From: ***Bioinformatics and Genomes: Current Perspectives***
Edited ***by:*** Miguel A. Andrade

Chapter 11

Pseudogenes and Genomes

David Torrents, Mikita Suyama
and Peer Bork

Abstract

The knowledge on the gene content of any organism is essential for the study and understanding of its biology. The recent sequencing of large and complex genomes has forced the scientific community to develop or improve computer programs in order to identify such genes. These algorithms are based on the identification of characteristic patterns of gene-related elements (such as promoters, splice sites, polyadenilation signals, and others) and present an estimated success rate of 80%. But, neither these programs nor their evaluation procedures normally take into consideration the presence of non-functional gene copies in the genome. These dispensable gene copies, known as pseudogenes, are formed either by retrotransposition or by tandem duplication. In some cases they are difficult to differentiate by using standard procedures since they share many sequence characteristics with their corresponding functional parental genes. The only criteria used so far to identify such non-functional elements depends on the detection of either disruptions in the open reading frame or any typical sign of retrotransposition. This leads to misclassification of some genes. In order to overcome this

situation, we have developed an independent strategy that is capable to differentiate many functional from non-functional sequences. This procedure takes advantage of the different selective constrains associated to pseudogenes and genes. Using this method we estimated that the human genome contains 40000 pseudogenes, doubling current approximations. We are also proposing an error rate of 23% in standard procedures of gene annotation regarding the classification of genes and pseudogenes.

INTRODUCTION

The development of sophisticated techniques that permit direct manipulation of DNA has changed the way researchers approach scientific questions related to biological processes. Around fifty years ago, classical biochemistry was restricted to the investigation of biological processes (enzymatic reactions) from a chemical point of view. Some years later, protein sequencing and purification techniques permitted the finding of relationships between these processes and particular peptide molecules. Nowadays, the capability of DNA manipulation (purification, sequencing, modification, expression in living cells, etc…) permits scientists to analyze many biological processes from a molecular point of view. The empirical generation of a large amount of information regarding DNA -> Protein -> Function relationships and the formulation of general biological rules offers the possibility of prediction. On the basis of this knowledge and its application to newly identified DNA sequences, bioinformaticians are able to make predictions about biological processes and relationships between macromolecules. In this sense, the recent arrival of complete genomic sequences is "happily" received by the bioinformatic community as very promising material.

GENE PREDICTION

In order to exploit a genome, the correct identification of the genes contained therein is required. In the case of prokaryotic genomes, the task is relatively simple. As protein-coding regions in bacterial genomes are not interrupted by intronic sequences, their identification is reduced to the detection of open reading frames by simple translation of the DNA. The key of this process is the definition of a gene in terms of the minimum accepted size, compromising the identification of the small ones. Nevertheless, this problem has been partially solved by evaluating the existing differences in nucleotide composition between coding and non-coding regions (Salzberg *et* al., 1998).

Gene prediction from eukaryotic genomes requires much more attention and effort owing to the large size of the sequences, the low gene density (estimated to be around 1 gene/100 kb on average in the human genome; International Human Genome Sequencing Consortium, 2001) and the complexity of the gene structures (mainly the presence of introns that in higher eukaryotes can be larger than 100 kb). Alternative splicing (the possibility of expressing differents sets of exons for a given gene under different conditions), and the presence of genes nested in the intronic sequences of other genes, complicate even more the predictions. The recent release of the draft sequence of the Human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.,* 2001) and the partially sequenced mouse genome constitute a challenge for the algorithms developed to find genes in complex scenarios.

Some collections of predicted genes have been lately proposed for the human genomes (Venter *et al.,* 2001 ;Yeh *et al.,* 2001; Hubbard *et al.,* 2002). The methodology used in each case is quite different yielding distinct pictures of the human gene content as revealed by the little overlap observed among these collections (Hogenesch *et al.,* 2001). Usually the automatic generation of gene collections for a particular organism on the basis of its genome analysis should find a compromise between quantity and quality. If the set is too small, despite its high accuracy, it will be considered non-informative and the scientific community will not use it. On the other hand, a too large gene index is likely to contain many false positives (not true genes) and users will be skeptical about this information.

The rapid growth of the number of known cDNA sequences permits a broader identification of new genes on the basis of sequence similarity analysis. In this sense, programs that combine the identification of intron-exon boundaries with sequence similarity to known cDNAs or derived proteins, such as BLAST (Altschul *et al.,* 1997) or GENEWISE (Birney and Durbin, 1997), are highly efficient in identifying genes but too expensive in terms of time and computer power when applied to large genomic sequences. Therefore, most gene prediction strategies tend to save time by sacrificing, at least in the initial steps, the sequence similarity analysis. The programs for *ab initio* identification of eukaryotic genes (i.e. without using sequence similarity) have improved along with the available empirical knowledge on sequence patterns associated to genomic elements (for review see Guigó, 1997; Burge and Karlin, 1998; Guigó *et. al.,* 2000). In this sense, these algorithms are designed to identify protein-coding genes basically upon detection of the first and last coding exons, intron-exon boundaries, promoter sequences (mainly transcription start sites and TATA-box signals), translational signals (Kozak, 1996), and by analysing their sequence composition (using Hidden Markov Models). It has been reported that these

programs, concretely GENESCAN (Burge and Karlin, 1997), present a sensitivity (proportion of true genes found) and specificity (proportion of predicted genes that are real) around 80% on average (Guigó *et* al., 2000). It should be though mentioned that this estimation was obtained considering artificial and controlled data sets that are quite distinct from real genomic sequences, i.e. they did not contain non-functional gene copies. Since these non-functional gene copies, known as pseudogenes, normally present many of the sequence characteristics found in genes, their presence should be taken into account when evaluating the efficiency of gene prediction methodologies.

Pseudogenes Complicate Gene Prediction

We define as pseudogenes all dispensable gene copies unable to code for functional proteins (for review see Vanin, 1985; Mighell *et* al., 2000). Those gene duplications that occur in germinal cell lines and are harmless for the organism will remain in the population and hence will be present in available genomic sequences. On the basis of the mechanism of such duplication two types of pseudogenes can be distinguished in eukaryotic genomes: processed and non-processed pseudogenes. Processed pseudogenes (also called retro-pseudogenes) are the result of a retro-transposition event in which single-stranded mRNA undergoes retro-transcription and integration in the genome with the help of the enzymatic machinery of retrotransposable elements (Esnault *et* al., 2000). Most of these pseudogenes have lost all or part of the introns present in the parental gene and are likely to be inactive right from the time of generation since they typically present no 5'-promoter sequence. Non-processed pseudogenes are formed by partial or complete gene tandem duplication. On the one hand, fragmented gene duplicates (i.e. lacking promoter sequences or relevant exons), are likely to be born as pseudogenes, like most processed pseudogenes. Complete gene duplicates have the chance to remain active by gaining a new function (neofunctionalization), but in most of the cases they are converted into pseudogenes by the acquisition of mutations that disrupt their expression. The characteristic feature of all non-processed pseudogenes is their partial or total conservation of the gene structure of the parental gene.

How do current gene prediction strategies differentiate genes from their non-functional copies? Since pseudogenes often accumulate all possible sequence alterations with time, they may include disruptions (STOP codons or frameshifts) in the corresponding open reading frame (ORF). Standard annotation strategies catalogue as pseudogenes all identified genomic regions whenever such disruptions are detected. In addition, due to the low probability

of a retrotransposed gene being correctly inserted in front an active promoter, any sign characteristic of a retrotranspositional origin (mainly the loss of introns) is also considered as indicative of non-functionality. Although the detection of these features can be used to correctly classify many real pseudogenes, we can propose some situations where these could lead to erroneous conclusions. For instance all functional processed genes (retrogenes; Brosius, 1999), all genes with pseudo-exons that are skipped during splicing, and those genes presenting sequencing errors leading to disruptions in their ORFs would be misclassified as pseudogenes. Moreover, non-processed pseudogenes with sequence alterations other than clear disruptions, e.g. amino acid substitutions or critical alterations at the promoter level, are likely to end up being annotated as genes. Although there is no quantification of such situations, as it is difficult to prove non-functionality given the infinite conditions under which a gene could be expressed, as much as 21% of current gene predictions may be affected by these cases (International Human Genome Sequencing Consortium, 2001). Therefore additional ways of discrimination between functional and non-functional sequences should be considered. On the basis of these simple criteria, pseudogenes have been identified in many organisms, from prokaryotes (Andersson and Andersson, 2001 ;Cole *et* al., 2001) to higher organisms (Goncalves *et al.*, 2000; Harrison *et al.*, 2001). At present, the total number of pseudogenes is not known for any organism.

In the case of the human genome, the complete sequence and annotation of human chromosome 22 and 21 indicate a proportion of one pseudogene every 4 or 5 functional genes (Dunham *et al.,* 1999; Hattori *et al.,* 2000) suggesting a total of 7000 to 9000 pseudogenes in the complete human genome by assuming a total of 35000 human genes. On the basis of a different study where the content of processed pseudogenes was analyzed in a restricted portion of the human genome, a ratio of one processed pseudogene every 3 functional genes has been also proposed (Goncalves *et al.,* 2000). Public databases, which contain gene predictions of the human genome, do not normally consider annotation of pseudogenic sequences as relevant. By simply searching with the keywords "human" and "pseudogene" we can retrieve up to 3534 and 767 sequences from the public DNA databases GeneBank and EMBL, respectively (by March, 2002). Since most of these entries correspond to side-products obtained from diverse genetic studies, their annotation quality is expected to be low. Despite all these surveys, there has been so far no extensive and accurate approach aiming at the identification and annotation of pseudogenes in the whole human genome. But, the number of pseudogenes, at least in mammal genomes, is likely to be high enough to demand careful consideration in gene identification procedures.

An exhaustive and accurate identification of pseudogenic sequences and an estimation of their content for available genomes are therefore needed. This information is not only relevant for an accurate identification of functional genes, but also offers the possibility to understand and quantify active genomic processes, such as gene duplication and non-functional DNA removal (''housecleaning''), which directly influence both the general evolution of the organism and the size of the genome.

PSEUDOGENE PREDICTION

In order to improve discrimination between functional and pseudogenic genomic regions we have considered in a recent study (manuscript under submission) a parameter corresponding to the ratio of innocuous to deleterious nucleotide substitutions associated to a particular problem sequence. According to the neutral evolution theory, pseudogenes, like other non-functional genomic regions, are unconstrained by selection (Kimura, 1977). This means that any kind of mutation affecting a pseudogenic sequence will be harmless for the organism, possibly fixed in the population and hence detectable in available sequences. On the contrary, most deleterious mutations, i.e. negatively affecting the function of a gene, will be selected against and hardly maintained in the population. In general, the evaluation of the ratio of neutral to deleterious substitutions arising in a particular sequence involves two basic steps: (i) estimation of the number of neutral and deleterious sites at the moment of sequence formation (right after duplication), and (ii) counting the number of neutral and deleterious sites substituted thereafter. This last step must include a correction for multiple substitutions occurring at same sites.

From a practical point of view, neutral mutations are defined as substitutions that do not change the amino acid composition of the gene product (synonymous), while deleterious mutations account for those that induce amino acid replacement (non-synonymous). It is likely that some point mutations occurring in functional genes, despite inducing no amino acid replacement and therefore considered synonymous or neutral, are indeed deleterious and consequently selected against. On the other hand, substitutions inducing replacement of irrelevant amino acids and therefore considered non-synonymous are in fact fixed in the population as neutral. We believe that, although it is not possible nowadays to identify and quantify with precision neutral and deleterious sites, the approximations obtained are fair indicators of the degree of selective pressure associated to a particular sequence. These ratios, designed as d_S/d_N (or K_S/K_A), where d_S = number of

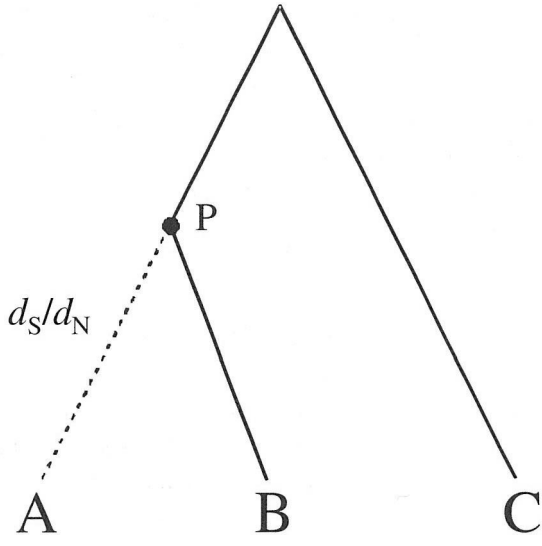

Figure 1. Assumed phylogenetic relationship between the problem sequence and the two closest functional homologues. The d_S/d_N associated to the problem sequence (A) is calculated along the dashed branch, i.e. from the parental sequence (P). P is inferred from the alignment of sequences A, B, and C using the parsimonian method (Yang, 1997).

synonymous substitutions / total number of synonymous sites, and d_N = number of non-synonymous substitutions / total number of non-synonymous sites, are expected to be about one for pseudogenes, and higher in the case of functional genes (Li *et al.*, 1981). Several analyses based on the calculation of d_S/d_N have been successfully applied to case studies to measure functional constraints associated to sequence evolution (Ohta and Ina, 1995; Nekrutenko *et al.*, 2002), but never before as a criterion to discriminate between genes and pseudogenes in genome annotation.

In this sense, we have developed a strategy to automatically obtain reliable d_S/d_N values for large data sets. Basically, this approach consists on predicting all point mutations that have been fixed in our problem DNA sequence (A in Figure 1) from the moment of its formation, i.e. from the duplication of the parental sequence (P). In order to do so, we need to deduce the nucleotide sequence of this parental gene and then to compare it with our problem sequence A. The inference of P requires two homologous sequences (B and C). The assumed phylogenetic relationship between all these sequences is shown in Figure 1. We believe that, in our study, this phylogeny accounts for the majority of the cases.

The application of similar protocols in studies of sequence evolution tends to force the comparison of the complete sequence A with very close

homologues B and C. In contrast, we opted to restrict this analysis to those regions of the sequences that appear to be more conserved, i.e. regions that are expected to be under stronger selective pressure in genes. Since we can, consequently, expect higher differences of d_S/d_N values between functional and pseudogenic sequences we have permitted the use of more remote homologues B and C. In this way we increased the number of problem sequences for which it was possible to compute d_S/d_N values.

We evaluated our strategy and the reliability of the resulting d_S/d_N ratios as criterion to discriminate functional from non-functional sequences using two confident sets of functional and pseudogenic human sequences, respectively. A non-redundant (up to 50% amino acid identity) data set consisting of 3034 well annotated human cDNA sequences (the human reviewed annotation fraction of the RefSeq database, Pruitt and Maglott, 2001) was taken as the *functional set*. The collection of pseudogenes was obtained through a homology search within intergenic regions and consisted of 1730 processed elements containing at least one stop codon or frameshift in the first half of the corresponding ORFs and thus likely to be non-functional. We applied to these two sets three different methods to acquire d_S/d_N values that use different calculation models (Nei and Gojobori, 1986; Ina, 1995; Yang and Nielsen, 2000) obtaining similar results. The logarithmic distributions of these two sets according to their associated d_S/d_N values are clearly distinct as shown in Figure 2 (using the method described by Yang and Nielsen, 2000). Most d_S/d_N values associated to either functional or pseudogenic sequences are clearly indicative of stronger and weaker selective constraints, respectively. It should be noticed that positive selection, theoretically observed if $d_S/d_N < 1$, is suspected in a very few cases (Endo *et. al.*, 1996) and therefore not considered as such in this study. But, why some d_S/d_N values do not strictly follow the theoretical expectation: d_S/d_N for pseudogenes = 1, and d_S/d_N for genes > I? Despite a certain level of inaccuracy of our method, we can think of some explanations accounting for these situations. (i) Fast evolving genes are expected to present d_S/d_N values close to one, as the number of amino acid replacement substitutions (d_N), under weaker selective constraints, can get close to the number of synonymous substitutions (d_S); and (ii) the restriction of our analysis to sequence regions with high amino acid conservation, forces d_N to remain low and therefore pushes d_S/d_N of some pseudogenes to higher values.

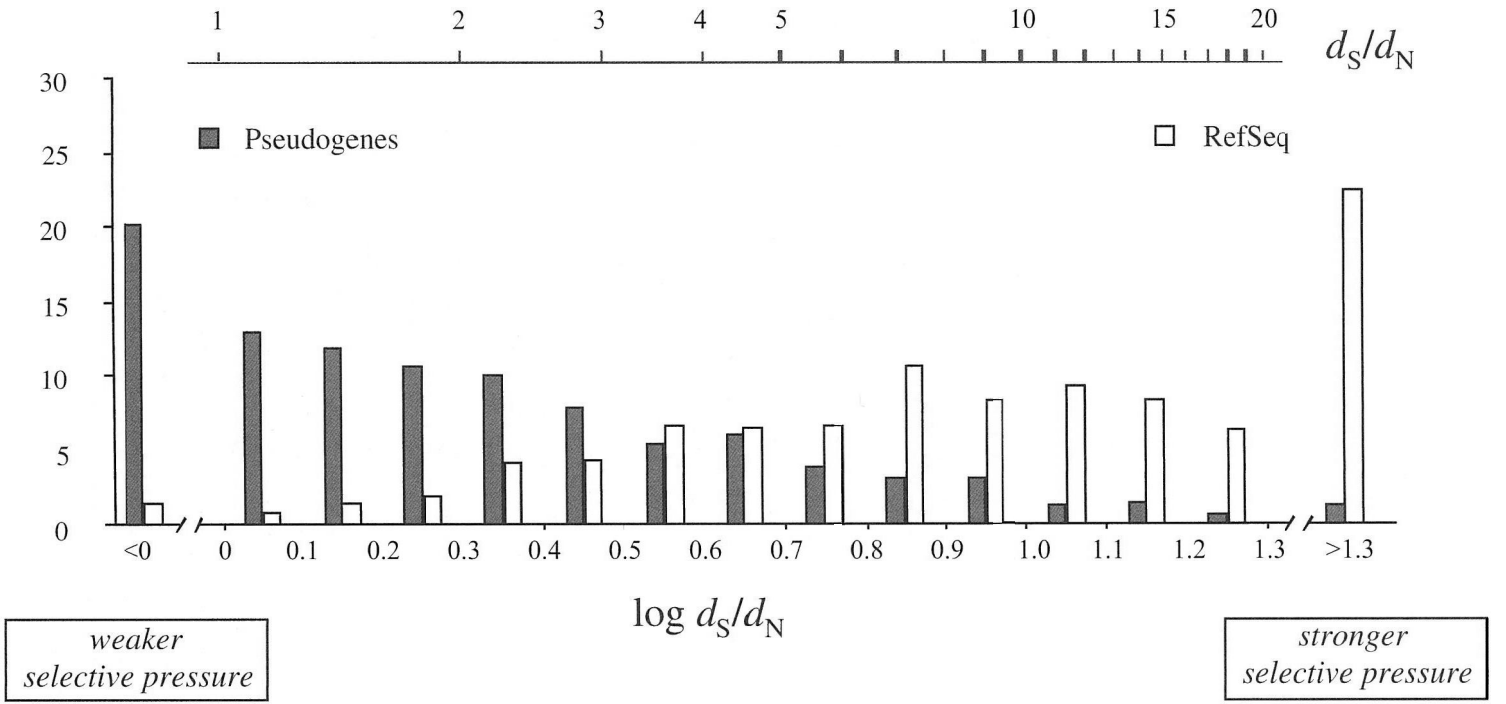

Figure 2. Distribution of reliable functional and pseudogenic sequences according to their associated d_S/d_N values.

How Many Pseudogenes Are In The Human Genome?

The information about the behaviour of d_S/d_N values obtained form functional and pseudogenic datasets can be used to estimate the portion of pseudogenes contained in any collection of human sequences with homology to known proteins. Accordingly, we carried out the same d_S/d_N analysis on a set of sequences obtained from a homology-based search through all intergenic regions in the human genome (according to ENSEMBL human gene database; Hubbard *et. al.*, 2002). Of all the sequences found, we estimated, based on the d_S/d_N values obtained, that around 10000 sequences corresponded to pseudogenes and around 2000 to functional genes. From a deeper analysis of two subsets containing either pseudogenes or genes regarding the presence of ORF disruptions with two reliable subsets of identified pseudogenes and functional sequences, our calculations indicated that up to 32% of the pseudogenes and 26% of the genes identified by standard annotation strategies (Hubbard *et. al.*, 2002) could be miscataloged as genes and pseudogenes, respectively.

We believe that this estimate of 10000 pseudogenes in the entire human genome is probably far too low. Taking into account the limitations internal to our homology search strategy (sequence similarity threshold applied and sequences lost by common DNA repeat masking), we can increase this estimate up to 35000 pseudogenes. Furthermore, we have reasons to suppose that some sequences annotated as genes in ENSEMBL database and therefore excluded from our search, could correspond to pseudogenes. Based again on the d_S/d_N analysis, we have estimated that this group of elements covers up to 23% of the whole database. We have found that most of these gene-catalogued pseudogenes correspond to non-functional partial tandem gene duplications, which have not yet acquired ORF disruptions and therefore difficult to index as such. If we add these elements, our estimate of human pseudogenes raises to 40000, covering at least a 5% of the genome.

Conclusion

We believe that 40000 is still an underestimate of the real number of pseudogenes in the human genome, since many pseudogenes are hidden to us due to their small size or degeneration under the possible level of detection by sequence similarity. The amount of non-functional DNA created by duplication seems to be higher than expected, as the high rate of pseudogene formation reflects, suggesting a relaxed evolutionary pressure on genome size in humans. This is in contrast to what has been proposed for the fly,

where the high rate of non-functional DNA removal suggests a higher selective constraint on the size of its genome. However we don't know whether our genome is still growing by means of DNA duplication, or whether we have reached the "allowed" genomic size, i.e. the formation and removal of non-functional DNA are at equilibrium. A deeper analysis regarding the age of the pseudogenic regions and the rate of DNA removal can bring light to this question. Considering gene duplication as one of the important driving forces of evolution, an accurate analysis of the pseudogene content in other organism and their comparison to the human pseudogene set is needed as it can offer hints regarding the speed of genome evolution.

REFERENCES

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

Andersson, J.O. and Andersson, S.G. 2001. Pseudogenes, junk DNA, and the dynamics of Rickettsia genomes. Mol. Biol. Evol. 18: 829-839.

Birney, E., and Durbin, R. 1997. Dinamite: a flexible code generating language for dynamic programming methods used in sequence comparison. Proc. Int. Conf. Intell. Syst. Mol. Biol. 5: 56-64.

Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene. 238: 115-134.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78-94.

Burge, C.B. and Karlin, S. 1998. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8: 346-354.

Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, PR., Honoré, N., Gamier, T., Churcher, C., Harris, D., et al. 2001. Massive gene decay in the leprosy bacillus. Nature 409: 1007-1011.

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

Endo, T., Ikeo K. and Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. Mol Biol Evol. 13: 685-690

Esnault, C., Maestre, J., and Heidmann, T. 2000. Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. 24: 363-367.

Gonçalves, I., Duret, L., and Mouchiroud, D. 2000. Nature and structure of human genes that generate retropseudogenes. Genome Res. 10: 672-678.

Guigó, R. 1997. Computational gene identification. J. Mol. Med. 75: 389-393.

Guigó, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. 2000. An assessment of gene prediction accuracy in large DNA sequences. Genome Res. 10: 1631-1642.

Harrison, P.M., Echols, N., and Gerstein, M.B. 2001. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic 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. Nature 405: 311-319.

Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R., Zhou, Y., Kay, S.A., Schultz, P.G., and Cooke, M. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell 106: 413-415.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., *et al.* 2002. The Ensembl genome database project. Nucleic Acids Res. 30: 38-41.

Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. 40: 190-226.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275-276.

Kozak, M. 1996. Interpreting cDNA sequences: some insights from studies on translation. Mamm. Genome 7: 563-574.

Li, W.-H., Gojobori, T., and Nei, M. 1981. Pseudogenes as a paradigm of neutral evolution. Nature 292: 237-239.

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

Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3: 418-426.

Nekrutenko, A., Makova, K.D., and Li, W.-H. 2002. The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res. 12: 198-202.

Ohta, T. and Ina, Y. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. J. Mol. Evol. 41: 717-720.

Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29: 137-140.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, **O.** 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26: 544-548.

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.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13: 555-556.

Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17: 32-43.

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