Molecular Fossils in the Human Genome: Identification and Analysis of the Pseudogenes in Chromosomes 21 and 22

Paul M. Harrison, Hedi Hegyi, Suganthi Balasubramanian, Nicholas M. Luscombe, Paul Bertone, Nathaniel Echols, Ted Johnson, and Mark Gerstein[1]

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8114, USA

We have developed an initial approach for annotating and surveying pseudogenes in the human genome. We search human genomic DNA for regions that are similar to known protein sequences and contain obvious disablements (i.e., mid-sequence stop codons or frameshifts), while ensuring minimal overlap with annotations of known genes. Pseudogenes can be divided into "processed" and "nonprocessed"; the former are reverse transcribed from mRNA (and therefore have no intron structure), whereas the latter presumably arise from genomic duplications. We annotate putative processed pseudogenes based on whether there is a continuous span of homology that is >70% of the length of the closest matching human protein (i.e., with introns removed), or whether there is evidence of polyadenylation. We have applied our approach to chromosomes 21 and 22, the first parts of the human genome completely sequenced, finding 190 new pseudogene annotations beyond the 264 reported by the sequencing centers. In total, on chromosomes 21 and 22, there are 189 processed pseudogenes, 195 nonprocessed pseudogenes, and, additionally, 70 pseudogenic immunoglobulin gene segments. (Detailed assignments are available at http://bioinfo.mbb.yale.edu/genome/pseudogene or http://genecensus.org/pseudogene.) By extrapolation, we predict that there could be up to ~20,000 pseudogenes in the whole human genome, with a little more than half of them processed. We have determined the main populations and clusters of pseudogenes on chromosomes 21 and 22. There are notable excesses of pseudogenes relative to genes near the centromeres of both chromosomes, indicating the existence of pseudogenic "hot-spots" in the genome. We have looked at the distribution of InterPro families and Gene Ontology (GO) functional categories in our pseudogenes. Overall, the families in both processed and nonprocessed pseudogene populations occur according to a similar power–law distribution as that found for the occurrence of gene families, with a few big families and many small ones. The processed population is, in particular, enriched in highly expressed ribosomal–protein sequences (~20%), which appear fairly evenly distributed across the chromosomes. We compared processed pseudogenes of different evolutionary ages, observing a high degree of similarity between "ancient" and "modern" subpopulations. This may be attributable to the consistently high expression of ribosomal proteins over evolutionary time. Finally, we find that chromosome 22 pseudogene population is dominated by immunoglobulin segments, which have a greater rate of disablement per amino acid than the other pseudogene populations and are also substantially more diverged.

Pseudogenes are disabled copies of genes that do not produce a functional, full-length copy of a protein (Mighell et al. 2000; Vanin 1985). They are of two types: First, processed pseudogenes result from reverse transcription of messenger RNA transcripts followed by reintegration into genomic DNA (presumably in germ-line cells) and subsequent degradation with disablements (premature stop codons and frameshifts) (Vanin 1985). Second, nonprocessed pseudogenes result from duplication of a gene, followed by an initial disablement if the duplicated copy is not "useful" (Mighell et al. 2000). These then also accumulate further coding disablements.

The extent of the pseudogene population in the human genome is unclear. Estimates for the number of human genes range from ~22,000 to ~75,000 (Crollius et al. 2000; Ewing and Green 2000; Lander et al. 2001; Venter et al. 2001; Wright et al. 2001). From previous reports, it is thought that up to 22% of these gene predictions may be pseudogenic (Lander et al. 2001; Yeh et al. 2001). It is important to characterize the human processed and nonprocessed pseudogene populations as their existence interferes with gene identification and prediction (particularly nonprocessed pseudogenes or individual pseudogenic exons). They are also an important resource for the study of the evolution of protein families (see, e.g., studies on the human olfactory receptor subgenome [e.g. Glusman et al. 2001]).

Here, we have performed a detailed analysis of the pseudogene populations of human chromosomes 21 and 22, which have been sequenced contiguously to high quality. This is similar in spirit to previous surveys we have performed on pseudogenes and other genomic features in other organisms (Harrison et al. 2001; Gerstein 1997, 1998; Hegyi and Gerstein 1999). We have examined the main populations and clusters of pseudogenes for the two chromosomes. Patterns of distribution of both nonprocessed and processed pseudo-

[1] **Corresponding author.**
E-MAIL Mark.Gerstein@yale.edu; FAX (360) 838 7861.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.207102.

genes indicate the existence of pseudogenic hot-spots in the human genome. In addition, we have estimated the total numbers and proportions of processed and nonprocessed pseudogenes in the whole human genome.

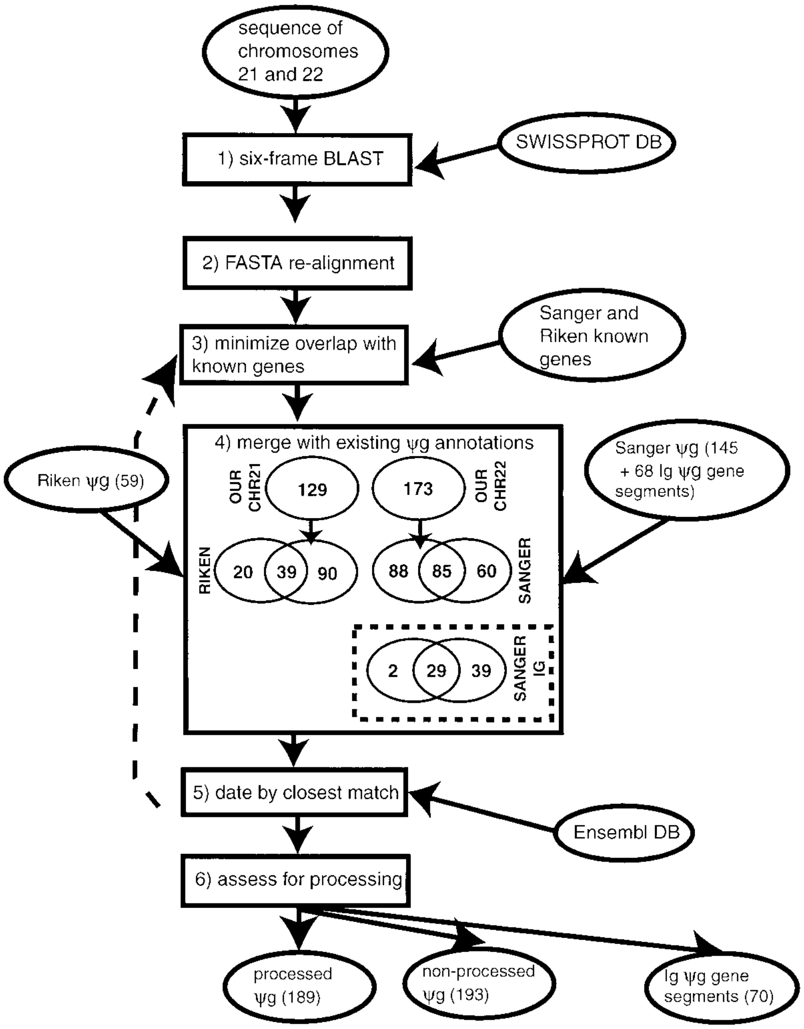


Figure 1 Flow diagram of the scheme for assignment of pseudogenes. The schematic shows the steps in assignment of pseudogenes. Ovals denote sources of data, and boxes denote operations. The term "ψg" denotes "pseudogene." The steps are as follows (described in detail in the text): (1) Six-frame blast. Comparison of SWISSPROT database to chromosome 21 and 22 genomic sequences using BLAST (Altschul et al. 1997) to find potential pseudogenic protein homologies (with stop codons in them). (2) FASTA realignment. Realignment with the FASTA package (Pearson et al. 1997) of the top-matching sequence for the potential pseudogene to find longest protein-homology fragment that has >1 disablement (frameshift or premature stop codon). (3) Minimize overlap with known genes. Overlap of putative pseudogenes with known human genes (from the Sanger center annotations for chromosome 22 and the Riken center annotations for 21) is minimized by choosing a suitable margin at the ends of pseudogenic homologies within which to ignore disablements. (4) Merge with existing ψg annotations. Pseudogene annotations from the Sanger and Riken centers are merged with those that are duplicates of these in our own set of annotations being deleted. (5) Date by finding closest matching Ensembl protein. For each pseudogene, the closest matching Ensembl human protein was found so that the pseudogene could be approximately dated. This was realigned to the genomic DNA sequence (backward step denoted by dotted arrow) and used as a replacement if it produced a longer pseudogene. (6) Assess for processing. All pseudogenes were then assessed for processing by searching for evidence of polyadenylation or extensive spans of protein homology in the absence of exon structure.

RESULTS AND DISCUSSION

We annotated both processed and nonprocessed pseudogenes, as described in Figure 1 and the Methods section. The numbers of processed and nonprocessed pseudogenes that we find are summarized in Table 1. As shown in Figure 1, there are 60 Sanger pseudogene annotations in excess of those pseudogenes that we find for chromosome 22, and 20 Riken pseudogene annotations in excess for chromosome 21.

Processed Pseudogenes on Chromosomes 21 and 22

We find a total of 189 processed pseudogenes (77 on chromosome 21, 112 on chromosome 22) (Table 1). The total for chromosome 22 is a combination of our own annotations and Sanger Center annotations for pseudogenes, whereas the total for chromosome 21 is a combination of our annotations and those obtained from the Riken genome-sequencing center. The number of processed pseudogenes for chromosome 22 relative to chromosome 21 is rather high (proportion = 112/77 ~1.45). When we remove the additional Riken and Sanger Center pseudogenes, the density of processed pseudogenes is still moderately higher for chromosome 22 relative to 21 (proportion = 83/65 ~1.30). The different numbers of processed pseudogenes for the two chromosomes is intriguing and may be related to the accessibility of genomic DNA for reintegration of a processed sequence, with chromosomes with more genes having more accessible genomic DNA because of transcriptional activity.

Non-Processed Pseudogenes on Chromosomes 21 and 22

For the counting of nonprocessed pseudogenes, we set aside the 70 κ and λ immunoglobulin gene segments on chromosome 22 as a separate population. We then have a total of 195 nonprocessed pseudogenes (72 on chromosome 21, 123 on 22). Considering all annotations, the number of nonprocessed pseudogenes on chromosome 22 relative to the number on chromosome 21 is higher (proportion = 123/72 ~1.71). As described above for processed pseudogenes, when those pseudogenes that arise only from Riken and Sanger Center annotations are excluded, the ratio of pseudogene numbers between the two chromosomes is more modest (proportion ~1.28), reflecting to less of an extent the corresponding relative gene density between the chromosomes (~2.13–2.24 for all sets of gene annotations).

Extrapolation to the Whole Human Genome

Based on our numbers of pseudogenes for chromosomes 21 and 22, we can tentatively

Table 1. Numbers of Pseudogenes and Genes

	Predictions for chromosomes			Extrapolation or prediction
Type of (pseudo)genes	**21**	**22**	**21 + 22**	**Whole human genome**
GenomeScan genes[a]	279	648 (593)[b]	927 (872)[b]	38,647 (~26,000–28,000)[c] (~20,000–25,000)[c]
Processed pseudogenes[d]	77 [65]	112 [83]	189 [148]	~6100–6600 (based on chromosome 21 data) ~8700–9400 (based on chromosome 22 data)[e]
Nonprocessed pseudogenes[d]	72 [64]	123 [83]	195 [147]	~5700–6200 (based on chromosome 21 data) ~9600–10,400 (based on chromosome 22 data)[e]
Pseudogenic Ig segments	—	70	70	—
GenomeScan genes, removing those that overlap with our pseudogenes (above)[f]	226–257	437–551	663–808	—

[a] Supplied by C. Burge and R.-F. Yeh (personal communication). See (Yeh et al., 2001) for details of the program GenomeScan.

[b] The figures in brackets give the totals omitting the Sanger (chromosome 22) and Riken (chromosome 21) pseudogene annotations.

[c] These are the ranges estimated for total number of genes by Yeh et al. (2001), first, if overparsing of gene structures is taken into account (i.e., splitting up genes into smaller genes) and, second, if the expected rates of false-positives and pseudogenes are taken into account [see Yeh et al. (2001) for details].

[d] The Riken and Sanger Centre annotations are merged with our own annotations. The pseudogenic immunoglobulin gene segments are taken out of these totals.

[e] Based on data for either chromosomes 21 and 22, omitting immunoglobulin gene segments for the nonprocessed pseudogene estimate.

[f] This does not include immunoglobulin gene segments on chromosome 22. The lower bound arises from discarding predicted genes that have any predicted pseudogenic exon. The upper bound is for predicted genes that are judged disabled in each exon, or only comprise an isolated disabled fragment (such as a processed pseudogene or an isolated disabled exon).

extrapolate to derive estimates of the pseudogene numbers in the whole human genome.

Using the total number of processed pseudogenes for either chromosome 22 or 21, we estimate the total number of processed pseudogenes in the human genome (see Table 1 footnote). The predicted ranges are ~8700–9400 (based on chromosome 22 data) and ~6100–6600 (chromosome 21) processed pseudogenes in the whole human genome. (The lower number in the range arises from using the total human genome size given by Lander et al. 2001; the higher from the size given by Venter et al. 2001.)

Also, as for processed pseudogenes, we estimate a predicted range of ~9600–10,400 nonprocessed pseudogenes in the human genome, extrapolating from the chromosome 22 data. Using the gene-poor chromosome 21, a much lower estimate is obtained (~5700–6200). Arguably, we would expect more of a relationship between nonprocessed pseudogene density and gene density than between processed pseudogene density and gene density, as the former type of pseudogenes arises from duplication of the genomic DNA. One could modify such estimates to account for lower gene density on chromosome 22 than on other human chromosomes (Dunham et al. 1999a; Lander et al. 2001; Venter et al. 2001), so the number of nonprocessed pseudogenes in the whole human genome may be even higher. However, as noted above, disregarding the κ and λ immunoglobulin variable-region gene segments, there does not seem to be a clear relationship between gene density and nonprocessed pseudogene density for these chromosomes.

Overlap with Known Sequencing Center (Riken/Sanger) Genes

During our pseudogene annotation procedure, we find that some potential pseudogenes overlap known genes; this overlap may be due to sequence alignment artifact or to a real phenomenon of discarded fragments of disabled protein homology near the extant parts of genes. As part of our assignment procedure, the allowed percentage of known gene exons overlapped by our pseudogene annotations is <5% (Methods section). Known genes are those labeled as "known" in either the Sanger or Riken annotations, that is, having a previously characterized genomic structure. For exons of known Sanger genes, this percentage of overlapped exons is 3.3%, and for Riken known gene exons, it is 2.6%. Similar levels of this overlap for exons (~3%) are found for all other (predicted) gene annotations from the Riken and Sanger centers (Hattori et al. 2000; Dunham et al. 1999a).

Overlap with Genes Predicted by GenomeScan

Genes predicted by the program GenomeScan (Yeh et al. 2001) were studied as a larger and more uniformly predicted set of genes than the gene annotations available from the Sanger and Riken centers. We examined the overlap of the pseudogene data sets with genes predicted by GenomeScan (Table 1). For the GenomeScan-predicted exons, on 21 and 22, there is only 5.2% and 6.2% overlap, respectively, of exons with pseudogenes.

Main Populations and Clusters of Chromosome 21 and 22 Pseudogenes

We now focus on the main populations and clusters of pseudogenes on chromosomes 21 and 22. To aid with our characterization of these, we determined the prevalence of InterPro motifs (Apweiler et al. 2000) in the pseudogene sets and used these to assign GO functional classes (Ashburner et al. 2000) for the processed and nonprocessed pseudogene populations (Methods section).

Power–Law Behavior of the Occurrence of Families for Pseudogenes

We examine the distribution of InterPro protein families in the processed and nonprocessed pseudogenes. For the overall groups of processed pseudogenes, nonprocessed pseudogenes and total pseudogenes, there is a power–law relationship between the number of InterPro families and the size of a family (the number of members in a family) (Fig. 2), if one removes the outliers that are labeled. These outliers are the zinc finger motif, which occurs in multiple copies (of up to 12) in a sequence, and the collagen triple helix repeat, which also occurs multiple times for the same sequence. There is also a point on the plot for the immunoglobulin (Ig) domain (which occurs in the Ig variable-region gene segments). That is, the trend can be fitted to a straight line when plotted on a log-log scale. This relationship has also been observed for genes in eukaryotes and other features of genomes (J. Qian et al. 2001). As described in the Methods section, we used the InterPro family assignments for the pseudogenes to assign a GO functional class when possible. However, the distribution of GO functional classes in the pseudogenes does not show the same sort of linearity on a log-log plot (data not shown).

The Largest Group of Processed Pseudogenes Is in the Ribosomal Class

Based on GO function classifications, the most common group of proteins in the processed pseudogenes is the ribosomal proteins, comprising 22% (42 out of 189 from GO classification) (Fig. 3). Over half of these are from the large subunit of the ribosome (60%, 25/42). This is close to the proportion of large subunit ribosomal proteins in the human ribosome (57%), implying that ribosomal proteins are evenly sampled for processed pseudogene formation. As a fraction of the estimates for the overall number of processed pseudogenes in the human genome (Table 1), this means there may be >1800 ribosomal–protein processed pseudogenes in the complete genome. In comparison, for ribosomal protein genes, there is only one actual gene for a ribosomal protein (*RPL3*) on chromosome 22, and no ribosomal protein gene on chromosome 21 (Uechi et al. 2001). Interestingly, there is a nonprocessed pseudogene for the ribosomal protein *RPL17*, on chromosome 22 (named bK440B3.1), although the "live" homolog for this pseudogene is on chromosome 18. The processed pseudogenes for the ribosomal proteins appear to be somewhat evenly distributed on the chromosomes (Fig. 4), although the sample is rather small for the two chromosomes surveyed, at the moment.

Ancient and Modern Processed Pseudogenes

We divided the processed pseudogene population by age into approximately equal-sized groups of ancient and modern processed pseudogenes using the median percentage identity value (which is 79%). This is based on the similarity of the pseudogenes to the closest matching human protein in the Ensembl database (Birney et al. 2001). Ancient pseudogenes have <79% sequence identity (see Fig. 5) to their closest matching human protein. The remainder comprise modern pseudogenes. We examined the ancient and modern processed pseudogene populations for their prevalent GO func-

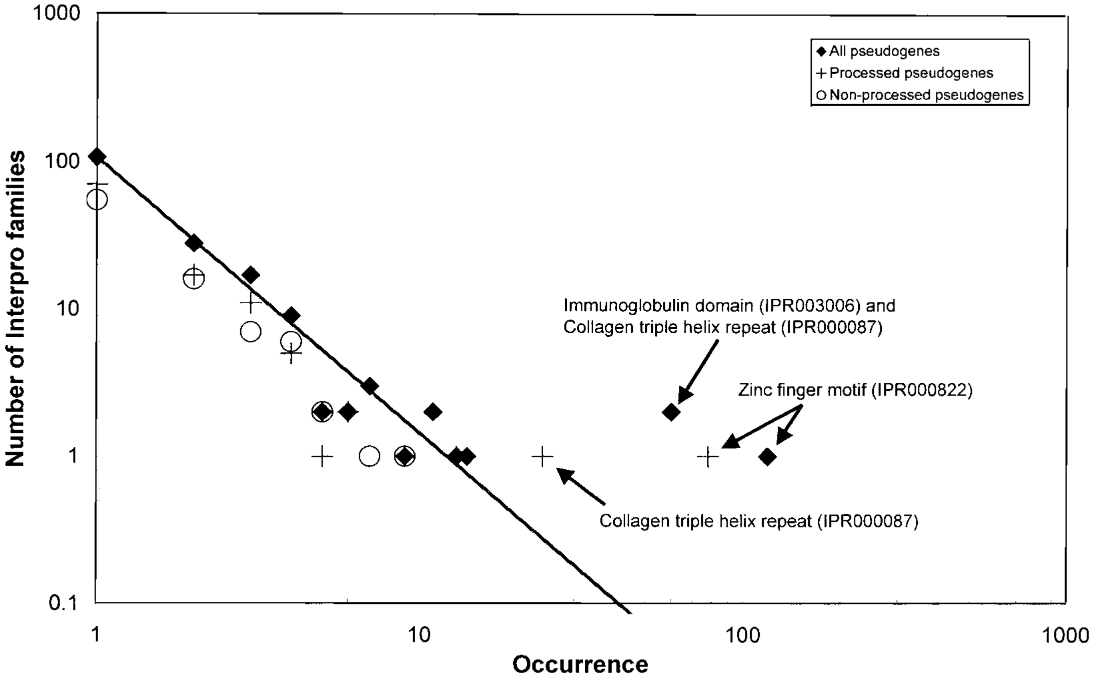
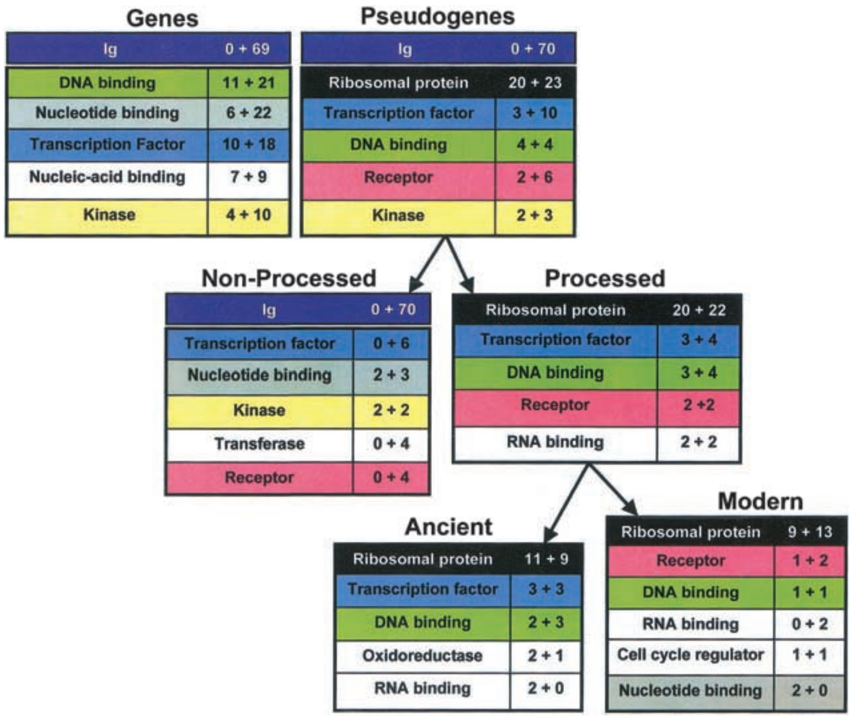


Figure 2 The relationship between the number of InterPro families for pseudogenes and their sizes shows a power–law behavior. The number of InterPro families is plotted vs. the size of a family on a log–log scale. "All pseudogenes" (processed and nonprocessed combined) are plotted with a filled diamond, processed pseudogenes with a cross, and nonprocessed pseudogenes with an unfilled circle. The straight line indicates the best least-squares linear fit to all points for all pseudogenes, except for the outliers that are labeled in the plot. This is indicative of a power–law relationship between the size of a protein family and the number of families that have this size.



Genes

Ig	0 + 69
DNA binding	11 + 21
Nucleotide binding	6 + 22
Transcription Factor	10 + 18
Nucleic-acid binding	7 + 9
Kinase	4 + 10

Pseudogenes

Ig	0 + 70
Ribosomal protein	20 + 23
Transcription factor	3 + 10
DNA binding	4 + 4
Receptor	2 + 6
Kinase	2 + 3

Non-Processed

Ig	0 + 70
Transcription factor	0 + 6
Nucleotide binding	2 + 3
Kinase	2 + 2
Transferase	0 + 4
Receptor	0 + 4

Processed

Ribosomal protein	20 + 22
Transcription factor	3 + 4
DNA binding	3 + 4
Receptor	2 +2
RNA binding	2 + 2

Ancient

Ribosomal protein	11 + 9
Transcription factor	3 + 3
DNA binding	2 + 3
Oxidoreductase	2 + 1
RNA binding	2 + 0

Modern

Ribosomal protein	9 + 13
Receptor	1 + 2
DNA binding	1 + 1
RNA binding	0 + 2
Cell cycle regulator	1 + 1
Nucleotide binding	2 + 0

Figure 3 Distribution of genes and pseudogenes on chromosomes 21 and 22 into GO functional categories. For each table within the figure, the total is split up into the number for chromosome 21 plus number for chromosome 22 and is sorted in decreasing order of this total. Each GO functional category is given a different color. The table at top *left* lists the top five GO functional categories for genes. The table at top *right* lists the top five GO categories for pseudogenes (processed and nonprocessed combined). This is split up into processed and nonprocessed; processed pseudogenes are further divided into ancient and modern sets, as described in the text. The number of immunoglobulin pseudogenic fragments is lumped in with the nonprocessed population. The GO functional categories in the figure are as follows (with GO numbers): (1) "Transcription factor," GO:0003700. For processed and nonprocessed pseudogenes these all arise from zinc finger C2H2 type. (2) "Other DNA-binding," all DNA-binding (GO: 0003677) except "Transcription factor." The proteins found for processed pseudogenes in this class are as follows: (IPR000910) HMG1/2 (high mobility group) box; (IPR000210) BTB/POZ domain; (IPR000637) HMG-I and HMG-Y DNA-binding domain (A + T-hook); (IPR002119) Histone H2A; (IPR000079) High mobility group proteins HMG14 and HMG17; (IPR001514) RNA polymerases D/30 to 40 Kd subunits; and (IPR001005) Myb DNA binding domain. (3) "Nucleotide-binding," GO:0000166. (4) "Nucleic-acid binding," GO:0003676 (this class arises from motifs or domains that cannot be classified specifically as "DNA-binding" or "RNA-binding"). (5) "Kinase," GO:0016301. (6) "Ribosomal protein," GO:0003735. (7) "Receptor," GO:0004930. (8) "Transferase," GO:0016740. (9) "RNA binding," GO:0003723. The proteins found for processed pseudogenes in this class are as follows: (IPR001014) Ribosomal L23 protein (also classed as ribosomal protein); (IPR001410) DEAD/DEAH box helicase; (IPR002942) S4 domain; (IPR001892) Ribosomal protein S13 (twice, also classed as ribosomal protein). The protein found for nonprocessed pseudogenes in this class is: (IPR001965) PHD finger. (10) "Oxidoreductase," GO:0016491. (11) "Cell cycle regulator," GO:0003750.

tional classes (Methods section). Ancient and modern processed pseudogenes do not have any overwhelmingly obvious prevalences; the "ribosomal structural protein" functional class dominates both sets. Transcription factors and other DNA-binding proteins tend to be in the ancient category (Fig. 3).

Main Nonprocessed Pseudogene Populations

We examined the nonprocessed pseudogene populations for chromosomes 21 and 22 for their prevalent functional classes and compared them with the classes for genes predicted using `GenomeScan` (Yeh et al. 2001). The total number of InterPro motif assignments and consequently GO-class assignments is much smaller for the nonprocessed pseudogenes in comparison with the gene totals and those for processed pseudogenes (Fig. 3). There is little similarity among all three lists (genes, processed pseudogenes, and nonprocessed pseudogenes for chromosomes 21 and 22 combined); the "transcription factor" functional classes occurs in the top five of all three (processed pseudogenes, nonprocessed pseudogenes, and genes, Fig. 3). The "receptor" class is common to the top five of both processed and nonprocessed pseudogenes. The nonprocessed pseudogenes share three of their top five functional categories with the top five for the `GenomeScan`-predicted genes. On a related note, in general, a large proportion of the nonprocessed pseudogenes are close to a homolog along the chromosomes: 28% have at least one homolog within 0.5 Mb and 31% have at least one within 1.0 Mb.

Immunoglobulin Gene Segments

There are a total of 70 κ and λ immunoglobulin (Ig) variable-region pseudogenic gene segments (65 λ, 5 κ) in the chromosome 22 loci for these gene segments. We find only an additional two (λ) pseudogenic gene segments relative to those already annotated by the Sanger Center (included in this total). Ig variable-region gene segments have a higher rate of nonsynonymous substitution in the germ line relative to synonymous substitution (Nei et al. 1997). We examined the variable-region Ig pseudogenic gene segments for the total number of disablements detected relative to the closest matching human protein sequences from the Ensembl database (Birney et al. 2001). We find a moderately increased rate of disablements per amino acid relative to the corresponding overall rate in pseudogenic sequences: 3.3% in Ig segments (106/3253) relative to 2.5% overall (1704/67965). This difference is statistically significant (the chance that it would arise randomly is P <0.002, assuming normal distribution statistics), and is unaffected by removing the five κ segments. This increased rate of disablement is consistent with the increased nonsynonymous substitution rate for Ig variable-region loci referred to in the literature (Nei et al. 1997). The nonprocessed pseudogene population on a whole has a slightly higher rate of disablement than the processed one, 2.6% (837/32843) versus 2.4% (759/31868). The degree of identity between the Ig pseudogenic gene segments and their closest matching Ensembl human protein sequences is also much lower on average (59.2% [+13.9]) than for either processed or nonprocessed pseudogenes (72.4% (+20.4) and 75.1% (+19.1), respectively); these latter two categories also have similar-shaped distributions (Fig. 5).

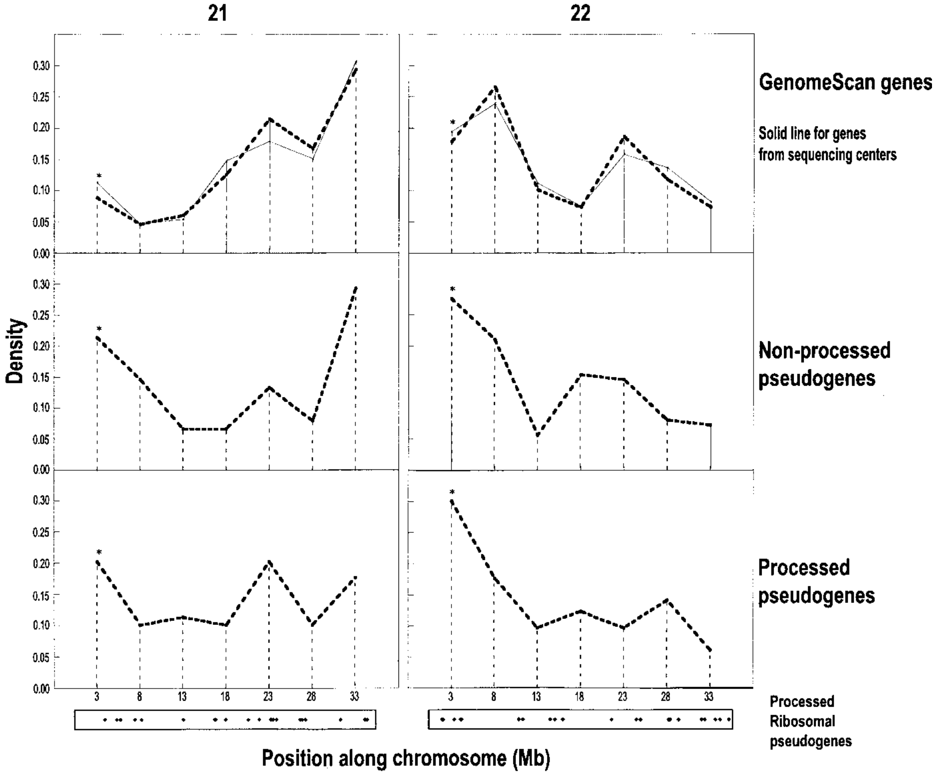


Figure 4 Distribution of pseudogene and gene densities for chromosomes 21 and 22. On the *left* are panels for chromosome 21, and on the *right* are panels for chromosome 22. For each chromosome, the panels are genes predicted by GenomeScan and the genome sequencing centers (Riken for chromosome 21 and Sanger for 22) (*top*), nonprocessed pseudogenes (*middle*), and processed pseudogenes (*bottom*). Each bin is named *x* for the interval *x* to *x* + 5 Mb. The first bin contains ~300,000 bases that are beyond the centromere (containing two genes and six pseudogenes). The final bin ends at the end of the telomere. The bins for the pseudogenic hot-spots referred to in the text are asterisked. For processed pseudogenes, we have added a representation of the distribution of ribosomal–protein processed pseudogenes along the chromosomes at the bottom of the panels, with a dot for each ribosomal–protein pseudogene at its approximate position along the chromosome.

Pseudogenic Hot-Spots

The density of genes or pseudogenes is defined as their number per interval of DNA. We have illustrated the trends for the largest interval for which we obtain any meaningful separation along the chromosomes (Fig. 4). We searched for the most notable differences in the pseudogene density and the gene density (either processed and nonprocessed), where they are observed for both the GenomeScan genes and the Riken/Sanger complete sets of gene annotations. We find that the most notable increased density for both processed and nonprocessed pseudogenes relative to the gene density is near the centromeres (in the first 5 Mb; difference in density, $\Delta D > 0.10$; Fig. 4). The most notable excess in gene density relative to the pseudogene density is at the telomere of chromosome 21, where there are few processed pseudogenes ($\Delta D = -0.13$); this area contains predicted collagen genes and nonprocessed pseudogenes. It will be interesting to see if regions of increased pseudogene density in the absence of increased gene density or pseudogenic hot-spots can be found on a larger scale in the total human genome. In general, pseudogenes in such regions may be more detectable because they take longer to be degraded; this may occur, perhaps, through local variations in DNA duplication rate relative to the rate of loss of genomic DNA (Petrov 2001).

The G + C content of genomic DNA is related to gene content, with G+C-rich regions having elevated numbers of genes relative to G+C-poor regions (Dunham et al. 1999; Lander et al. 2001; Venter et al. 2001). There does not appear to be any obvious relationship between pseudogene content and G + C content that can be readily decoupled from the known link between gene density and G + C content. On chromosome 21, the most G + C-poor region, which is between 5 and 12 Mb from the start of the sequence, has low G + C (35%) compared with the rest of the chromosome (43%), and has low pseudogene content as well as low gene content (Hattori, et al. 2000); on chromosome 22, the most notable G + C-poor region, the 2 Mb closest to the centromere (<40% G + C; Dunham, et al. 1999), has elevated pseudogene content relative to gene content (Fig. 5). Evidently, this topic will be amenable to in-depth study with a larger data set of pseudogenes derived from the whole human genome.

CONCLUSIONS

We have derived a procedure for the assessment of processed and nonprocessed pseudogenes in genomic DNA by looking for disabled protein homologies while minimizing the overlap with known genes; using this, we have predicted the pseudogene populations of chromosomes 21 and 22, finding 180 pseudogenes additional to existing available annotations. Also, we have tentatively extrapolated that there are up to

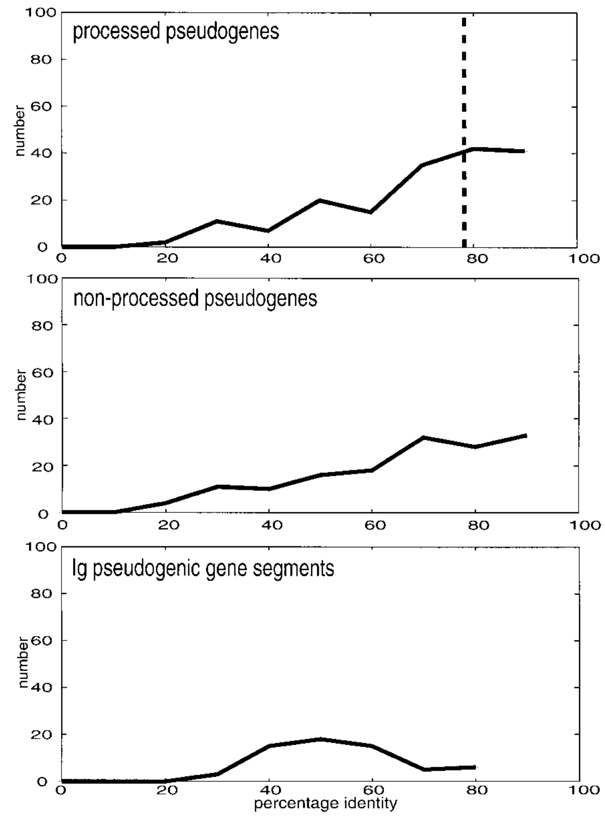


Figure 5 Distribution of the percent identity to closest-matching Ensembl proteins for processed, nonprocessed, and immunoglobulin gene segment pseudogenes. The percentage identity to the closest matching Ensembl human protein for pseudogenes. The bin named x contains every value y such that $x < y < (x + 10)$%. The panels are processed pseudogenes (*top*), nonprocessed pseudogenes (*middle*), and immunoglobulin pseudogenic gene segments (*bottom*).

~9000 processed and ~10,000 nonprocessed pseudogenes in the human genome. Up to 6% of annotated exons in these two chromosomes may be pseudogenic. Based on GenomeScan gene predictions, modified totals for the actual number of genes on chromosomes 21 and 22 are given in Table 1.

Other types of protein-related pseudogenes that are not accounted for in the present work are semiprocessed (pseudo)-genes (arising from aberrant mRNAs that contain an intron) and pseudogenes that produce transcripts (but not protein chains). Surveys of the literature by the authors indicate, however, that the occurrence of either of these is relatively rare and is not likely to affect gene prediction significantly. From examination of the distribution of pseudogenes along chromosomes 21 and 22, there is some evidence of the existence of pseudogenic hot-spots; this will remain to be confirmed upon examination of the whole human genome. This study serves as preparation for such a whole-genome survey.

The density of pseudogenes relative to genes derived in this study seems very high (one processed and one nonprocessed pseudogene for every ~ four genes), with a total of ~390 found in the 70 Mb of chromosomes 21 and 22, with >97% noncoding DNA. By comparison, a moderately-sized complement of > ~1100 verified pseudogenes (corresponding to ~19,000 genes) was found in the 100-Mb worm genome (Harrison et al. 2001), which has ~70% noncoding DNA. Estimates for the other eukaryote genomes are at present unavailable, although for the fly genome (which has 120 Mb of euchromatic DNA with ~80% noncoding DNA), a survey by the authors indicates ~100 pseudogenes (P. Harrison and M. Gerstein, submitted). There appears to be no obvious relationship between the proteome size or genome size or the amount of noncoding DNA and the number of pseudogenes for worm, fly, and human. Contributing factors would include the rate of gene duplication, the occurrence of transposable elements, and the overall rate of genomic DNA loss (Petrov 2001). Also, for prokaryotes, there does not appear to be a clear relationship between the amount of noncoding DNA and the number of pseudogenes. There are two reported cases of prokaryotic genomes with high proportions of noncoding DNA: *Rickettsia prowazekii* has 24% noncoding DNA and 12 pseudogenes (Andersson et al. 1998); whereas *Mycobacterium leprae* has 51% noncoding DNA, 27% of which is composed of a population of 1100 pseudogenes (corresponding to a proteome of 1604 coding sequences) (Cole et al. 2001). Evidently, surveys of more eukaryote and prokaryote genomes are required to give a fuller picture.

METHODS

Determination of a Set of Pseudogenes for Human Chromosomes 21 and 22

We developed an initial scheme for identifying pseudogenes in human genomic DNA; this is depicted as a flow diagram in Figure 1. Genomic annotation is an inherently dynamic process in which it is necessary to make use of many different sources of data (represented by ovals in the flow diagram), which are not updated in a concerted fashion. (Detailed files listing our assignments are available at http://bioinfo.mbb.yale.edu/genome/pseudogene and http://genecensus.org/pseudogene.)

Six-Frame BLAST

Using the BLAST alignment package (Altschul et al. 1997) with repeats masked using RepeatMasker (Bedell et al. 2000), we compared the SWISSPROT database of protein sequences (version 40) (Bairoch and Apweiler 2000) with the complete available sequences of human chromosomes 21 and 22 in six frames (Dunham et al. 1999b; Hattori et al. 2000). Chromosome 21 was downloaded from GenBank on August 22, 2000; chromosome 22 is the May 9, 2000, version). We then took all of the significant protein matches to the genomic DNA (e-value < 1×10^{-4}), and reduced them for mutual overlap by picking matches in decreasing order of significance and deleting any matches that overlap substantially with a picked match (i.e., more than 10 amino acids).

FASTA *Realignment*

For each match, we then realigned the matching SWISSPROT sequence to the same region of genomic DNA using the FASTA program (Pearson et al. 1997), expanded on either side by the length of the matching sequence (in nucleotides). From these alignments, we picked any matching sequences that had more than one disablement (either a frameshift or a premature stop codon) as a potential pseudogene. At this stage, these potential sequences were filtered for low complexity by comparing with the same SWISSPROT database but using masking with SEG (settings "25 3.0 3.3" and "45 3.4 3.75") (Wootton and Federhen 1996).

Minimize Overlap with Known Genes

To ensure that we are not considering disablements at the end of alignment subsequences that are artifactual, we examined how these potential pseudogenic sequences overlapped with the known genes on human chromosome 22 (Dunham et al. 1999a). We calculated the position of the disablement that is nearest the middle of each potential pseudogenic sequence and the distance (*d*) from this particular disablement to the closest end of the sequence. We then determined an appropriate margin (*m*) for the ends of the pseudogenic sequences, which allows us to discard pseudogenes so that the total set of pseudogenes overlaps only a small proportion (<5%) of the set of known gene exons. Following this criterion, all pseudogenes were discarded if $d < m$, where $m = 16$ residues. This gave us a total of 3.3% of all known gene exons that are overlapped by our set of pseudogene predictions.

Merging

At this stage, the pseudogene predictions for chromosome 22 were merged with previous pseudogene predictions provided by the Sanger Center (Dunham et al. 1999b); those for chromosome 21 were merged with data downloaded from the Riken sequencing Center Web site (Hattori et al. 2000). Where a predicted pseudogene from the present data set was duplicated by a Riken or Sanger Center pseudogene, the Riken/Sanger Center annotation is chosen in preference. The nature of the overlap of the data sets is shown in box 4 of the flow diagram (Fig. 1).

Date by Closest Match Ensembl Protein

For each pseudogene sequence, we searched through the most current version of the Ensembl database (http://www.ensembl.org; [Birney et al. 2001]) of human coding sequences to find the closest known human sequence homolog. If a matching sequence from Ensembl was found (only 84% of Sanger Center annotations have a corresponding Ensembl database protein; 92% of our own pseudogenes have a match), each matching Ensembl sequence was then realigned, using the FASTA program (Pearson et al. 1997), to the region of genomic DNA corresponding to the pseudogene sequence but expanded on either side by the length of the Ensembl sequence (in nucleotides). Any pseudogene that was lengthened as a result of this realignment was replaced with the new Ensembl-derived sequence, and the new updated list of pseudogenes was reduced for overlap, as above. We then rechecked through the complete set of pseudogenes for overlap with known genes as described previously.

Assess for Processing

We inspected the genomic DNA around the potential pseudogenes for any evidence of exon structure from existing Riken or Sanger Center gene and pseudogene parsing, from gene annotations made using the program `GenomeScan` (Yeh et al. 2001), or from evidence of exon structure from our own pseudogene analysis. For this third option, we considered an intron to occur if the gap in the sequence was >126 nt (from inspection of the distribution of intron lengths for known genes of chromosome 22, only 5% of introns would be shorter than this). From our visual curation, we made a decision as to whether the predicted pseudogene fragment was part of a larger exon structure. We also checked for pseudogenic duplications of any single-exon genes.

For any potential pseudogene that does not have such evidence of exon structure, we used two lines of evidence for assessing whether it was processed:

(1) We labeled as candidate-processed pseudogenes all those matches that comprise >70% of the length of the closest-matching human Ensembl or SWISSPROT database sequence in a continuous segment. This criterion was used previously by Venter et al. (Venter et al. 2001). We checked over the list for matches to known single-exon human genes (which, for example, comprise ~6% of the known genes on chromosome 22), that do not have any evidence for processing from the analysis for a polyadenine tail (see below).
We checked the utility of this criterion for a set of 46 previously identified primate processed pseudogenes that we collated from Genbank (Benson et al. 2000) in August 2000. Almost all of these processed pseudogenes (42/46, 91%) were detected by this criterion.

(2) Recently processed pseudogenes can be identified through the existence of a polyadenine tail of at least 15–20 nt that is preceded by a polyadenylation signal (AATAAA), usually about 15–20 nt upstream. We searched a 1000-nt region that was 3′ to the pseudogene homology segment, with a sliding window of 50 nt for a region of elevated polyadenine content (>30 nt), and picked the most adenine-rich 50-nt segment as the most likely candidate. An interval of 1000 nt was used because of the possible existence of 3′-untranslated regions (3′-UTRs); 90% of 3′-UTRs are of length <942 nt (Makalowski et al. 1996). In addition, we searched in the same 1000-nt region for candidate AATAAA polyadenylation signals and checked whether they were upstream to the candidate polyadenine tail site. When a polyadenylation signal was found within 50 nt upstream of the candidate polyadenine tail, the pseudogene is labeled as a “class 1” candidate processed pseudogene and “class 2” if the signal is found between 51 and 100 nt upstream. The latter class may arise if there is a genomic DNA insertion event; in reality, there are very few of them found (eight in total). All other pseudogenes with a detected candidate polyadenine tail are labeled as “class 3.” This last class (exclusively) only accounts for 17% of the candidate-processed pseudogenes; their removal does not change the main results and trends reported in the paper.

There is considerable overlap between criteria (1) and (2) for detecting processing; about half (52%) of the candidate-processed pseudogenes assigned by criterion (1) are also classified as having polyadenylation. A quarter (26%) of the candidate-processed pseudogenes only have evidence for polyadenylation (criterion [2]). Of the set of previously annotated primate-processed pseudogenes, downloaded from Genbank (see [1] above), 54% are detected to have evidence of polyadenylation.

All remaining pseudogene sequences were categorized as nonprocessed; some of these were merged into single nonprocessed pseudogenes from visual examination of the sequence fragments involved.

Analysis for Protein Function

Each pseudogene sequence was run through the InterPro sequence motif assignment package `InterProScan` (Apweiler et al. 2000). Functional categories were then assigned using the GO classification (Ashburner et al. 2000) with a list of correspondences between InterPro motifs and GO functional classes that is available from the InterPro Web site (http://www.ebi.ac.uk/interpro). A substantial proportion (~75%) of the pseudogene annotations were assigned to an InterPro motif and ~45% were able to be mapped onto GO function classifications. The GO categories given by InterPro were merged into a higher level if they were judged to be too specific, for example, all “receptors” are merged into one higher GO category. These proportions are at a level that is within the range of proportions of proteomes that have automatic reliable functional assignment in the GeneQuiz database (Hoersch et al. 2000), a well-known standard of automated functional classification. We did not try to map pseudogenes that do not have InterPro motifs to the GO classification because this introduces an extra degree of judgmental bias.

ACKNOWLEDGMENTS

We thank Ru-Fang Yeh and Chris Burge (MIT) for providing GenomeScan gene predictions and Sam Karlin (Stanford) for discussions. M.G. acknowledges support from the NIH Protein Structure Initiative (P50 grant GM62413-01).

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R.M., Naslund, A.K., Eriksson, A.S., Winkler, H.H., and Kurland, C.G. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396:** 133–140.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. 2000. InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16:** 1145–1150.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

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

Bedell, J.A., Korf, I., and Gish, W. 2000. MaskerAid: A performance enhancement to RepeatMasker. *Bioinformatics* **16:** 1040–1041.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. 2000. GenBank. *Nucleic Acids Res.* **28:** 15–18.

Birney, E., Bateman, A., Clamp, M.E., and Hubbard, T.J. 2001. Mining the draft human genome. *Nature* **409:** 827–828.

Cole, S.T., Eigimeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D., et al. 2001. Massive gene decay in the leprosy bacillus. *Nature* **409:** 1007–1011.

Crollius, H.R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet.* **25:** 235–238.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **232:** 232–233.

Gerstein, M. 1988. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* **33:** 518–534.

Gerstein, M. 1997. A structural census of genomes: comparing bacterial eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274:** 562–576.

Glusman, G., Yanai, I., Rubin, I., and Lancet, D. 2001. The complete human olfactory subgenome. *Genome Res.* **11:** 685–702.

Harrison, P., Echols, N., and Gerstein, M. 2001 Digging for Dead Genes: An Analysis of the Characteristics of the Pseudogene Population in the *C. elegans* Genome. *Nuc. Acids. Res.* **29:** 818–830.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., et al. 2000. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405:** 311–319.

Hegyi, H. and Gerstein, M. 1999. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288:** 147–164.

Hoersch, S., Leroy, C., Brown, N.P., Andrade, M.A., and Sander, C. 2000. The GeneQuiz Web server: Protein functional analysis through the Web. *Trends Biochem. Sci.* **25:** 33–35.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature* **409:** 860–921.

Makalowski, W., Zhang, J., and Boguski, M. S. 1996. Comparative analysis of 1,196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* **6:** 846–857.

Mighell, A.J., Smith, N.R., Robinson, P.A., and Markham, A.F. 2000. Vertebrate pseudogenes. *FEBS Lett.* **468:** 109–114.

Nei, M., Gu, X., and Sitnikova, T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci.* **94:** 7799–7806.

Pearson, W.R., Wood, T., Zhang, Z., and Miller, W. 1997. Comparison of DNA sequences with protein sequences. *Genomics* **46:** 24–36.

Petrov, D.A. 2001. Evolution of genome size: New approaches to an old problem. *Trends Genet.* **17:** 23–28.

Qian, J., Luscombe, N.M., and Gerstein, M.B. 2001. Protein family and fold occurrence in genomes: Power-law behavior and evolution model. *J. Mol. Biol.* **313:** 673–681.

Uechi, T., Tanaka, T., and Kenmochi, N. 2001. A complete map of the human ribosomal protein genes: Assignment of 80 genes to the cytogenetic map and implications for human disorders. *Genomics*, **72:** 223–230.

Vanin, E.F. 1985. Processed pseudogenes: Characteristics and evolution. *Annu. Rev. Genet.* **19:** 253–272.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The Sequence of the Human Genome. *Science* **291:** 1304–1351.

Wootton, J.C. and Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266:** 554–571.

Wright, F.A., Lemon, W.J., Zhao, W.D., Sears, R., Zhuo, D., Wang, J.-P., Yang, H.-Y., Baer, T., Stredney, D., Spitzner, J., et al. 2001. A draft annotation and overview of the human genome. *Genome Biol.* **2:** research0025.1–0025.18.

Yeh, R.-F., Lim, L. P., and Burge, C. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11:** 803–816.

Received July 23, 2001; accepted in revised form November 28, 2001.