Digging for Dead Genes: An Analysis of the Characteristics of the Pseudogene Population in the *C. elegans* Genome

Paul M. Harrison, Nathaniel Echols and Mark B. Gerstein *

Dept. of Molecular Biophysics & Biochemistry,

Yale University,

260 Whitney Ave.,

P.O. Box 208114,

New Haven, CT 06511-8114,

U.S.A.

**corresponding author*

Telephone: (203) 432-6105, (203) 432-5065

Fax: (360) 838 7861

E_mail: Mark.Gerstein@yale.edu

Original version submitted to Nucleic Acids Research, August 17th, 2000

Revised version submitted November 20th, 2000

Abstract

Pseudogenes are non-functioning copies of genes in genomic DNA, which may either result from reverse transcription from a mRNA transcript (processed pseudogenes) or from gene duplication and subsequent disablement (non-processed pseudogenes). As pseudogenes are apparently 'dead', they usually have a variety of obvious disablements (*e.g.* insertions, deletions, frameshifts and truncations) relative to their functioning homologues. We have derived an initial estimate of the size, distribution and characteristics of the pseudogene population in the *Caenorhabditis elegans* genome, performing a survey in 'molecular archaeology'. Corresponding to the 18,576 annotated proteins in the worm (i.e., in Wormpep18), we have found an estimated total of 2,168 pseudogenes, about one for every eight genes. Few of these appear to be processed. Details of our pseudogene assignments are available from http://bioinfo.mbb.yale.edu/genome/worm/pseudogene. The population of pseudogenes differs significantly from that of genes in a number of respects: (i) Pseudogenes are distributed unevenly across the genome relative to genes, with a disproportionate number on chromosome IV; (ii) The density of pseudogenes is higher on the arms of the chromosomes; (iii) The amino-acid composition of pseudogenes is midway between that of genes and (translations of) random intergenic DNA, with enrichment of *Phe, Ile, Leu* and *Lys,* and depletion of *Asp, Ala, Glu* and *Gly* relative to the worm proteome; And (iv) the most common protein folds and families differ somewhat between genes and pseudogenes -- whereas the most common fold found in the worm proteome is the immunoglobulin fold, the most common 'pseudofold' is the C-type lectin. In addition, the size of a gene family bears little overall relationship to the size of its corresponding pseudogene complement, indicating a highly dynamic genome. There are in fact a number of families associated with large

populations of pseudogenes. For example, one family of seven-transmembrane receptors (represented by gene B0334.7) has one pseudogene for every four genes, and another uncharacterized family (represented by gene B0403.1) is approximately two-thirds pseudogenic. Furthermore, over a hundred apparent pseudogenic fragments do not have any obvious homologs in the worm.

Keywords: bioinformatics, genomics, nematode, molecular evolution, gene finding, protein folds, proteome

Introduction

Over the course of evolution, genes duplicate in the genome, gradually accumulating mutations that may lead to the acquisition of new functions, or to the modification of existing functions. However, some duplications of genes acquire deleterious mutations that disable them so that they can no longer be translated into a functioning protein. The disablement may occur at either or both the transcription and translation levels. These copies of genes are called non-processed pseudogenes. Pseudogenes may also arise by a process of retrotransposition, where a messenger RNA transcript is reverse transcribed and re-integrated into the genome [1-3]. These are termed processed pseudogenes or retropseudogenes and occur in a variety of plants and animals.

Some pseudogenes are evidently transcribed. A possible case of a 'functioning' pseudogene transcript has been described recently for neural nitric oxide synthase in the snail *Lymnaea stagnalis* [4]. Here, the pseudogene has a segment that is the inverse complement of the normal gene, and interferes through RNA duplex formation with the expression of nitric oxide synthase [4]. Interestingly, the expression of pseudogene transcripts can vary markedly with

respect to the expression of the transcripts of their homologous living genes. For example, for the 5-HT7 receptor, transcripts of a pseudogene can be detected in various tissues whereas transcripts for the corresponding functioning gene are absent [5]. Pseudogene transcripts can have raised expression in tumour cells, *e.g.* in laryngeal squamous cell carcinoma [6] or in glioblastoma [7].

Pseudogenes are important in the study of molecular evolution, since they generally acquire mutations, insertions and deletions without any apparent evolutionary pressures. (However, in *Drosophila* for example, many putative pseudogenes appear to have patterns of mutation that are inconsistent with a lack of functional constraints [8-10].) In evolutionary studies, pseudogenes have been used to derive underlying rates of nucleotide substitution [11-13] and rates of insertion and deletion in genomic DNA [14, 15]. In particular, Averof *et al.* [13] used eta-globin pseudogenes to show that double-nucleotide substitutions occur more often than would be expected from independent single-nucleotide substitutions. Gu and Li [14] noted that the pattern of insertions and deletions in processed pseudogenes implies that a logarithmic gap penalty dependence on gap size in sequence alignment is more appropriate than the more commonly used linear dependence. Ophir and Graur [15] performed a survey of processed pseudogenes in human and mouse and found evidence for distinctly different mechanisms underlying gene truncations, insertions and deletions each occur by different mechanisms. Pseudogenes are also useful in determining rates of genomic DNA loss for an organism: a smaller complement of pseudogenes in a genome implies a greater net loss of genomic DNA [10, 16]. Petrov *et al.* [16] demonstrated experimentally, using dead copies of retrotransposons as ‘pseudogene surrogates’, that the rates of DNA loss in *Drosophila* and the cricket *Laupala* are key determinants of genome size. In certain circumstances, pseudogenes can be conserved by a process of gene conversion, such as for immunoglobulin V_H pseudogenes in the chicken [17]. Goncalves *et al.* [18] surveyed human

retropseudogenes and found that genes with a high number of retropseudogene copies tend to be widely expressed, highly conserved and low in (G+C) content.

With the complete genomes of more than 30 prokaryotes and 4 eukaryotes (including the *Caenorhabditis elegans* genome [19]) now published, we have the opportunity to investigate pseudogenes on the genomic scale. Surveys have recently been performed on the genes and pseudogenes of families of G-protein coupled receptors [20, 21]. We have conducted a global survey of the population of pseudogenes in the *Caenorhabditis elegans* genome. Our survey highlights some surprising characteristics of the pseudogene population, such as a markedly uneven chromosomal distribution. In a sense, our survey is a form of 'molecular archaeology', focussing on the characteristics of the 'dead' genes that can be uncovered in a genome. We see it as logically following upon a number of global surveys of the characteristics of the 'living' protein population in the newly sequenced genomes [22-24].

Results and Discussion

General definitions: G, ΨG, and related terms

Given the gene population of the worm genome, what is the size and distribution of the corresponding pseudogene population? To answer this question, we need to define several populations and subpopulations of genes and pseudogenes in the worm. These are described in detail in Table 1 and Figure 1a. We denote by G the total population of confirmed and predicted protein-encoding genes which are taken from the Wormpep18 database. We denote by ΨG the estimated population of pseudogenes that correspond to G. In general, the symbol Ψ before any gene name or gene population name denotes the corresponding pseudogene or pseudogene population. The use of the term pseudogene here does not imply any attempt at parsing the exon

structure, but refers loosely to any pseudogene or pseudogenic fragment readily detected by homology matching and the occurrence of a simple disablement (a premature stop codon or a frameshift). The total ΨG population is thus an initial estimate somewhat in the spirit of recent attempts to estimate the gene complement of the human genome [25 , 26].

We have clustered all genes in G into paralog families. Pseudogenes are assigned to the paralog family of the gene with closest homology to it. (*Singleton genes* are those genes that do not have an obvious paralog.) Pseudogenes are assigned to the paralog familiy of the gene with the closest homology to it. An example of a paralog family with its associated pseudogenes is illustrated (Figure 1b).

As summarized in Figure 1a, we compiled various subsets of ΨG. We denote by ΨG_R a particularly 'reliable' subset of ΨG that is supported by a variety of information such as a complete cDNA match, or a matching protein homology in another organism (see *Methods* for complete description).

We also generated subsets of G and ΨG were generated that relate to levels of gene expression. The set of genes with at least one verifying EST match was derived (G_E). G_E was expanded by including all of the paralogs of G_E proteins to give the $(G_E)_P$ set. A set of genes that were adjudged to be highly expressed was derived from microarray expression data [27] and denoted G_M. The corresponding predicted pool of pseudogenes is denoted ΨG_M.

Estimated size of pseudogene population

The pseudogene population (denoted ΨG) arising from the decay of protein-coding genes in the worm is estimated to comprise 2,168 sequences which is about 12% of the total gene complement (G). This is only an initial estimate of the pseudogene population, that may be

examined for broad trends and characteristics. One should keep in mind that there are a number of obvious factors that may affect the size of ΨG, causing over- or under-estimation:

(i) Dead copies of transposable elements would lead to an over-estimate of ΨG. However, these may be considered validly as pseudogenic fragments, and have been used as such in studies of DNA loss in *Drosophila* [10, 16]. Nonetheless, we do not find any abundant patterns of multiple protein-homology hits in the genomic DNA that would be indicative of a major unknown transposable element (see below). Only ~5% of our total potential pseudogene matches are deleted because of matches to known transposable element proteins (see *Methods*).

(ii) The size of ΨG here may be an underestimate as we do not include pseudogenes that *only* have the less obvious coding disablements, such as damaged splicing signals. However, our search for only frameshifts and premature stops is supported by the fact that 6% of the Sanger Centre-annotated pseudogenes would be missed by this procedure.

(iii) Some annotated genes may in fact be pseudogenes, as the disablement is undetectable by gene prediction procedures (such as a disabled promoter).

(iv) Conversely, some of our pseudogenes might be parts of real functioning genes that were not annotated in Wormpep. In particular, it is conceivable that some premature stops or frameshifts may indicate a shortened protein that lacks all or part of a domain. However, a search of the scientific literature has revealed that reported cases of this phenomenon are rare, and where they occur they may be pathogenic (and thus unlikely to be conserved, *e.g.*, a germline mutation in the human prion protein gene in a single Japanese patient [28].)

(v) Some of the pseudogenes may arise because of sequencing errors (and so should be annotated as genes). However, the reported overall error rate in sequencing is low (<1 in 10,000 bases) [29].

(vi) Some of our pseudogene fragments that are extrapolated from Wormpep may comprise genomic-level repeats; however, we have taken measures to avoid this problem (see *Methods*).

(vii) Some pseudogenes may be fragments of two separate pseudogenes; this problem is minimized in the present work by merging some pseudogene matches along the genomic DNA, with a procedure described below in *Method*s.

Pseudogene subpopulations

Highly expressed genes appear to have fewer dead gene copies or fragments. When only EST-matched genes are considered, ΨG_E corresponds to 5% of G_E (363 predicted pseudogenes) (Table 1). (Intermediate between these, there are 1,165 predicted pseudogenes that correspond to a gene with an EST match or that are paralogous to a gene with an EST match, $\Psi(G_E)_P$.). For pseudogenes related to genes that are highly expressed according to microarray data (*i.e.*, those that comprise the G_M data set), the corresponding pseudogene complement is about 7% of the size of G_M. Interestingly, singleton genes (*i.e.* those with no close paralogs) have a smaller relative population of pseudogenes (corresponding to 11% of the total number of singleton genes) yet constitute 32% of the gene population. The most reliable subset of the pseudogene population (ΨG_R) is about half of the total for ΨG (Table 1). The sizes of ΨG_R, the most reliable subset of pseudogenes, and G_R are not directly comparable as ΨG_R is compiled from a variety of sources (Table 1).

Intronic pseudogenes are pseudogenes that are contained completely within a single intron. A substantial fraction of ΨG is estimated to be intronic (39%) (Table 1). Interestingly, there is no preference for sense or antisense alignment for an intronic pseudogene relative to the exons of the surrounding gene (53% are antisense). This indicates that the existence of pseudogenes in an intron has no relation to the transcription and splicing of a gene.

A key consideration is the proportion of ΨG that are processed pseudogenes. Processed pseudogenes are derived originally from messenger RNA transcripts that have been reverse-transcribed and re-integrated into the genome. In a sense, these pseudogenes are not indications of ailing families of proteins, but rather the opposite; one might expect more processed pseudogenes for genes that are highly or widely expressed. They have the following features: (i) they lack the introns of the gene from which they are derived; (ii) they tend to have a characteristic polyadenine tail; (iii) they lack the promoter structure of the gene from which they are derived; (iv) they have short direct repeats (about 9-15 base pairs) at their N- and C-termini [2]. We could not find any mention of processed pseudogenes in the worm in the scientific literature. We estimated the proportion of processed pseudogenes in ΨG using a simple heuristic that involved looking for stretches of coding sequence that could not be in the pseudogene without processing (which we have termed ‘exon seams’) and also for evidence of a polyadenine tail (see *Methods*). According to the exon seams identification, there appear to be few pseudogenes that result from processing in ΨG (totalling 208, 10%). We could not find any obvious subpopulation of pseudogenes with an elevated adenine content 3’ to their homology segment that would indicate a polyadenine tail. The size of the estimated population of processed pseudogenes here contrasts substantially with the human genome, where about 80% of the pseudogenes are predicted to be processed [30].

Chromosomal distribution of pseudogenes

We mapped the positions of pseudogenes and genes along each of the six worm chromosomes. Pseudogenes appear to be more abundant nearer the ends or 'arms' of the chromosomes (Figure 2). When the distributions for the individual chromosomes are merged, we find that 53% of the pseudogenes are in the first and last 3 Mb of the chromosomes, compared to only 30% of the genes. It was previously noted [19] that the proportion of genes with similarities to other organisms tends to be lower on the chromosomal arms. The pseudogene distribution along the chromosomes correlates with this observation and supports the idea of more rapidly evolving genomic DNA towards the ends of the chromosomes [19]. The same trend for increased occurrence of pseudogenes is observed for the various pseudogene subpopulations. In particular, for the G_E subset, 50% of the pseudogenes are in the first and last 3Mb of genomic DNA (depicted in Figure 2). The analogous number for $(G_E)_P$ is 53%, and one also gets similar results for the highly expressed subset G_M. For the most reliable subset ΨG_R, this proportion is lower (40% in the first and last 3Mb). This may be related to the fact that genes with homology to proteins from other organisms are more prevalent towards the center of the chromosomes [19].

As is also shown in Figure 2 (legend), the distribution of pseudogenes between the chromosomes is also uneven. For each chromosome, we calculated the proportion of 'dead' genes (equal to $|\Psi G_n| / [|\Psi G_n| + |G_n|]$, where $|G_n|$ is the size of the gene population G_n for chromosome *n* and $|\Psi G_n|$ is the number of pseudogenes). Chromosome IV appears to be the most 'dead', chromosome II the least. (The same trend is also found for the ΨG_R subset, as noted.) This variation in the proportion of pseudogenes between chromosomes may be due to specific gene families, or perhaps recently defunct families of genes.

We looked for recurrent pairs of predicted pseudogenes distributed along the chromosomes that may perhaps indicate some undiscovered transposable element. The most frequent pair patterns are tabulated (Table 2). The most common pseudogene pair is for a seven-transmembrane gene family, represented by the gene B0334.7. None of the top recurrent pairs appears indicative of a transposable element.

Disablements, length and composition of ΨG matches

The obvious disablements in the pseudogene population (*i.e.*, frameshifts or premature stops) are tallied in Figure 3a. A high proportion of ΨG has only one disablement over the length of genomic sequence aligned (44%). This may indicate an evolutionarily young pseudogene population that is rapidly deleted from genomic DNA. (Alternatively, it may just reflect the fact that pseudogenes with more than one disablement tend to have less similarity to known proteins than those with a single disablement.) In general, non-coding frameshifts (of either one or two bases) and premature stop codons are approximately evenly represented in the pseudogene fragments detected (Figure 3a). A similar trend for disablements is seen for ΨG_R (data not shown).

The length distribution for the homology matches for pseudogenes is shown in Figure 3b, compared to the length distribution for exons in known worm genes. The modes of these distributions are similar. The mean length of these matches is 338 nucleotides, somewhat larger than the mean length of a worm exon (210 nucleotides), because the distribution for the pseudogenes has a somewhat longer tail. Over medium-range lengths (300 to 500 nucleotides) pseudogenic fragments are about twice as prevalent as exons. The long tail is probably due firstly to processed pseudogenes and secondly to matches against genomic DNA where a gap has been introduced over the length of a small intron. The maximum length is 3,156 nucleotides for a

pseudogene that is most similar to the gene W08D2.5. This is probably a processed pseudogene as there is no evidence of the exon structure of one of its paralogs. In general, however, these matches will not correspond to the length of the whole pseudogene, and are only used to detect the presence of a pseudogene at a particular genomic locus (see below in *Methods*). We do not observe any preference in the pseudogenic homology matches for the N- or C-termini of the corresponding worm protein, for either the processed or unprocessed pseudogenes (50% of those estimated to be unprocessed and 57% of those estimated to be processed tend toward the C-terminus).

As shown in Figure 3c, we measured the amino-acid composition of G (the Wormpep18 protein complement) and the implied amino-acid composition of both ΨG and random non-repetitive genomic sequence (Figure 3c). The amino-acid composition for the pseudogenes is generally intermediate between the composition of random genomic sequence and the composition of the Wormpep18 proteins (Figure 3c), being closer to random than to Wormpep18 (14 out of 20 residues). One would expect older pseudogenes to be closer to random sequence than younger ones, so study of the amino acid composition in this way may indicate from genome to genome of the overall age of the pseudogene population. (However, of course, the actual age of a pseudogene subpopulation will be dependent in a complex way on rates of genomic deletion / insertion and point mutation.)

In our composition analysis, we find that the most enriched residues in ΨG relative to G are *Phe, Ile, Leu* and *Lys*, and the most depleted residues are *Asp*, *Ala*, *Glu* and *Gly* relative to the worm proteome. The enrichment in Phe is particularly interesting as the number of codons for this residue is small (two, TTT and TTC) (Figure 3c). Moreover, the enrichment of Phe and Lys in the ΨG and random sequences relative to G is perhaps related to an underlying trend for local A/T mononucleotide repeats in the genome (data not shown). Also, Lys is preferred to the physico-

chemically similar Arg in the *C. elegans* proteome even though the former has only two codons, compared with six for the latter (Figure 3c and Ref. [31]). Lys, in fact, has been found to be the amino acid that varies most in composition between various genomes [32]. The amino-acid composition of ΨG_R was also derived and yields the same results as described above (data not shown).

Distribution in terms of gene paralog families

We clustered the genes and pseudogenes in the worm genome into paralog families. An example of a paralog family is illustrated in Figure 1a. For each family, as shown in Figure 4, we plotted the number of genes versus the number of pseudogenes. Clearly, the number of pseudogenes per family is not correlated with the number of genes. The large families that have an extensive graveyard of pseudogenes relative to their living population, or vice versa, are labelled with their family representatives. Some of these larger families are 'outliers' that deviate from the overall ratio, indicating a dynamic genome. The family represented by the gene B0403.1 is uncharacterized, but comprises twice as many pseudogenes (31 in total) as genes (16 in total).

In Table 3a, we list the largest sequence families in the worm, ranked by their number of genes and pseudogenes. There are named for their particular family representative. Four of the top ten paralog gene families when ranked by number of pseudogenes are functionally uncharacterized. Moreover, three of the pseudogene top-10 are amongst the biggest families when we rank according to number of genes. These large, evolutionarily dynamic seven-transmembrane (7-TM) receptor families are represented by the genes B0334.7, B0213.7 and C03A7.3. The B0334.7 family is the largest and has about one pseudogene for every four genes, which is close to the overall ratio for genes and pseudogenes in the genome (Figure 4). The occurrence of the reverse-

transcriptase and the TcA transposase families in the top ten list may indicate parts of an unknown transposable element that we failed to mask for.

The pseudogene family rankings are similar for the EST-matched genes (ΨG_E) (Table 3a). If the higher e-value threshold of 0.01 is used instead of 0.001 for worm protein homology matching, there is little change in the most prevalent families for pseudogenes (Table 3a footnote). This suggests a fundamental robustness to these rankings of gene paralog families. The additional pseudogenes pulled in by the less stringent e-value threshold (0.01) presumably represent more ancient pseudogenes. Thus, the fact that the rankings change little suggests that the older pseudogenes have the same distribution of families as more modern ones.

In addition, we found 150 pseudogenic fragments that were similar to representative sequences from the PROTOMAP database but did not have detectable homology to a worm protein (Table 3b). These 'PROTOMAP pseudogenes' either result from horizontal transfer or have diverged too far for the homology to their parent worm protein to be detected. Or perhaps they are even remnants of gene families that have completely died out in the worm. We list the biggest families of PROTOMAP pseudogenes in Table 3b. The top match is an uncharacterized ORF of yeast (yja7_yeast, yeast ORF name YJL007C), which has no other reported homologs, whereas the second and third are similar to mammalian proteins with known functions (Table 3b).

Protein 'pseudofolds' and transmembrane assignments

The proteins encoded by the worm genome have previously been assigned to globular protein domain folds from the SCOP (version 1.39) database and 'top 10' lists of the most common folds in the worm have been constructed [23, 33]. Here, we tried to perform the analogous procedure on the pseudogene population. Where possible, we assigned one of the known protein

folds to each identified pseudogene based on standard approaches. In particular, for every pseudogene, the structural assignments of its closest gene homolog were considered as implied structural assignments (see *Methods*). Then we ranked the pseudogenes in terms of these implied structural assignments or 'pseudofolds' (Figure 5). Overall, there is a decrease in assignability to a SCOP domain for the pseudogene population (12% have an assignment) compared to the gene population (24%). This may be due to truncation or deletion of genomic DNA.

In Figure 5, we ranked the pseudogenes in terms of these implied structural assignments or 'pseudofolds'. The prevalence of different globular folds is somewhat different for the gene and pseudogene populations, although six folds occur in both top-ten lists (Figure 5). Examination of 'pseudofolds' may give an indication of protein structures that have fallen out of favour evolutionarily. Two of the top ten pseudofolds occur infrequently in the worm proteome and thus may be folds that have lost some utility for the worm; the DNase-I-like fold ($\alpha+\beta$ class) and the ovomucoid PCI-like inhibitor fold, which is small and disulphide-rich (Figure 5). The immunoglobulin-like fold, which is in the all–β folding class, is the top fold in G, but is the second-ranking fold for ΨG. This fold is much more abundant in the worm than in any completely sequenced microbial organism [23]. The most common pseudofold for ΨG is C-type lectin fold, which has only been found in eukaryotes [34].

Previously, the worm gene population was surveyed for the presence of transmembrane (TM) segments [23]. We tried to perform a similar survey here for the pseudogene population. The proportion of pseudogenes corresponding to a predicted transmembrane protein is the same in ΨG (22%) as in G (22%). In addition, outside of the homology-based pseudogene fragments, transmembrane helices were assigned on six-frame translations of the raw genomic sequence to locate other regions that are transmembrane-protein-like and pseudogenic (see *Methods* for details).

There is a small number of such pseudogenic transmembrane segments with 4 or more predicted transmembrane helices (174 in total). These may be additional deceased transmembrane protein genes.

Conclusions

Our goal in this study was to provide an initial estimate of the size, distribution and characteristics of the pseudogene population in a large metazoan genome, that of *Caenorhabditis elegans*, in the spirit of recent attempts to estimate the total number of genes in the human genome [25 , 26]. We have found 2,168 homology fragments in the worm genome (about 1 for every 8 genes) that appear to be pseudogenic. About a half of these (totalling 1,100) form a most 'reliable' subset of the data. These figures for ΨG may be an over-estimate due to inclusion of dead copies of transposable elements, or of 'unpredicted' genes with disablements that are due to sequencing errors. Contrarily, it may be an under-estimate due to disregard for pseudogenes with only the less likely disablements, such as a damaged splicing signal, or because some annotated genes are in fact pseudogenes.

We found few pseudogenes that are apparently due to processing in the worm genome. This is in marked contrast to the situation for the human genome, where 80% of the pseudogenes are thought to be processed [30].

The distribution of the proportion of pseudogenes relative to genes for different gene families is notably uneven, indicative of a highly dynamic genome. There are some examples of gene families with an extensive panel of dead fragments, most notably for families of chemoreceptors and other seven-transmembrane receptors [20, 21]. A future detailed study of the complete chemoreceptor worm 'subgenome' that includes these pseudogenes may shed light on the

evolution of these largely worm-specific proteins. We also found one large functionally uncharacterized gene family that comprises about two-thirds dead genes. Such genes or gene families may be falling out of usage due to removal of the evolutionary pressure for their conservation, or due to recent functional redundancy with another gene family. This may partly explain why fewer pseudogenes occur for genes / gene families that are EST-matched.

There are more pseudogenes relative to genes on the arms of the chromosomes, suggesting that many duplications at the ends of the chromosomes tend to produce unusable genes. This may be because the arms of chromosomes undergo more recombination relative to the overall rate of genomic DNA loss. These areas may be thus more 'unreliable' for encoding genes and functions, but conversely are more likely to spawn new proteins. This may also explain the general depletion of genes homologous to other organisms on the arms of the chromosomes [29].

Acknowledgements

Thanks to Nathan Bowen (University of Atlanta) for LTR retrotransposon data in the worm genome, to Richard Durbin (Sanger Centre, UK) for helpful advice, to Hedi Hegyi for worm protein assignment data and to Valerie Reinke (Stanford University) and Ronald Jansen for microarray expression data of genes in the worm, and to the NIH Structural Genomics Initiative and the Keck Foundation for support.

Methods

Data files used and Pseudogene Annotation Pipeline

We downloaded the following data from the Sanger sequencing center ftp site (ftp://www.sanger.ac.uk, versions present in December 1999): the complete sequences of the six

worm chromosomes, the most current worm protein sequence database (Wormpep18) and GFF data files with annotations for genes and other genomic features that correspond to this Wormpep version. The *C. elegans* genome sequence data is constantly updated and certain regions will undoubtedly be revised in future versions; it should be stressed therefore that our survey results here are just an initial estimate of the pseudogene population. We have arranged our ΨG identification procedure in the form of a pipeline schematized in Figure 1a.

Pipeline Step 1: Sanger Center pseudogene annotations

We started off with a list of 332 pseudogenes annotated in the Sanger Center. This original list is small compared to the final size of ΨG, as the Sanger center annotators did not set out to find all of the pseudogenes in the worm genome (R. Durbin, personal communication). These pseudogenes are included in the clustering procedure for derivation of paralog families described below. Our pseudogene population was derived by looking for a simple disablement (a frameshift or premature stop codon; see below). We calculate that 6% of the Sanger Centre-annotated pseudogenes would not be detectable by looking for a simple disablement.

Pipeline Step 2: FASTA matching to find potential pseudogenes

After Wormpep18 was initially masked for low complexity regions with the program SEG [35], the sequence alignment programs TFASTX and TFASTY (version 3.1t13) [36] were used to compare the complete Wormpep18 against the worm genome (in six-frame translation). A list of representatives for the SCOP database (version 1.39) and for sequence clusters from PROTOMAP (version 1) (ref. [37]) was also compared against the 99-megabase worm genomic DNA. (PROTOMAP is a database that comprises the whole of SWISSPROT clustered into families.)

Sequences were checked for an obvious (sequence-length dependent) coding disablement (*i.e.*, either a frameshift or a premature stop codon) indicative of a pseudogene. The potential pseudogene matches were then further filtered and refined as described below.

Pipeline Step 3: Reduction for overlap on the genomic DNA

Initial significant matches of the protein sequences to the genomic DNA (with e-value $\leq$ 0.01) were reduced for redundancy where homologs match the same segment of DNA. A normal e-value of 0.01 was used at this stage as it is consistent with that used in previous genome analyses [22-24]. Firstly, matches were sorted in a list in decreasing order of significance. Then, if a match was selected, any matches extensively overlapping it were excluded from subsequent selection (allowing for a small margin of overlap of thirty nucleotides). This (de)selection procedure was continued until the end of the list was reached.

Pipeline Step 4: Prevention of over-counting for adjacent matches

Some of these initial matches may correspond to the same pseudogene. Therefore, to avoid over-counting for these worm protein matches, the initial matches were further aligned. The genomic DNA fragment *f* corresponding to each matching protein was extracted. The predicted genomic sequence *g* for each paralog of the initial matching worm protein in the Wormpep18 database was aligned against *f*. The length of genomic sequence (g_{top}) for the top-matching paralog relative to the fragment *f* gives an interval on the genomic DNA within which other less significant matches *f* can be discarded. This second alignment stage insures that two or more initial consecutive matches of a Wormpep18 protein to genomic DNA are not counted as separate

pseudogenes. The gene for g_{top} was also used as the final assignment as the closest homolog/paralog for a particular pseudogene.

Pipeline Step 5: Masking against Sanger Centre annotation and a transposon library

The potential pseudogenes were then filtered for overlap with any other annotations in the Sanger Centre GFF files such as exons of genes, tandem or inverted repeats and transposable elements. We masked for further transposable elements and their associated repeats by comparing a library of sequences for reported (retro)transposons against the complete *C. elegans* genome sequence (including the Tc DNA transposons, the Rte-1 retrotransposon and LTR retrotransposons [38-40]).

Pipeline Step 6: Reduction for possible additional repeat elements

At this point in the pipeline, we have a set of 3,814 pseudogenic fragments, which we denote $\Psi G_{1\text{-}6}$. To delete any possible unknown repeat elements from the total estimated ΨG, any matches to a Wormpep18 protein that recurred more than three times to the same exon (in the absence of additional supporting homology) were deleted.

Pipeline Step 7: Reducing threshold stringency

At this point, we had a set of 2,069 pseudogenes, denoted $\Psi G_{1\text{-}6}$. Next, we reduced the e-value match threshold for pseudogene matches to Wormpep18 from 0.01 to 0.001. However, matches supported by other evidence (such as cDNA or protein homology match) were allowed for e-value up to 0.01. This gave us a new total pseudogene population (denoted $\Psi G_{1\text{-}7}$ or simply ΨG throughout the paper). We had a number of rationales for doing this: (i) comparison of $\Psi G_{1\text{-}7}$ and

$\Psi G_{1\text{-}6}$ gives some indication of the sensitivity of pseudogene annotation to the thresholds; (ii) it also potentially allows one to identify a set of more ancient pseudogenes ($\Psi G_{1\text{-}7} - \Psi G_{1\text{-}6}$); (iii) a FASTA e-value cutoff of 0.01 is expected to give 1 false positive per 100 matches, not a particularly high value, but one that would give a substantial number of false positives in tens of thousands of comparisons that underlie our pseudogene identification.

Processed pseudogenes

We developed a heuristic to assess whether a pseudogene was processed. We estimated whether a pseudogene was processed by looking for 'exon seams' in the DNA segment *f* that contains an homology match to a protein. An exon seam is a short stretch of coding sequence that would not be found uninterrupted in the genomic DNA without processing. We found that ten amino acids was a suitable length for an exon seam. If all but one of the exon seams for any paralogous protein are found in the translation of *f* then the pseudogene is identified as a *possible* processed pseudogene. Processed pseudogenes have a polyadenine tract 3' to their protein homology segment [2]. Polyadenosine tracts are added during messenger RNA processing and are usually between 50 and 200 nucleotides long. Therefore, in addition, we analysed a 50-nucleotide stretch 3' to the pseudogene fragments found in the genomic DNA for any evidence of an elevated adenine content relative to the overall distribution of polyadenine content for predicted genes in the same region.

Clustering of Wormpep18 proteins

The 18,576 proteins on Wormpep18 were clustered using a modification of the algorithm of Hobohm *et al.* [41] for deriving representative lists of protein chains. Pairwise alignment using the

FASTA (version 3.1t13) algorithm [36] was performed to compare proteins. Two proteins were judged similar if they had an e-value for alignment $\leq$ 0.01. Clusters are formed in increasing order of the number of relatives that a sequence has in order to minimize false linkage of multidomain proteins. These clusters are termed *paralog families*. Each cluster is named after its representative Wormpep18 protein. Genes with no close relatives according to this method are termed *singleton genes*.

Fold assignments

For the worm proteome, matches to SCOP (version 1.39) domains and to transmembrane proteins are extrapolated onto Wormpep18 from assignments made previously on Wormpep17 proteins [23]. For the pseudogene complement, implied assignments to SCOP domains and transmembrane proteins are taken from the closest matching Wormpep18 protein for each individual pseudogene or pseudogene fragment.

In addition, we performed transmembrane helix prediction directly on six-frame translations of the raw genomic DNA using a hydropathy scale and 20-residue window as described in previous work [23, 42]. Based on an analysis of the distribution of length of interhelical segments in existing membrane protein structures, we joined two predicted transmembrane helices into the same ‘exon’ if they were separated by less than 40 amino acids. We only flagged the resulting assemblage as a pseudogene if it contained a single stop codon in one of the predicted transmembrane helices. These predicted transmembrane protein regions are masked for overlap with other described genomic features as for the pseudogene homology matching.

Subsets of worm genes

Some of the gene sequences in the Sanger Centre worm genome data are noted as matched to ESTs or full-length cDNA. A further set of EST- and cDNA-confirmed worm gene structures is available from the Intronerator database (http://www.cse.ucsc.edu/~kent/intronerator; [43]). We merged these two sets of notations and derived two sets of EST-verified genes. Firstly, the set of genes with at least one verifying EST were compiled (G_E). Secondly, G_E was expanded by including all of the paralogs of G_E proteins ($(G_E)_P$).

Microarray expression data at four time points in the development of the worm (from egg to adult) is available for a substantial cross-section of worm genes [27]. The average of this expression level may be a rough indicator of whether a gene is more highly expressed or more lowly expressed. (However, microarray data, unlike that from GeneChips or SAGE, gives only approximate qualitative indications of the degree to which various genes are differentially expressed. It is much more accurate in highlighting the genes that change considerably in expression.) A suitable threshold for this average expression was used to compile a data set that comprises about half of the ~18,500 worm genes (totalling 9,991 more highly expressed genes, denoted G_M). The corresponding data set of pseudogenes is ΨG_M.

A subset of more 'reliable' pseudogenes (ΨG_R) was compiled that are supported by a variety of evidence. They are pseudogenes that: (i) are verified by a full-length cDNA or have complete EST coverage; or (ii) are noted as confirmed genes in the Wormbase database (http://www.wormbase.org), excluding those which upon inspection have obviously incorrect genomic structure; or (iii) have been previously annotated by the Sanger Centre annotators as a pseudogene using a gene prediction algorithm; or (iv) have a homology match to another non-worm protein over the length of the pseudogene homology match; or (v) have fifty or more matches to a

worm coding sequence of substantial length (>400 nucleotides). This last condition mainly applies to homologies to chemoreceptor genes and other G-protein-coupled receptors. The corresponding set of whole genes for these is denoted G_R, but is not directly comparable as some of the conditions above do not relate to them.

Data on website

We have constructed a web site (http://bioinfo.mbb.yale.edu/genome/worm/pseudogene) for browsing the pseudogene annotations, along with other genomic features downloaded from the Sanger Centre website. The ΨG_R data can be viewed either by searching for a particular ORF or protein name, by viewing the region around an ORF, or simply by viewing a specified range in the chromosome. The sense and alignment score of all pseudogenes is displayed, and the genomic sequences of aligned segments (along with their amino acid translations) are viewable. We have also linked the results to a variety of available internal and external resources including online databases and structural annotations.

References

1. Weiner, A.M., P.L. Deininger, and A. Efstratiadis. *Non-viral retroposons: genes, pseudogenes and transposable elements generated by the reverse flow of genetic information* (1986) *Annu. Rev. Biochem.*, **55**, 631-661.
2. Vanin, E.F. *Processed pseudogenes: characteristics and evolution* (1985) *Ann. Rev. Genet.*, **19**, 253-272.
3. Mighell, A.J., N.R. Smith, P.A. Robinson, and A.F. Markham. *Vertebrate pseudogenes* (2000) *FEBS Letts.*, **468**, 109-114.

4. Korneev, S.A., J.-H. Park, and M. O'Shea. *Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene* (1999) *J. Neurosci.*, **19**, 7711-7720.
5. Olsen, M.A. and L.E. Schechter. *Cloning, mRNA localization and evolutionary conservation of a human 5HT7 receptor pseudogene* (1999) *Gene*, **227**, 63-69.
6. Feenstra, M., J. Bakema, M. Verdaasdonk, E. Rosemuller, J. van den Tweel, P. Slootweg, R. Weger, and M. Tilanus. *Detection of a putative hla-a*31012 Processed pseudogene in a laryngeal squamous cell carcinoma* (2000) *Gen. Chrom. Cancer*, **27**, 26-34.
7. Fujii, G.H., A.M. Morimoto, A.E. Berson, and J.B. Bolen. *Transcriptional analysis of the PTEN1/MMAC1 pseudogene, pisPTEN* (1999) *Oncogene*, **18**, 1765-1769.
8. Currie, P.D. and D.T. Sullivan. *structure, expression and duplication of genes which encode phosphoglycerate mutase of D. melanogastrer* (1994) *Genetics*, **138**, 353-363.
9. Sullivan, D.T., W.T. Starmer, S.W. Curtiss, M. Menotti, and J. Yum. *unusual molecular evolution of an Adh pseudogene in Drosophila* (1994) *Mol. Biol. Evol.*, **11**, 443-458.
10. Petrov, D., E. Lzovzkaya, and D. Hartl. *High intrinsic rate of DNA loss in Drosophila* (1996) *Nature*, **384**, 346-349.
11. Gojobori, T., W.H. Li, and D. Graur. *Patterns of nucleotide substitutions in pseudogenes and functional genes* (1982) *J. Mol. Evol.*, **18**, 360-369.
12. Li, W.H., C.I. Wu, and C.C. Luo. *Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications* (1984) *J. Mol. Evol.*, **21**, 58-71.
13. Averof, M., A. Rokas, K.H. Wolfe, and P.M. Sharp. *Evidence for a high frequency of simultaneous double-nucleotide substitutions* (1999) *Science*, **287**, 1283-1285.

14. Gu, X. and W.-H. Li. *The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment* (1995) *J. Mol. Evol.*, **40**, 464-473.

15. Ophir, R. and D. Graur. *Patterns and rates of indel evolution in processed pseudogenes from humans and murids* (1997) *Gene*, **205**, 191-202.

16. Petrov, D., T.A. Sangster, J.S. Johnston, D.L. Hartl, and K.L. Shaw. *Evidence for DNA loss as a determinant of genome size* (2000) *Science*, **287**, 1060-1062.

17. Ota, T. and M. Nei. *Evolution of immunoglobulin VH pseudogenes in chickens* (1995) *Mol. Biol. Evol.*, **12**, 94-102.

18. Goncalves, I., L. Duret, and D. Mouchiroud. *Nature and structure of human genes that generate retropseudogenes* (2000) *Genome Res.*, **10**, 672-678.

19. Consortium, T.C.e.S. *Genome sequence of the nematode C. elegans: A platform for investigating biology* (1998) *Science*, **282**, 2012-2018.

20. Robertson, H.M. *Two large families of chemoreceptor genes in the nematodes C. elegans and C. briggsae reveal extensive gene duplication, diversification, movement and intron loss* (1998) *Genome Res.*, **8**, 449-463.

21. Robertson, H.M. *The large srh family of chemoreceptor genes in Caenorhabditis nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses* (2000) *Genome Res.*, **10**, 192-203.

22. Gerstein, M. *A structural census of genomes: Comparing Bacterial, Eukaryotic and Archaeal Genomes in terms of Protein Structure* (1997) *J. Mol. Biol.*, **274**, 562-576.

23. Gerstein, M., J. Lin, and H. Hegyi. *Proteins Folds in the Worm Genome* (2000) *Pac. Symp. Biocomputing*, **5**, 30-42.

24. Jansen, R. and M. Gerstein. *Analysis of the yeast transcriptome with broad structural and functional categories: Characterizing highly expressed proteins* (2000) *Nucleic Acids Res.*, **28**, 1481-1488.

25. Ewing, B. and P. Green. *Analysis of expressed sequence tags indicates 35,000 human genes* (2000) *Nat. Genet.*, **25**, 232-234.

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

27. Reinke, V., et al., *A global profile of germline expression in C. elegans* (2000) *Mol. Cell*, **6**, 1-12.

28. Kitamoto, T., R. Iuszuka, and I. Takeishi. *An amber mutation of prion protein in Gerstmann-Straussler-Scheinker syndrome with mutant PrP plaques* (1993) *Biochem. Biophys. Res. Commun.*, **191**, 709-714.

29. Chervitz, S.A., et al. *Comparison of the complete protein sets of worm and yeast: Orthology and divergence* (1998) *Science*, **282**, 2016-2022.

30. Dunham, I., et al. *The DNA sequence of human chromosome 22* (1999) *Nature*, **402**, 489-495.

31. Nishizawa, M. and K. Nishizawa. *Biased usages of arginines and lysines in proteins are correlated with local-scale fluctuations of the G+C content of DNA sequences* (1998) *J. Mol. Evol.*, **47**, 385-393.

32. Gerstein, M. *How representative are the sequences in a genome? A comprehensive structural census* (1998) *Fold Des*, **3**, 497-512.

33. Murzin, A.G., S.E. Brenner, T. Hubbard, and C. Chothia. *SCOP: A structural classification of proteins database for the investigation of sequences and structures* (1995) *J. Mol. Biol.*, **247**, 536-540.

34. Gerstein, M. and M. Levitt. *A structural census of the current population of protein sequences* (1997) *Proc Natl Acad Sci USA*, **94**, 11911-11916.

35. Wootton, J.C. and S. Federhen. *Analysis of compositionally biased regions in sequence databases* (1996) *Methods Enzymol.*, **266**, 554-571.

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

37. Yona, G., N. Linial, and M. Linial. *ProtoMap: automatic classification of protein sequences and hierarchy of protein families* (2000) *Nucleic Acids Res.*, **28**, 49-55.

38. Youngman, S., H.G.A.M. van Luenen, and R.H.A. Plasterk. *Rte-1, a retrotransposon-like element in Caenorhabditis elegans* (1996) *FEBS Letters*, **380**, 1-7.

39. Bigot, Y., C. Auge-Gouillou, and G. Periquet. *Computer analyses reveal a hobo-like element in the nematode C. elegans, which presents a conserved transposase domain common with the Tc1-Mariner transposon family* (1996) *Gene*, **174**, 265-271.

40. Bowen, N. and J. McDonald. *Genomic analysis of C. elegans reveals ancient families of retro-viral-like elements* (1999) *Genome Res.*, **9**, 924-935.

41. Hobohm, U., M. Scharf, R. Schneider, and C. Sander. *Selection of representative protein data sets* (1992) *Protein Sci.*, **1**, 409-417.

42. Gerstein, M. *Patterns of protein fold usage in eight microbial genomes: a comprehensive structural census* (1998) *Proteins*, **33**, 518-534.

43. Kent, W.J. and A.M. Zahler. *The intronerator: Exploring introns and alternative splicing in C. elegans* (2000) *Nucleic Acids Res.,* **28**, 91-93.

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

Figure Legends

Figure 1:

(a) Schematic showing the derivation of the ΨG data set and its breakdown into subsets. The steps in the derivation of ΨG are summarized in the Methods section. The size of ΨG is indicated for the last two steps in this procedure. The name $\Psi G_{1\text{-}x}$ indicates ΨG after *x* steps. The final ΨG data set comprises 2,168 sequences. The subsets ΨG_M, ΨG_R, ΨG_E and $\Psi(G_E)_P$ that are mentioned in the text are indicated as a Venn diagram.

(b) An example of a paralog family with associated pseudogenes. The positions of genes for the paralog family whose representative is the sequence C02F4.2, are indicated by grey ovals (totalling 40). The pseudogenes are marked with black ovals (totalling 4). A pseudogene fragment (ΨC02F4.2) from chromosome II is shown along with an example of a gene from this paralog family W09C3.6 (which is for a serine/threonine protein phosphatase PP1) with the homologous segment underlined. The pseudogene is interrupted by a frameshift relative to this gene (marked by a # symbol). The corresponding sequence in the gene paralog is boxed in black. This corresponds to one exon of the gene paralog. The stop codon of the gene is marked by an asterisk (*).

Figure 2: The estimated chromosomal distribution of pseudogenes. Each panel depicts the distribution of genes (left hand side) and pseudogenes (right hand side) for the chromosomes I, II, III, IV, V, X. The EST-matched subsets for each chromosome are binned as a dark grey bar with the remainder of the genes pseudogenes as a light grey bar. The bin size is 250,000 bases. The axis for number of pseudogenes is scaled by two (*X2*) relative to the same axis for genes. The total estimated sizes of the chromosomal populations of pseudogenes are as follows (the columns are chromosome name, total number of genes, total number of exons for genes, total number of pseudogenes and the proportion of 'dead' gene copies) :-

Chromosome	$\|G_{chromosome}\|$	**|Exons|**	$\|\Psi G_{chromosome}\|$	$\|\Psi G_{R,\ chromosome}\|$	$\|\Psi G_{chromosome}\| / [\|G_{chromosome}\| + \|\Psi G_{chromosome}\|]$
I	2645	17641	245	116	0.08
II	3338	19931	305	147	0.08
III	2347	15243	253	114	0.10
IV	2757	16824	507	285	0.16
V	4737	26756	616	359	0.12
X	2684	19508	242	79	0.08

Figure 3: Disablements, length and composition for ΨG.

(a) Simple Disablements. This data is only for the ΨG population directly derived from Wormpep18.

(b) Length distribution of pseudogene matches. The distribution of pseudogene match lengths (in nucleotides) is shown as an intermittent line, and of lengths for worm gene exons by a continuous line. The lengths of the Sanger center annotated genes are not included as these are more carefully parsed predictions arising from a gene prediction algorithm. Each point *n* denotes the count of exons or matches for an interval from *n* to 50-*n*. Every fourth point is indicated on the x-axis.

(c) Composition for ΨG. The amino-acid composition of the Wormpep18 database is compared to the implied amino-acid composition of random non-repetitive genomic sequence and the ΨG population. The percentage composition for each of the twenty amino acids is graphed in decreasing order of the implied amino acid composition in the pseudogene set. In the bottom part of the figure, the *ΨG difference* for each amino acid composition is indicated by a bar. This is defined as $(|w-p| + |p-r|) / p$, where w is the amino-acid composition value for the Wormpep18 proteins, r is the implied composition for random genomic sequence and p is the implied pseudogene composition. The asterisk (*) in this graph represents the termination codons. The number of codons for each amino-acid type is written below the one-letter code for the residue.

Figure 4: Plot of the number of genes in a paralog family (G_{family}) versus the number of pseudogenes in a paralog family (ΨG_{family}). The families from the G_E set are marked as grey filled points, with the remainder as unfilled points. The lines indicate the overall ratio of the number of genes to the number of pseudogenes for the whole genome and for the G_E subset. Families with large numbers of genes and/or pseudogenes are labelled with the name of their family representative.

Figure 5: The folds and pseudofolds in the worm genome. The SCOP domain matches (part (a) of the figure) are extrapolated onto Wormpep18 from assignments made previously on Wormpep17 proteins [23]. 'Pseudofold' assignments (part (b)) are taken from the closest matching gene paralog for each pseudogene. The columns are as follows: Rank for folds or pseudofolds (with total numbers in brackets); corresponding rank for pseudofolds or folds; a fold cartoon; the

representative domain, the SCOP 1.39 domain number and a brief description of the fold. The fold cartoons are coloured in a sliding gradient from blue for the N-terminus to red for the C-terminus.

Table 1: Overall statistics for ΨG

					Subsets relating to expression					
	Category	**Total Number in *Category***	**Most reliable subset**	**Previous column as percentage of *Category***	**Genes with EST match**	**Previous column as percentage of *Category***	**Genes in paralog families with EST match**	**Previous column as percentage of *Category***	**Numbers for highly expressed genes derived from microarray data**	**Previous column as percentage of *Category***
Genes	Total	**18,576 (G)**	2,154 (G_R)	12%	7,829 (G_E)	42%	13,417 (G_P)	72%	9,991 (G_M)	54%
	Singletons	5,913	682	12%	2,788	47%	---	---	3,199	54%
Pseudogenes and pseudogene fragments	Total	**2,168 (ΨG)**	1,100 (ΨG_R)	51%	363 (ΨG_E)	17%	1,165 (ΨG_P)	54%	746 (ΨG_M)	34%
	Singletons	269	25	9%	66	25%	---	---	146	54%
	Intronic pseudogenes*	518	181	35%	110	21%	285	55%	196	38%

* The estimated numbers of sense and antisense intronic pseudogenes are 274 (53%) and 244 respectively.

Table 2: Most common pair patterns for predicted pseudogenes along chromosomes

Pair of pseudogenes*	**Number of occurrences**
ΨB0334.7, ΨB0334.7	10
ΨAC3.1, ΨAC3.1	8
ΨB0213.7, ΨB0213.7	5
ΨB0035.13, ΨB0035.13	5
ΨC09E9.2, ΨC09E9.2	4

*Each pseudogene or pseudogene fragment is named according to the paralog family representative of its matching protein, with the prefix *Ψ*. All pair that occur more than three times are shown.

Table 3:

(A) Top paralog families for ΨG and G *

Rankings for ΨG **			**Rankings for G**		
Name of family representative	**ΨG_{family}**	**Note on family**	**Name of family representative**	**G_{family}**	**Note in family**
B0281.2 E	59	Reverse transcriptase	**B0280.8 E**	216	Ligand-binding domain of Nuclear Hormone receptor
B0334.7 E	51	7TM receptor	**B0334.7 E**	193	7TM receptor
B0403.1 E	31	uncharacterised	**B0213.7 E**	188	7TM receptor
B0213.7 E	27	7TM receptor	**B0205.7 E**	124	Casein-kinase protein kinase
AC3.1	22	uncharacterised	**B0047.1 E**	93	MATH domain
C04G2.4 E	21	Major sperm-specific proteins	C03A7.3	70	7 TM receptor
B0205.2	20	uncharacterised	**AH6.1 E**	70	Guanylyl cyclase / receptor tyrosine kinase
B0462.3 E	19	uncharacterised	**B0213.10 E**	70	Cytochrome P450
B0213.1 E	19	TcA transposase family	**B0207.1 E**	70	Protein tyrosine phosphatase
C03A7.3	17	7TM receptor	**AC3.2 E**	68	UDP-glucosyltransferase

*Paralog families for EST-matched proteins are in bold and are labeled with a subscript E. Family representatives are underlined if they are common to both lists.

**The family rankings for ΨG including additional worm protein matches for 0.01 < e-value <= 0.001 were also derived. The total for ΨG including these matches is 2,401 predicted pseudogenes (see Figure 1a). The top ten rankings, in decreasing order, are as follows (# of pseudogenes in brackets): 1. B0281.2 (61); 2. B0334.7 (53); 3. B0403.1 (33); 4. AC3.1 (29); 5. B0213.7 (28); =6. B0205.2 (24); =6. B0213.1 (24); 8. C04G2.4 (22); 9. B0462.3 (20); =10. C03A7.3 (18); =10. B0302.3 (18).

(B) Other pseudogenic homology fragments that match a PROTOMAP family representative but with no detected homology to WormPep

Rank	Name of PROTOMAP family representative	Number of matches	Organism of closest match*	Note on family representative
#1	YJA7_YEAST	7 *******	Yeast	Hypothetical protein in yeast
#2 =	XPD_MOUSE	5 *****	Human	Xeroderma pigmentosum group D complementing protein
#2 =	CPSA_BOVIN	5 *****	Bovine	Cleavage and polyadenylation specificity factor
#4 =	THB_RANCA	4 ****	Xenopus laevis	Thyroid hormone receptor beta
#4 =	SEX_HUMAN	4 ****	Human	SEX gene
#4 =	MDR1_RAT	4 ****	Drosophila	Multidrug resistance protein 1
#7 =	YVFB_VACCC	3 ***	Vaccinia virus	Hypothetical vaccinia virus protein
#7 =	VHRP_VACCC	3 ***	Drosophila	Host range protein from vaccinia
#7 =	IF4V_TOBAC	3 ***	Human	Eukaryotic initiation factor 4A
#7 =	ACRR_ECOLI	3 ***	E. coli	Acrab operon repressor

*Determined by a database search with the PSI-BLAST alignment program [44].

Figure 1a

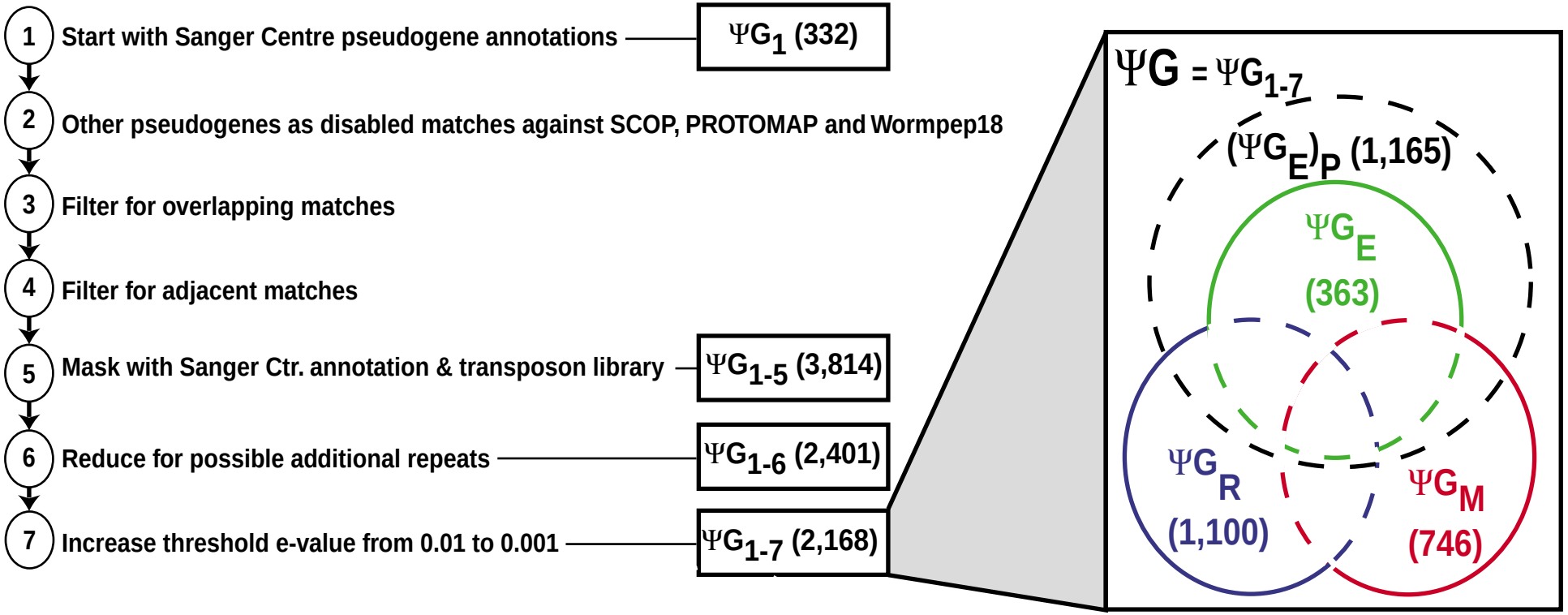
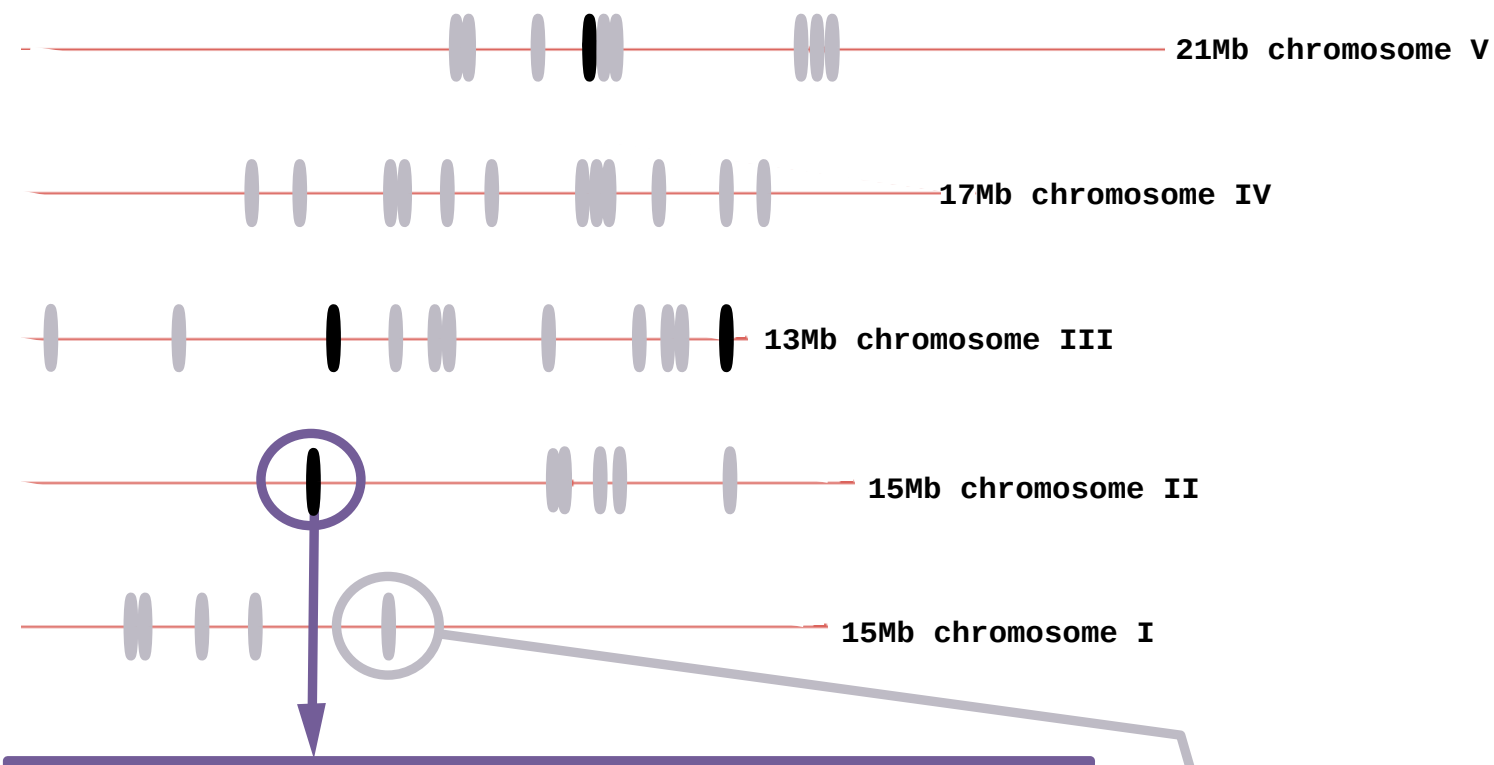

1 Start with Sanger Centre pseudogene annotations
ΨG1 (332)
2 Other pseudogenes as disabled matches against SCOP, PROTOMAP and Wormpep18
3 Filter for overlapping matches
4 Filter for adjacent matches
5 Mask with Sanger Ctr. annotation & transposon library
ΨG1-5 (3,814)
6 Reduce for possible additional repeats
ΨG1-6 (2,401)
7 Increase threshold e-value from 0.01 to 0.001
ΨG1-7 (2,168)
ΨG = ΨG1-7
(ΨGE)P (1,165)
ΨGE
(363)
ΨGR
(1,100)
ΨGM
(746)

Figure 1b



pseudogene fragment on worm chromosome II

TKRTSNGFGQDVVVDLFSILDSGLVARAHXVLQDIFEFFAS
KKMVTIFS#APHSPHSAPHYCAQFDNSAATVKV

a paralog with the homologous segment highlighted (from chromosome I)
(W09C3.6, serine/threonine protein phosphatase PP1)

MTAPMDVDNLMSRLLNVGMSGGRLTTSVNEQELQTCCAVAKSVFASQASLLEVEPPIIVC
GDIHGQYSDLLRIFDKNGFPPDVNFLFLGDYVDRGRQNIETICLMLCFKIKYPENFFMLR
GNHECPAINRVYGFYEECNRRYKSTRLWSIFQDTFNWMPLCGLIGSRILCMHGGLSPHLQ
TLDQLRQLPRPQDPPNPSIGIDLLWADPDQWVKGWQANTRGVSYVFGQDVVADVCSRLDI
DLVARAHQVVQDGYEFFASKKMVTIFSAPHYCGQFDNSAATMKVDENMVCTFVMYKPTPK
SMRRG*

Figure 2

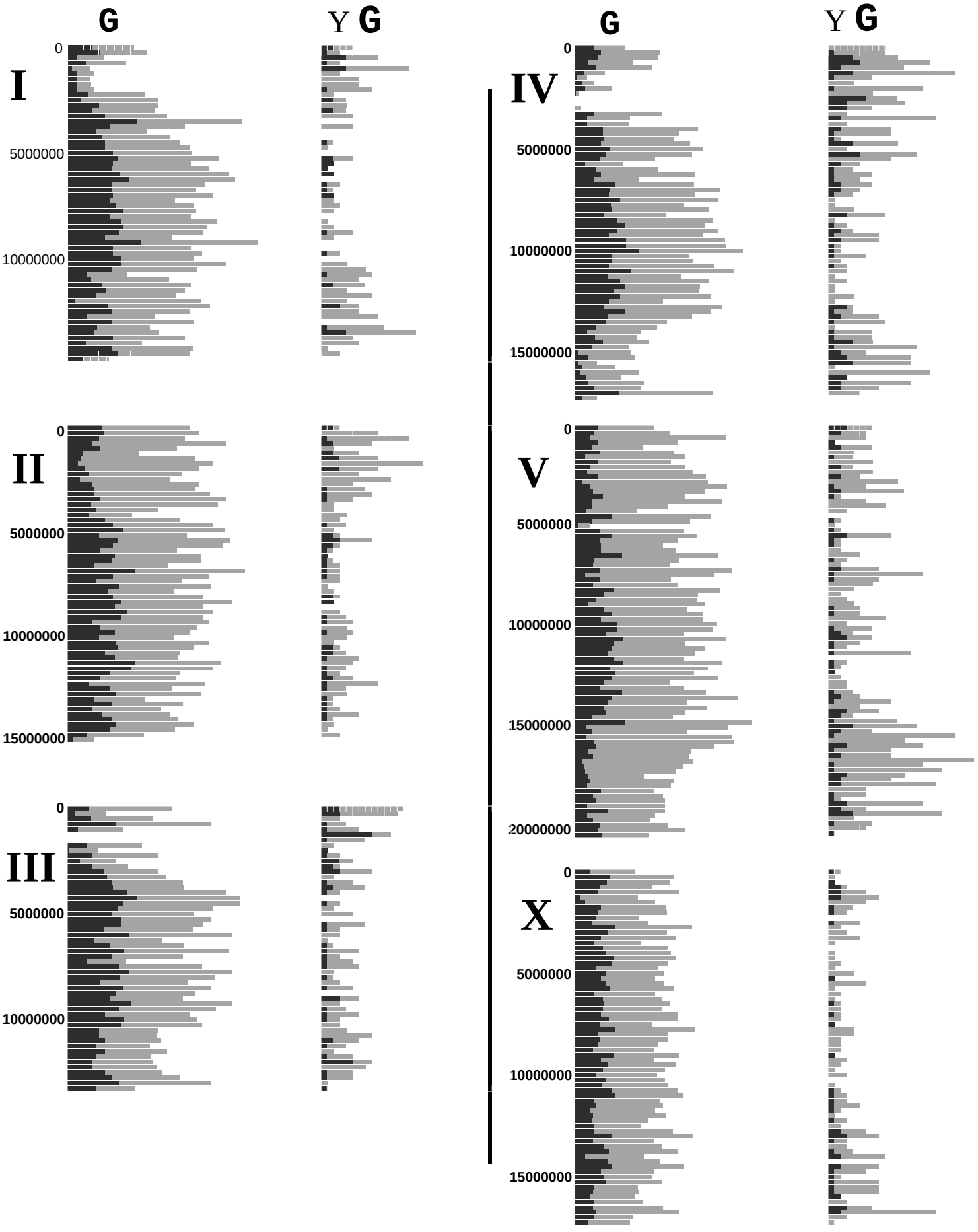

G
Y G
I
0
5000000
10000000
II
0
5000000
10000000
15000000
III
0
5000000
10000000
G
Y G
IV
0
5000000
10000000
15000000
V
0
5000000
10000000
15000000
20000000
X
0
5000000
10000000
15000000

Figure 3(a):

Category	**Numbers**	**Total**
Frameshifts	982 (mononucleotide) 643 (dinucleotide)	1,625
Premature stop codons	---	2,201
Sequences with one disablement	355 (frameshift) 360 (premature stop codon)	715

Figure 3b

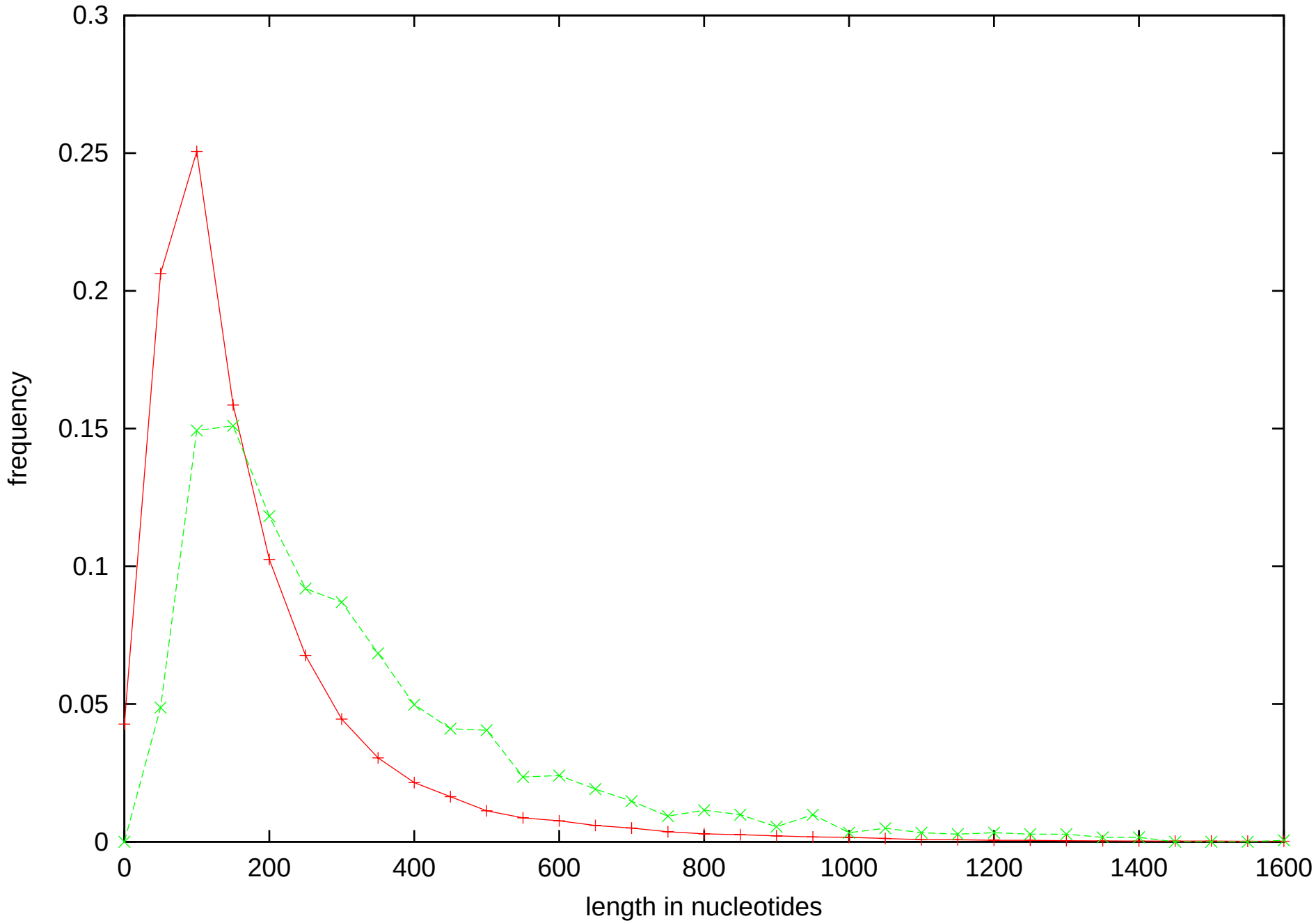

0.3
0.25
0.2
0.15
0.1
0.05
0
frequency
0
200
400
600
800
1000
1200
1400
1600
length in nucleotides

Figure 3c

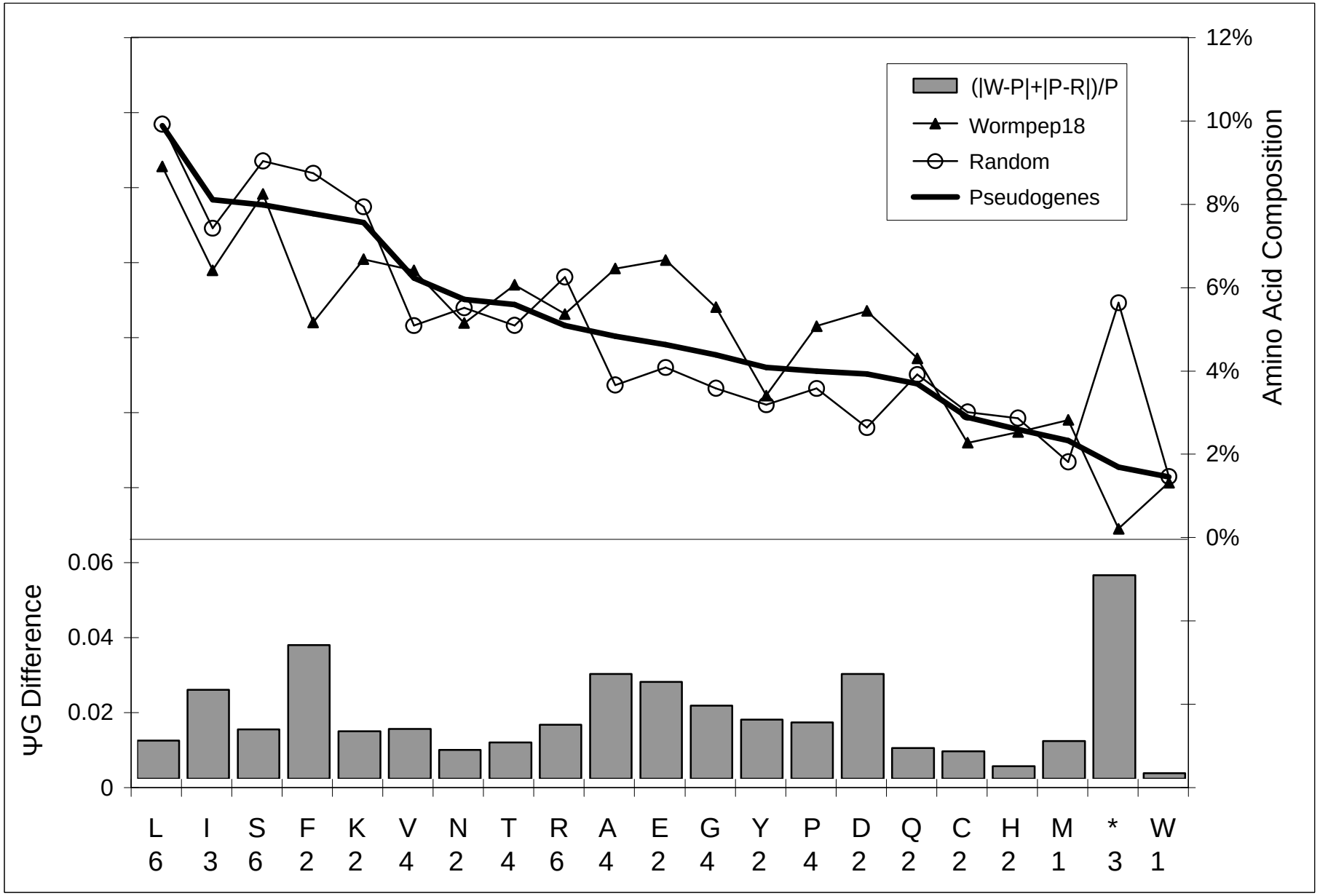

(|W-P|+|P-R|)/P
Wormpep18
Random
Pseudogenes
Amino Acid Composition
12%
10%
8%
6%
4%
2%
0%
ΨG Difference
0.06
0.04
0.02
0
L 6
I 3
S 6
F 2
K 2
V 4
N 2
T 4
R 6
A 4
E 2
G 4
Y 2
P 4
D 2
Q 2
C 2
H 2
M 1
* 3
W 1

Figure 4

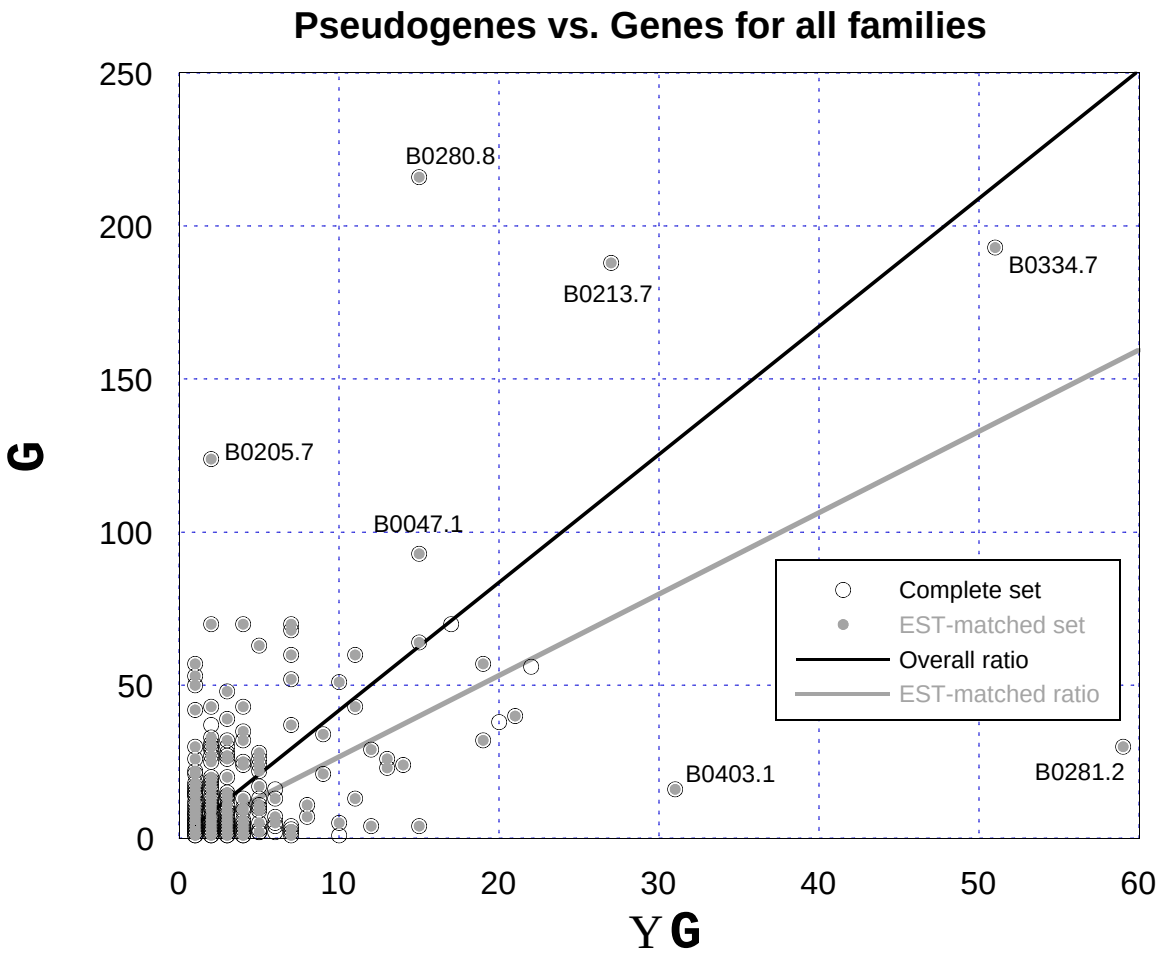
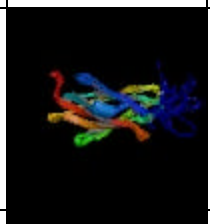
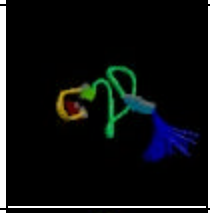
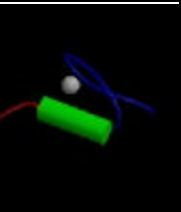
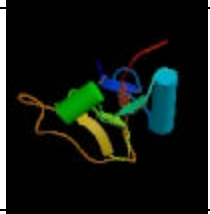
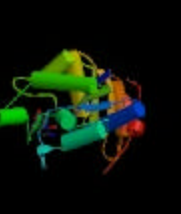
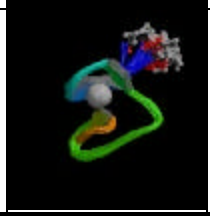

Pseudogenes vs. Genes for all families
250
200
150
100
50
0
G
0
10
20
30
40
50
60
Y G
B0280.8
B0213.7
B0334.7
B0205.7
B0047.1
B0403.1
B0281.2
Complete set
EST-matched set
Overall ratio
EST-matched ratio

Figure 5

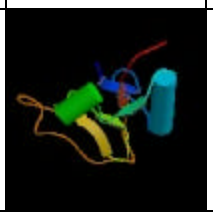
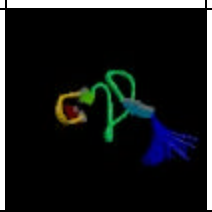
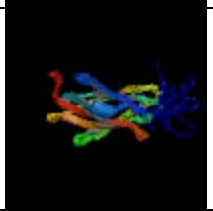
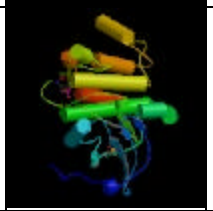
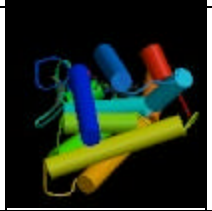
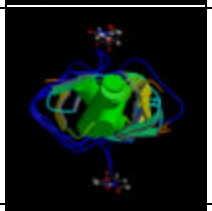
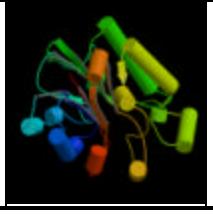
Part *(a)*

Fold Rankings for *G*

G Rank (Number matches)	ΨG Rank	Fold	Representative Domain, SCOP 1.39 Number, Description	G Rank (Number matches)	ΨG Rank	Fold	Representative Domain, SCOP 1.39 Number, Description
1 (769)	**2**		d1ajw__ 2.1 Immuno-globulin	**6** (246)	**8**		d2lbd__ 1.95 Nuc. receptor ligand-binding domain
2 (555)	**6**		d1dec__ 7.3 Knottin	**7** (243)	34		d1a17__ 1.91 Alpha/alpha superhelix
3 (434)	**3**		d3lck__ 5.1 Protein kinase	**8** (227)	17		d1sp2__ 7.31 Classic zinc finger
4 (302)	**1**		d1tsg__ 4.105 C-type lectin	**9** (215)	20		d1dai__ 3.29 P-loop NTP hydrolase
5 (274)	**7**		d1zfo__ 7.33 Glucocorticoid receptor DNA-binding dom.	**10** (197)	13		d2aw0__ 4.34 Ferredoxin

Part *(b)*

Pseudofold rankings for ΨG

ΨG Rank (Number matches)	G Rank	Fold	Representative Domain, SCOP 1.39 Number, Description	ΨG Rank (Number matches)	G Rank	Fold	Representative Domain, SCOP 1.39 Number, Description
1 (39)	**4**		**d1tsg__ 4.105** C-type lectin	**6** (18)	**2**		**d1dec__ 7.3** Knottin
2 (32)	**1**		**d1ajw__ 2.1** Immuno-globulin	**7** (17)	**5**		**d1zfo__ 7.33** Glucocorticoid receptor DNA-binding dom.
3 (27)	**3**		**d3lck__ 5.1** Protein kinase	**8** (15)	**6**		**d2lbd__ 1.95** Nuc. receptor ligand-binding domain
4 (25)	11		**d1cvl__ 3.56** Alpha/beta-hydrolase	**9** (13)	58		**d1bus__ 7.14** Ovomucoid PCI inhibitor fold
5 (23)	63		**d1ako__ 4.93** DNAse-I fold	**9** (13)	19		**d2bnh__ 3.7** Leu-rich, right-handed $\beta\alpha$ superhelix