Analysis of expressed sequence tags indicates 35,000 human genes

Brent Ewing & Phil Green

The number of protein-coding genes in an organism provides a useful first measure of its molecular complexity. Single-celled prokaryotes and eukaryotes typically have a few thousand genes; for example, *Escherichia coli*[1] has 4,300 and *Saccharomyces cerevisiae*[2] has 6,000. Evolution of multicellularity appears to have been accompanied by a several-fold increase in gene number, the invertebrates *Caenorhabditis elegans*[3] and *Drosophila melanogaster*[4] having 19,000 and 13,600 genes, respectively. Here we estimate the number of human genes by comparing a set of human expressed sequence tag (EST) contigs with human chromosome 22 and with a non-redundant set of mRNA sequences. The two comparisons give mutually consistent estimates of approximately 35,000 genes, substantially lower than most previous estimates. Evolution of the increased physiological complexity of vertebrates may therefore have depended more on the combinatorial diversification of regulatory networks or alternative splicing than on a substantial increase in gene number.

In contrast to the situation with more compact genomes, completion of the human genome sequence will not immediately provide definitive gene counts because *de novo* identification of human genes from finished genomic sequence is not reliable[5], and will be even less so for 'rough draft' sequence in which the relatively high error rate may prevent discrimination of genes from pseudogenes (which are relatively frequent[5]). To estimate the number of human genes, we adapted a method originally applied to *C. elegans*[6] that involves determining the overlap between two independently derived sets of gene sequences. The first set should contain essentially full-length sequences for an unbiased representative sample of genes from the genome; the second set need not be an unbiased sample, and it may have sequences that are incomplete or redundant provided they are accurate enough to reliably determine matches to genes in the first set. Under these assumptions, if the first set has n_1 genes, which constitute a fraction f G of the total number G of genes in the genome, then it should match a corresponding fraction f n_2 of the n_2 sequences in the second set. So if m_2 is the number of sequences in the second set that are matched by the first, G may be estimated as $G=n_1/f=n_1n_2/m_2$. In contrast to some other methods for estimating G, no assumption about the size of the genome is required here. For the first (representative) set of gene sequences, we use either the genes in the 33.5-Mb sequence[5] of chromosome 22 (estimated to number ~680) or a set of 7,662 genes obtained by clustering mRNA sequences obtained from GenBank. For the second set, we constructed contig sequences by assembling approximately 1 million ESTs from 168 cDNA libraries (generated at the Washington University Genome Sequencing Center[7]). These contigs do not randomly sample the set of all genes, because expression level and the spectrum of tissues from which the libraries were derived affect the probability that a particular gene is represented; however, random sampling is not required for our calculation.

To eliminate the artefactual and contaminant sequences in the ESTs (refs 7,8), we determined the high-quality part of each read (using phred (refs 9,10) quality values) and used only those parts of the contig sequences that were confirmed by the high-quality parts of reads from at least two independent clones. There were 62,064 confirmed, high-quality contig sequences, averaging 540 bases in length. Of these, 43,278 include the putative 3′ end of a cDNA clone; there can be several such contigs for a single gene due to internal priming during the construction of cDNA libraries (the normalization procedure used for some libraries in fact tends to enrich for such events[11]), alternative splicing or the presence of multiple polyadenylation sites for the same gene.

We compared the 3′ EST contigs to chromosome 22 and to the mRNA clusters using stringent criteria for a match, and used the results to estimate the total number of genes in the genome as described above (Table 1). The two comparisons yielded similar estimates of less than 35,000 genes. Similar results (data not shown) were obtained if we used all contigs or 'scaffolds' consisting of contig groups linked by forward-reverse read pairs in place of the 3′ contigs. In the comparison of 3′ contigs with mRNAs, 6,169 of the 7,662 mRNA clusters were matched, so we estimate that the EST contigs may represent approximately 80% of all genes.

In a 3-gigabase genome, the presence of 35,000 genes would imply an average density of 1 gene per 85 kb. This is generally consistent with what is currently known about how genes are distributed in the genome[12]. Genes in A+T-rich regions, which constitute over 90% of the genome, tend to have very large introns and frequently extend over hundreds of kilobases. Gene density is significantly higher in G+C-rich regions, but these appear to comprise less than 10% of the genome. Our estimate is also consistent with older estimates of 30,000–40,000 genes derived from mRNA reassociation studies (for review, see ref. 13), but is less than other published estimates[14,15].

Table 1 • Gene estimates from sequence comparisons

Reference set	Genes (n_1)	3′ EST contigs (n_2)	Matched contigs (m_2)	Predicted no. genes ($n_1 n_2/m_2$)
Chromosome 22	680	43,278	848	34,700
mRNAs	7,662	43,278	9,859	33,630

Department of Molecular Biotechnology, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to P.G. (e-mail: phg@u.washington.edu).

The main caveat to our calculation is our assumption that the chromosome 22 genes and mRNA clusters are representative of all genes, and in particular that these sets are not biased by expression level. The mRNA set almost certainly has some expression-level bias, because early cDNA cloning efforts were more likely to succeed with highly expressed genes. However, the current size of this set, the diversity of research avenues leading to cloned genes and the improved methods for cloning rarely expressed messages all make this set much more representative than it was a few years ago. Regarding the chromosome 22 gene set, it is tempting to assume that any chromosome or large genomic region is likely to be typical of the genome as a whole, but failure of that assumption caused the application of the current method to underestimate (by ~25%) the number of genes in *C. elegans*[6]; the genomic regions selected for comparison were in the gene-rich central chromosomal clusters, which were subsequently found[3] to be enriched for highly expressed genes. Although no similar expression bias has yet been observed for entire human chromosomes, it should be noted that chromosome 22 is more G+C rich and gene dense than the genome as a whole[5], and these properties may correlate with expression level. In the event that this chromosome and the mRNA set both have an expression bias relative to the whole-genome average that is comparable to that seen for the *C. elegans* gene-rich clusters, our estimate could be too low by a comparable fraction, in which case the true number of genes may be closer to 50,000. More generally, it is possible that there are substantial numbers of rapidly evolving, rarely expressed genes (these two properties being generally correlated[3,16]) under-represented or underannotated on chromosome 22 and under-represented in the mRNA and EST sets that are effectively invisible to our calculations. In this case, however, our estimate should still give a reasonable picture of the number of genes likely to be readily identifiable using current methods and resources.

The fact that humans apparently have less than twice as many genes as the 959-cell nematode *C. elegans* is notable. The protein 'parts list' of an organism may be substantially larger than its set of genes due to post-translational modifications and alternative splicing, which for humans at least is now believed to be relatively frequent[17]. Thus the relative molecular complexities of the two organisms may differ by a substantially greater ratio than the gene numbers would indicate. We speculate, however, that the greater physiological complexity of vertebrates has instead been generated primarily by regulatory combinatorics, particularly the diversification of gene regulatory networks through signals encoded in the genome.

Methods

Details regarding EST assembly, mRNA clustering, chromosome 22 gene count and sequence comparisons, as well as the confirmed contig sequences, are available (http://www.phrap.org). In brief, we obtained chromatograms for 1,043,599 human ESTs from the Washington University Genome Sequencing Center, derived base calls and quality values (log-transformed error probabilities) using phred (refs 9,10), removed cloning vector sequences, assembled the ESTs using phrap, and identified the portions of contig sequences confirmed by high-quality segments from at least two ESTs. We eliminated probable chimaeras and a small number of *E. coli* and mouse contaminants.

By clustering mRNA sequences from GenBank, we obtained a set of 7,662 'genes', defined as mRNA clusters in which at least one mRNA has an annotated full-length coding sequence at least 100 bases long and at least 20 bases of 3′ untranslated sequence.

For comparisons of the contig sequences with chromosome 22 and with the mRNA clusters, we only accepted matches in which the aligned regions were at least 98% identical and did not lie within a repeat. For the chromosome 22 comparison, we additionally required the matches to (collectively) constitute at least 90% of the contig length. For the mRNA comparison this was relaxed to a minimum of 100 aligned bases because some of the mRNAs were incomplete or represented different alternatively spliced forms of the same gene.

Technical comments on previous gene-number estimates. Using a method similar to ours in which EST clusters were compared with mRNAs, an estimate of 60,000–70,000 genes was obtained[15]. A major difference with our calculation is that these authors included clusters consisting of single unconfirmed ESTs. We believe this may have spuriously inflated the estimate due to the inclusion of contaminant, artefactual or inaccurate ESTs. The cDNA library normalization process tends to enrich for contaminants and aberrant clones, and even if they represent only a tiny percentage of the total EST set, they can constitute a large fraction of the clusters. For this reason, we used conservative criteria for deriving contig sequences. Although many of the unconfirmed ESTs are undoubtedly real, eliminating them does not affect the validity of our calculation, because the set of EST contigs is not assumed or required to be a random sample of all genes, and part of it can therefore be removed without biasing the estimate. We believe that even some of what we consider 'confirmed' EST contigs are cases in which a single contaminant or artefactual clone spuriously confirms itself because it has been duplicated due to library amplification or well-to-well or plate-to-plate cross-contamination. Such cases may account in part for recent unpublished estimates of a very high gene number based on counting EST contigs assembled from proprietary databases[18]. A low rate of such artefacts still may produce large absolute numbers when the data set is very large.

It has been estimated[14] that there are 45,000 unmethylated CpG islands in the human genome, which, under the assumptions that 56% of genes have an island and that all islands are associated with genes, extrapolates to an estimate of 80,000 genes. Several aspects of this calculation are open to question and may have inflated this estimate. First, in converting from tritiated counts to genome percentage (Table 1 of ref. 14), it was assumed that DNA fractions 1 and 2 have the same overall G+C content as the CpG-island fragments in these fractions. These fractions, however, were known to contain a significant fraction (28.6%) of non-island DNA, which may reasonably be assumed to have a G+C content equal to the genome average (40%) rather than that of island fragments (67.1%). Making this correction reduces the estimated number of islands from 45,000 to 34,200. Second, our own analyses (unpublished data) of CpG islands suggest that the estimated length of an island is somewhat sensitive to the precise definition that is applied. Using a definition more sensitive to CpG dinucleotide frequency (which should be the best indicator of methylation status) than to C+G content, we obtained an average island length of ~1.3 kb rather than 1 kb. Use of this value would reduce the island count further to 26,300 and the estimated number of genes to ~47,000. Third, there may be islands not associated with genes. For example, many transposable elements and pseudogenes are known to have CpG islands[19], some of which may be unmethylated, and as some tissue-specific genes have one or more islands associated with cryptic internal promoters that do not produce translatable transcripts[19], it seems possible that other islands are not associated with any gene. Fourth, it is possible that the size of the gene-bearing, euchromatic portion of the genome, which must be assumed in the CpG-island calculation but is irrelevant in ours, is substantially smaller than has been assumed, as in fact was the case with chromosome 22 (ref. 5).

Acknowledgements

We thank C. Wilson and A. Nichols for programming assistance. This work was supported by a grant from the National Human Genome Research Institute.

Received 16 March; accepted 2 May 2000.

1. Blattner, F.R. *et al*. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
2. Goffeau, A. *et al*. Life with 6000 genes. *Science* **274**, 563–567 (1996).
3. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
4. Adams, M.D. *et al*. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
5. Dunham, I. *et al*. The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
6. Waterston, R. *et al*. A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genet.* **1**, 114–123 (1992).
7. Hillier, L. *et al*. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6**, 807–828 (1996).
8. Wolfsberg, T.G. & Landsman, D. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **25**, 1626–1632 (1997).
9. Ewing, B., Hillier, L., Wendl, M. & Green, P. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
10. Ewing, B. & Green, P. Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
11. Bonaldo, M., Lennon, G. & Soares, M.B. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **6**, 791–806 (1996).
12. Bernardi, G. The human genome: organization and evolutionary history. *Annu. Rev. Genet*. **29**, 445–476 (1995).
13. Lewin, B. *Genes IV* 466–481 (Oxford University Press, Oxford, 1990).
14. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
15. Fields, C., Adams, M.D., White, O. & Venter, J.C. How many genes in the human genome? *Nature Genet.* **7**, 345–346 (1994).
16. Green, P. *et al*. Ancient conserved regions in new gene sequences and the protein databases. *Science* **259**, 1711–1716 (1993).
17. Mironov, A.A., Fickett, J.W. & Gelfand, M.S. Frequent alternative splicing of human genes. *Genome Res.* **9**, 1288–1293 (1999).
18. Dickson, D. Gene estimate rises as US and UK discuss freedom of access. *Nature* **401**, 311 (1999).
19. Larsen, F., Gundersen, G., Lopez, R. & Prydz, H. CpG islands as gene markers in the human genome. *Genomics* **13**, 1095–1107 (1992).