 20 bp were placed into 69,318 clusters (316,852 EST and 4,015 gene sequences remained as singletons). Some singleton ESTs may represent rare transcripts, but most probably represent contaminating sequences including ESTs with unspliced introns, genomic DNA and spurious transcription. Sequences comprising each cluster were assembled using CAP3 (ref. 7), which has been demonstrated to be very tolerant to EST sequencing errors, faithfully reconstructing transcript sequences without creating additional consensus sequences. This assembly produced 75,424 tentative human consensus (THC) sequences containing 1,188,592 ESTs and 47,834 NP and ET sequences; an additional 340,127 singleton ESTs were identified and eliminated from further analysis. (The statistics for the latest build are available, see Table 1, http://genetics.nature.com/supplementary_info/.)

Our first estimate of the number of human genes is based on the observation that many of the annotated genes in GenBank are not represented in the EST data. Of the 54,506 NP and ET sequences, 39,798 appear in 10,224 THCs containing ESTs, 8,036 appear in 1,769 THCs containing only NP sequences, and 6,672 remain as singletons (see Table 2, http://genetics.nature.com/supplementary_info/). This suggests that the gene sequence data represent 18,665 (10,224+1,769+6,672) unique genes, of which only 10,224 (54.8%) have been sampled by EST sequencing projects. If this trend holds true for the remainder of the genes, it implies that the 73,655 THCs that contain ESTs represent only 54.8% of the total number of genes, suggesting an upper limit of approximately 134,000 human genes.

A number of factors may influence this estimate by 'splitting' ESTs from the same gene into multiple THCs. One potential source of splitting is the relatively low quality of the EST data. CAP3 outperforms other assemblers in its ability to produce a single, high-fidelity consensus sequence from ESTs with 1–8% error rates at various depths of coverage (F.L. *et al.*, manuscript submitted). Consequently, we do not believe misassembly to be a major source of error. Alternative splicing, however, may generate multiple transcripts from a single gene, and more than one of these may be represented in the ESTs. Additionally, not all ESTs represent the 'correct' mature mRNA; some ESTs may contain artefacts such as unspliced introns. To refine our approximation, we estimated the amount of splitting that may have occurred at any stage of the process. The THCs were clustered, grouping sequences that share greater than or equal to 96% identity over 100 bp or more (criteria that should be sufficient to group most gene-family members and alternative splice forms). Of the 75,424 THCs, 20,299 were grouped into 6,966 clusters. This suggests that HGI is as much as 1.22-fold redundant. Using the same assumptions as above, this suggests a lower estimate of 110,000 genes.

The recently completed sequence of chromosome 22 provided an opportunity to validate our estimate. The Chromosome 22 Sequencing Consortium identified 679 genes (and pseudogenes) that were identical to known genes or contained a region similar to known genes in humans or other species, or similar to only ESTs (ref. 4). Given that chromosome 22 represents approximately 1.1% of the genome, these data suggest the genome contains, at a minimum, 62,000 genes. As previously noted[4], EST radiation-hybrid mapping data[8] imply that chromosome 22 is gene-rich by a factor of 1.38 relative to the genome average, giving a revised minimum estimate of 45,000 genes. (This enrichment factor may be an overestimate. If we use electronic PCR (e-PCR; ref. 9) to map the most recent radiation-hybrid mapping data to the chromosome 22 sequence, we find that 715 (1.35%) of the 52,825 localized markers

*The Institute for Genomic Research, Rockville, Maryland, USA. Correspondence should be addressed to J.Q. (e-mail: johnq@tigr.org).*

map to chromosome 22, implying 1.23-fold enrichment relative to the genome average and consequently a minimum estimate of 50,000 genes. Radiation-hybrid mapping data are available in e-PCR format at ftp://ncbi.nlm.nih.gov/repository/genemap/Mar1999/genemap99.sts. The e-PCR program can be found at ftp://ncbi.nlm.nih.gov/pub/schuler/e-PCR/.) The consortium concedes that the data from the annotated genes represent a lower limit on the actual number of genes. Gene prediction programs identified a further 325 candidate genes, of which they assume 100 are correct. The presence of 553 CpG islands and their correlation with the annotated genes suggests the existence of 271 additional genes. These 371 additional genes raise the estimate of the minimal gene content to 69,000.

We searched the 679 annotated genes in chromosome 22 using HGI 5.0 and identified 331 genes (48.8%) that have greater than or equal to 95% identity with one or more THCs having more than 80% of the THC length. This is consistent with our observation that approximately one-half of the identified genes have EST support. It also provides an independent assessment of the redundancy in the THCs. The 331 genes matched 459 THCs. After assembling these at high stringency, 14 of the THCs and 2 of the genes remained as singletons. This implies that 445 THCs actually represent 329 distinct genes, or a level of redundancy of about 1.35-fold, slightly greater than our previous estimate. After masking repeats using RepeatMasker (A.F.A. Smit and P. Green, unpublished data; http://repeatmasker.genome.washington.edu), we searched HGI against chromosome 22. We identified 1,326 THCs that had high-scoring hits (>98%, >150 bp), including 1,153 that did not hit genes annotated in the published sequence. Among these are completely sequenced genes that had previously been mapped to the chromosome (but had been omitted from the annotation) as well as THCs that contain only ESTs. The average length of these 1,153 THCs was 793 bp, of which 492 bp matched an average of 99.2% identity over an unspliced genomic length of 1,879 bp. Examples of alignments between these THCs and the chromosome 22 sequence are shown (see Fig. 1, http://genetics.nature.com/supplementary_info/).

The 1,326 THCs mapped to the chromosome allow us to estimate the number of genes in the genome. If we assume, as the data from the annotated genes imply, that the THCs are 1.35-fold redundant, there is EST support for approximately 980 genes on the chromosome. If we assume, as previously, that only 54.8% of genes are represented by ESTs, this suggests that the chromosome contains as many as 1,790 genes. Assuming that the chromosome represents 1.1% of the genome, but that it is 1.38-fold gene rich, the genome would contain approximately 118,000 genes. It is noteworthy that this is within 3.2% of the arithmetic mean of our previous estimates.

As with all estimates, one should be wary of the conclusions that are drawn and cognizant of the limitations of the supporting data. EST data may have limitations, including the presence of sequences from alternative-spliced and unspliced transcripts, spurious transcription, DNA contamination and cryptic poly(A) sequences. The EST sequences represent the most extensive (and widely used) survey available of transcribed genes, and consequently our estimate must in some way rely on the EST data. More than 26% of the initial pool of 1,610,947 ESTs were eliminated in our cleaning and assembly process. The remaining ESTs represent the highest quality sequences available, and have all been sequenced independently multiple times or match independently annotated genes. It is this refined set, represented in the 73,655 EST-containing THCs, that was used in our estimates. This provides a measure of confidence in the THC assemblies and the estimates derived from them.

Finding genes in genomic sequence is a significant challenge, and the EST data represent a substantial resource that can be applied to this problem. The TIGR Gene Indices provide a reliable reduction of the EST data and can simplify annotation by providing fewer, more accurate sequences that can be queried against genomic sequence. Our estimate of the gene content of the genome represents a hypothesis that can be used to stimulate further analysis of the genome sequence. Our analysis of chromosome 22 suggests that, as the consortium concedes, there are likely many more genes awaiting discovery within the sequence. With the imminent completion of the sequence of the human genome, the challenge will be to use all available resources to accurately catalogue and characterize the genes encoded within it.

**Acknowledgements**

*We thank the remaining members of the TIGR Gene Index Team, T. Hansen and J. Upton; A. Glodek for database development efforts; M. Heaney and S. Lo for database support; V. Sapiro, B. Lee, S. Gregory, R. Karamchedu, C. Irwin, L. Fu and E. Arnold for computer system support; and C. Ronning, R. Buell, J. White, and C.M. Fraser for thoughtful comments and suggestions. This work was supported by a grant from the U.S. Department of Energy.*

Received 10 March; accepted 2 May 2000.

1. Fields, C., Adams, M.D., White, O. & Venter, J.C. How many genes in the human genome? *Nature Genet.* **7**, 345–346 (1994).
2. Green, P. Interpreting the genome. Presentation at *Bridging the Gap Between Sequence and Function*, Cold Spring Harbor Laboratory, NY, September, 1999.
3. Scott, R. The future in understanding the molecular basis of life. Presentation at Eleventh International Genome Sequencing and Analysis Conference, Miami, 1999.
4. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
5. Quackenbush, J., Liang, F., Holt, I., Pertea, G. & Upton, J. The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* **28**, 141–145 (2000).
6. Huang, X., Adams, M.D., Zhou, H. & Kerlavage, A.R. A tool for analyzing and annotating genomic sequence. *Genomics* **46**, 37–45 (1997).
7. Huang, X. & Madan, A. CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
8. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
9. Schuler, G.D. Sequence mapping by electronic PCR. *Genome Res* **7**, 541–550 (1997).